

# Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)



Linköping Electronic Conference Proceedings Nr. 211

eISSN 1650-3740 (Online)

ISSN 1650-3686 (Print)

ISBN 978-91-8075-774-4 (Print)

2024



Proceedings of the

# 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)

edited by

Thomas Gaillat, Cyriel Mallart, Fabienne Moreau, Jen-Yu Li, Griselda Drouet,  
David Alfter, Elena Volodina and Arne Jönsson

Proceedings and all papers therein  
published under a CC BY 4.0 license:  
<https://creativecommons.org/licenses/by/4.0>

Front cover image by ykaiavu  
via Pixabay

Linköping Electronic Conference Proceedings Nr. 211

eISSN 1650-3740 (Online)  
ISSN 1650-3686 (Print)  
ISBN 978-91-8075-774-4 (Print)

2024



## Preface

The workshop series on Natural Language Processing (NLP) for Computer-Assisted Language Learning (NLP4CALL) is a meeting place for researchers working on the integration of Natural Language Processing and Speech Technologies in CALL systems and exploring the theoretical and methodological issues arising in this connection. The latter includes, among others, the integration of insights from Second Language Acquisition (SLA) research, and the promotion of “Computational SLA” through setting up Second Language research infrastructures.

The intersection of Natural Language Processing (or Language Technology / Computational Linguistics) and Speech Technology with Computer-Assisted Language Learning (CALL) brings “understanding” of language to CALL tools, thus making CALL intelligent. This fact has given the name for this area of research –Intelligent CALL, or for short, ICALL. As the definition suggests, apart from having excellent knowledge of Natural Language Processing and/or Speech Technology, ICALL researchers need good insights into second language acquisition theories and practices, as well as knowledge of second language pedagogy and didactics. This workshop therefore invites a wide range of ICALL-relevant research, including studies where NLP-enriched tools are used for testing SLA and pedagogical theories, and vice versa, where SLA theories, pedagogical practices or empirical data are modeled in ICALL tools. The NLP4CALL workshop series is aimed at bringing together competences from these areas for sharing experiences and brainstorming around the future of the field.

### We invited submissions:

- that describe research directly aimed at ICALL
- that demonstrate actual or discuss the potential use of existing Language and Speech Technologies or resources for language learning
- that describe the ongoing development of resources and tools with potential usage in ICALL, either directly in interactive applications, or indirectly in materials, application, or curriculum development, e.g. Large Language Model exploitation, learning material generation, assessment of learner texts and responses, individualized learning solutions, provision of feedback
- that discuss challenges and/or research agenda for ICALL
- that describe empirical studies on language learner data

In this edition of the workshop a special focus was given to systems relying on AI trained for ICALL tasks. This included, but not only, fine tuning Large Language Models (LLMs) and supervised-learning methods based on learning analytics. Issues related to data in SLA and learner corpus such as collection and feature extraction were also welcome. We encouraged paper presentations and software demonstrations describing the above-mentioned themes for all languages.

### Invited speakers

This year, we had the pleasure to welcome two invited speakers: Helen Yannakoudakis (King’s College London) and Kristopher Kyle (University of Oregon).

**Helen Yannakoudakis** is an Assistant Professor at King’s College London and Affiliated Staff at the University of Cambridge. She is also a Turing Fellow and a Fellow of the Higher Education Academy. Helen is working on machine learning for natural language processing

with a focus on few-shot learning, lifelong learning, multilingual NLP, and societal and health applications. Helen’s work has been deployed under the Cambridge brand (Write & Improve), and has been published in leading venues in the field such as NeurIPS and ACL. She has received funding awards from industry and academia, has served as a keynote speaker and a panelist, and has won international competitions such as the NeurIPS 2020 Hateful Memes Challenge. Among others, she has been invited for spotlight interviews (e.g., DrivenData) and comments by media channels such as Reuters and TechCrunch. Recently, she was invited to stay at Windsor Castle to talk about AI in a two-day consultation on threats and opportunities.

In her talk, Helen Yannakoudakis focused on the potential for integrating large language models (LLMs) into AI-powered language teaching and assessment systems. She explored various research areas including content creation, automated grading, and grammatical error correction, while also addressing the risks and ethical concerns surrounding the use of generative AI in language learning technology. Further, she highlighted the need for further research to better understand the strengths and limitations of LLMs and to address foreseeable risks such as misinformation and harmful bias, and explored several directions for future work.

**Kristopher Kyle** is an Associate Professor of Linguistics and Director of the Learner Corpus Research and Applied Data Science Lab. His research interests include natural language processing, corpus linguistics, second language writing, second language assessment, and second language development. (Norwegian Computing Center).

In his talk, Kristopher Kyle provided a brief overview of the use of natural language processing in research related to language learning and assessment over the past 50 years, culminating in the advent of large language models (LLMs). He then briefly discussed some recent (mis)uses of LLMs in language learning and assessment research. He argued that while some black-box LLM-based systems can achieve results with reasonable metrics, such systems tend to have particularly weak validity arguments. He then argued for the development and use of LLM-based systems that increase the construct validity of common CALL applications such as automated evaluation and feedback systems and introduced some working examples of these systems.

## Previous workshops

This workshop follows a series of workshops on NLP4CALL organized by the NEALT Special Interest Group on Intelligent Computer-Assisted Language Learning (SIG-ICALL<sup>1</sup>). The workshop series has previously been financed by the Center for Language Technology at the University of Gothenburg, the SweLL project<sup>2</sup>, the Swedish Research Council’s conference grant, Språkbanken Text<sup>3</sup>, L2 profiling project<sup>4</sup>, itec<sup>5</sup>, the CENTAL<sup>6</sup> and the Analytics for Language Learning (A4LL) project<sup>7</sup> at LIDILE - Univ Rennes.

Submissions to the thirteen workshop editions have targeted a wide range of languages, ranging from well-resourced languages (Chinese, German, English, French, Portuguese, Russian, Spanish) to lesser-resourced languages (Erzya, Arabic, Estonian, Irish, Komi-Zyrian, Meadow Mari, Saami, Udmurt, Võro). Among these, several Nordic languages have been targeted, namely Danish, Estonian, Finnish, Icelandic, Norwegian, Saami, Swedish and Võro. The wide scope of the workshop is also evident in the affiliations of the participating authors as illustrated in Table 1.

---

<sup>1</sup><https://spraakbanken.gu.se/en/research/themes/icall/sig-icall>

<sup>2</sup><https://spraakbanken.gu.se/en/projects/swell>

<sup>3</sup><https://spraakbanken.gu.se>

<sup>4</sup><https://spraakbanken.gu.se/en/projects/l2profiles>

<sup>5</sup><https://itec.kuleuven-kulak.be>

<sup>6</sup><https://cental.uclouvain.be>

<sup>7</sup><https://sites-recherche.univ-rennes2.fr/lidile/articles/a4all/>

<b>Country</b>	<b>Count</b>	<b>Country</b>	<b>Count</b>
Algeria	1	Japan	7
Australia	2	Lithuania	1
Belgium	18	Netherlands	4
Canada	4	Norway	16
Cyprus	3	Portugal	6
Czech Republic	1	Romania	1
Denmark	5	Russia	10
Egypt	1	Slovakia	1
Estonia	3	Spain	5
Finland	15	Sweden	82
France	29	Switzerland	13
Germany	130	UK	23
Iceland	6	Uruguay	5
Ireland	5	US	14
Israel	1	Vietnam	3
Italy	15		

Table 1: NLP4CALL speakers’ and co-authors’ affiliations, 2012–2024

<b>Workshop year</b>	<b>Submitted</b>	<b>Accepted</b>	<b>Acceptance rate</b>
2012	12	8	67%
2013	8	4	50%
2014	13	13	77%
2015	9	6	67%
2016	14	10	72%
2017	13	7	54%
2018	16	11	69%
2019	16	10	63%
2020	7	4	57%
2021	11	6	54%
2022	23	13	56%
2023	18	12	67%
2024	23	19	82%

Table 2: Submissions and acceptance rates, 2012-2024

The acceptance rate has varied between 50% and 82%, the average being 65% (see Table 2). Although the acceptance rate is rather high, the reviewing process has always been very rigorous with two to three double-blind reviews per submission. This indicates that submissions to the workshop have usually been of high quality.

## Program committee

We would like to thank our Program Committee for providing detailed feedback for the reviewed papers:

- Alfter David - University of Gothenburg (Sweden)
- Ar Rouz David - Université Rennes 2 (France)
- Ballier Nicolas - Université Paris Cité (France)
- Balvet Antonio - Université de Lille (France)
- Belan Sophie - Université de Nantes (France)
- Bexte Marie - FernUniversität in Hagen (Germany)
- Bibauw Serge - Université catholique de Louvain (Belgium)
- Caines Andrew - University of Cambridge (United Kingdom)
- Cornillie Frederik - Katholieke Universiteit Leuven (Belgium)
- De Kuthy Kordula - University of Tübingen (Germany)
- Drouet Griselda - Université Rennes 2 (France)
- El Ayari Sarra - CNRS (France)
- Evain Christine - Université Rennes 2 (France)
- Gaillat Thomas - Université Rennes 2 (France)
- Graën Johannes - Universität Zürich (Switzerland)
- Hamilton Clive - Université Paris Cité (France)
- Horbach Andrea - FernUniversität in Hagen (Germany)
- Jönsson Arne - Linköping University (Sweden)
- Laarmann-Quante Ronja - Ruhr-Universität Bochum (Germany)
- Lange Herbert - University of Gothenburg (Sweden)
- Ljunglöf Peter - University of Gothenburg (Sweden)
- Mallart Cyriel - Université Rennes 2 (France)
- Mieskes Margot - Darmstadt University of Applied Sciences / Hochschule Darmstadt (Germany)
- Moreau Fabienne - Université Rennes 2 (France)



- Munoz Garcia Margarita - Université Rennes 2 (France)
- Muñoz Sánchez Ricardo - University of Gothenburg (Sweden)
- Nicolas Lionel - Eurac Research (Italy)
- Pado Ulrike - Hochschule für Technik Stuttgart / University of Applied Sciences of Stuttgart (Germany)
- Paquot Magali - Université catholique de Louvain (Belgium)
- Sarré Cédric - Sorbonne Université (France)
- Stemle Egon - Eurac Research (Italy)
- Vajjala Balakrishna Sowmya - University of Tübingen (Germany)
- Valdez Cristian - Université Paris Cité (France)
- Volodina Elena - University of Gothenburg (Sweden)
- Zesch Torsten - FernUniversität in Hagen (Germany)

We intend to continue this workshop series, which so far has been the only ICALL-related recurring event based in the Nordic countries, Belgium and France. Our intention is to co-locate the workshop series with the two major LT events in Scandinavia, the Swedish Language Technology Conference (SLTC) and the Nordic Conference on Computational Linguistics (NoDaLiDa), thus making this workshop an annual event. Through this workshop, we intend to profile ICALL research in Nordic countries as well as beyond, and we aim at providing a dissemination venue for researchers active in this area.

## Workshop website

[https://nlp4call.github.io/current/past\\_editions.html](https://nlp4call.github.io/current/past_editions.html)

## Workshop organizers

### Université Rennes 2, France

Thomas Gaillat, Cyriel Mallart, Fabienne Moreau, Jen-Yu Li, Griselda Drouet - Linguistique Ingénierie et Didactique des Langues (LIDILE)

### University of Gothenburg, Sweden

David Alfter, Gothenburg Research Infrastructure in Digital Humanities (GRIDH)  
Elena Volodina, Språkbanken Text

### Linköping University, Sweden

Arne Jönsson

## Acknowledgments

We gratefully acknowledge the financial support from the French *Agence Nationale de la Recherche* and the funded project *Analytics for Language Learning A4LL* - ANR-22-CE38-0015.

## Content

Preface	i
<i>Thomas Gaillat, Cyriel Mallart, Fabienne Moreau, Jen-Yu Li, Grisela Drouet, David Alfter, Elena Volodina and Arne Jönsson</i>	
Out-of-the-Box Graded Vocabulary Lists with Generative Language Models: Fact or Fiction?	1
<i>David Alfter</i>	
Investigating Acoustic Correlates of Whisper Scoring for L2 Speech Using Forced alignment with the Italian Component of the ISLE corpus	20
<i>Nicolas Ballier and Adrien Méli</i>	
Leading by Example: The Use of Generative Artificial Intelligence to Create Pedagogically Suitable Example Sentences	33
<i>Jasper Degraeuwe and Patrick Goethals</i>	
Potential of ASR for the study of L2 learner corpora	49
<i>Sarra El Ayari and Zhongjie Li</i>	
Enhancing a multi-faceted verb-centered resource to help a language learner: the case of breton	59
<i>Annie Foret, Erwan Hupel and Pêr Morvan</i>	
Evaluating Automatic Pronunciation Scoring with Crowd-sourced Speech Corpus Annotations	67
<i>Nils Hjortnaes, Daniel Dakota, Sandra Kübler and Francis Tyers</i>	
Opinions Are Buildings: Metaphors in Secondary Education Foreign Language Learning	78
<i>Anna Huelsing and Andrea Horbach</i>	
Investigating strategies for lexical complexity prediction in a multilingual setting using generative language models and supervised approaches	96
<i>Abdelhak Keliou, Mathieu Constant and Christophe Coeur</i>	
Developing a Pedagogically Oriented Interactive Reading Tool with Teachers in the Loops	115
<i>Mihwa Lee, Björn Rudzewitz and Xiaobin Chen</i>	
Developing a Web-Based Intelligent Language Assessment Platform Powered by Natural Language Processing Technologies	126
<i>Sarah Löber, Björn Rudzewitz, Daniela Verratti Souto, Luisa Ribeiro-Flucht and Xiaobin Chen</i>	
Jingle BERT, Jingle BERT, Frozen All the Way: Freezing Layers to Identify CEFR Levels of Second Language Learners Using BERT	137
<i>Ricardo Muñoz Sánchez, David Alfter, Simon Dobník, Maria Szawerna and Elena Volodina</i>	
Generating Contexts for ESP Vocabulary Exercises with LLMs	153
<i>Iglika Nikolova-Stoupak, Serge Bibaw, Amandine Dumont, Françoise Stas, Patrick Watrin and Thomas François</i>	
Automatic Text Simplification: A Comparative Study in Italian for Children with Language Disorders	176
<i>Francesca Padovani, Caterina Marchesi, Eleonora Pasqua, Martina Galletti and Daniele Nardi</i>	

A Conversational Intelligent Tutoring System for Improving English Proficiency of Non-Native Speakers via Debriefing of Online Meeting Transcriptions	187
<i>Juan Antonio Pérez-Ortiz, Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Felipe Sánchez-Martínez, Roman Chernysh, Gabriel Mora-Rodríguez and Lev Berezhnoy</i>	
Evaluating the Generalisation of an Artificial Learner	199
<i>Bernardo Stearns, Nicolas Ballier, Thomas Gaillat, Andrew Simpkin and John P. McCrae</i>	
Semantic Error Prediction: Estimating Word Production Complexity	209
<i>David Strohmaier and Paula Buttery</i>	
GRAMEX: Generating Controlled Grammar Exercises from Various Sources	226
<i>Guillaume Toussaint, Yannick Parmentier and Claire Gardent</i>	
LLM chatbots as a language practice tool: a user study	235
<i>Gladys Tyen, Andrew Caines and Paula Buttery</i>	
Sailing through multiword expression identification with Wiktionary and Linguse: A case study of language learning	248
<i>Till Überrück-Fries, Agata Savary and Agnieszka Dryjańska</i>	



# Out-of-the-Box Graded Vocabulary Lists with Generative Language Models: Fact or Fiction?

David Alfter

Gothenburg Research Infrastructure in Digital Humanities (GRIDH)

University of Gothenburg, Sweden

first.last@gu.se

## Abstract

In this paper, we explore the zero-shot classification potential of generative language models for the task of grading vocabulary and generating graded vocabulary lists. We expand upon prior research by testing five different language model families on five different languages. Our results indicate that generative models can grade vocabulary across different languages with moderate but stable success, but producing vocabulary in a language other than English seems problematic and often leads to the generation of non-words, or words in a language other than the target language.

## 1 Introduction

Vocabulary lists have long been a cornerstone in language learning, offering learners a structured approach to building their vocabulary and improving reading comprehension (Laufer, 2006; Webb and Nation, 2017; Miralpeix and Muñoz, 2018). Resources like the Academic Word List (AWL; Coxhead 1998) and the New General Service List (NGSL; Brezina and Gablasova 2015) have proven useful for both learners and teachers.

Graded vocabulary lists are a subset of vocabulary lists that include a *grade* for each vocabulary item, indicating its *difficulty* level for learners. This information empowers learners to understand words at their current level, build their vocabulary progressively, and improve their reading skills. For teachers and curriculum developers, graded lists are essential tools for lesson planning and textbook creation, ensuring learners encounter vocabulary appropriate for their proficiency level (Kilgarriff et al., 2014). The importance of graded vocabulary lists is especially clear in the second language learning (L2) context. They are used in language assessment tests (Coxhead, 2011), as vocabulary learning strategies (LaBontee, 2019), in automated essay grading systems (Pilán et al., 2016; Wilkens

et al., 2022), in text simplification systems (Tack et al., 2016; Yancey and Lepage, 2018), for automatic exercise generation (Alfter et al., 2019; Alfter and Graěn, 2019), to search for appropriate reading materials (Lee and Yeung, 2018; Ehara et al., 2018), or in intelligent tutoring systems (Avdiu et al., 2019).

While graded vocabulary lists have undeniable value, they also come with some limitations. Static vocabulary lists can become outdated as language evolves, and they cannot dynamically adjust to individual learner needs. Furthermore, compiling graded vocabulary lists often requires access to specific textbooks or learning materials, which may not always be readily available or affordable.

The emergence of Generative Language Models (GLMs) presents a potential paradigm shift (Creely, 2024; Godwin-Jones, 2024). These models have demonstrated impressive capabilities in tasks relevant to the L2 context. For example, GLMs can generate difficulty-adapted definitions for words (Kong et al., 2022; Yuan et al., 2022), which helps learners with unfamiliar words. ; simplify complex texts and tailor the difficulty to the learners' needs (Baez and Saggion, 2023); assess essays and provide feedback (Bannò et al., 2024); and perhaps most importantly, GLMs can generate new texts specifically adapted to different difficulty levels (Bezirhan and von Davier, 2023; Kianian et al., 2024; Zualkernan and Shapsough, 2024).

While GLMs hold immense promise, approaching or surpassing human-level performances in some areas (for example in cloze tasks; Rego Lopes et al. 2024), they are not without their drawbacks. Some studies show that current models do not yet outperform task-specific models (Kocoń et al., 2023), that they struggle with vocabulary in an L2 setting (Farr, 2024; Żerkowska, 2024) and lexical complexity prediction (Kelious et al., 2024). Additionally, achieving optimal results with GLMs often requires significant computational resources,

potentially limiting their accessibility.

However, now that it is possible to train GLMs on consumer GPUs without strategies such as off-loading, model parallel, check-pointing (Zhao et al., 2024), the question arises: In the age of GLMs, do we still need graded vocabulary lists? Can end users easily use GLMs for vocabulary grading purposes, and if so, how well do these models perform? In order to shed light on these questions, we formulate and explore the following hypothesis: *GLMs are effective at grading vocabulary*.

Our contributions are:

1. We investigate the utility of generative language models on the task of grading vocabulary for language learners in a zero-shot setting
2. We test five generative language model families on five (European) languages
3. We show that all models show comparable yet underwhelming performance across the five languages

The rest of the paper is structured as follows: Section 2 contextualizes our work and points to the gaps in current research. Section 3 explains the methodology, including data, experimental setup and evaluation criteria. Section 4 presents and discusses the results. Sections 5 and 6 round off the paper with conclusion and future work.

## 2 Related Work

There are two research strands that are closely connected to this line of research: complex word identification (Paetzold and Specia, 2016) and lexical complexity prediction (North et al., 2023b). Complex word identification is concerned with identifying *complex* words with downstream applications such as lexical text simplification (Shardlow, 2013; Maddela and Xu, 2018). It is a binary task (is a word complex or not), and is not specifically targeting the L2 context.

Lexical complexity prediction emerged from complex word identification and aims at classifying the complexity of words on a *graded* scale (e.g., how complex is a word, on a scale from 1 to 4). Lexical complexity prediction is also mainly used for downstream tasks like text simplification (North et al., 2023a; Shardlow et al., 2024b), and is not specifically targeting the L2 context. However, as demonstrated by the ongoing list of shared tasks on the topic (Paetzold and Specia, 2016; Yimam et al.,

2018; Ortiz-Zambrano and Montejo-Ráez, 2020; Shardlow et al., 2024a), it is still an active area of research. The latest lexical complexity prediction shared task was a sub-task of the BEA shared task on multilingual text simplification (Shardlow et al., 2024a).

Recent work on complex word identification and lexical complexity prediction found that ChatGPT only sometimes outperforms task-specific models, mostly in cases when the contexts are dissimilar enough to allow for the discovery of a difference; task-specific models tend to perform better at discriminating the complexity of words even with smaller context variations (Kelious et al., 2024). In the recent shared task on multilingual lexical complexity prediction and lexical simplification, the winning team of sub-task 1 (lexical complexity prediction) used GPT4, with an average Pearson correlation of 0.62 (Enomoto et al., 2024).

On the other hand, generative language models and their potential for on-the-fly generation of learning material is increasingly being investigated. However, the focus of these studies is mostly on text passage generation (Attali et al., 2022; Bezirhan and von Davier, 2023; Peng et al., 2023; Boras et al., 2024) and personalization (Leong et al., 2024; Pesovski et al., 2024).

We fill a critical gap in the literature by investigating the potential of GLMs for graded vocabulary lists and by extending the analysis to multiple different models and multiple languages on comparable data.

## 3 Methodology

In this paper we explore two use cases for GLMs and graded vocabulary lists. First, we suppose that a researcher/learner/teacher is in possession of an ungraded word list that they might want to grade using GLMs. Second, we suppose that no vocabulary list exists, and the researcher/learner/teacher wants to create a graded vocabulary list from scratch using GLMs. In both cases, we compare the output of the GLMs to existing vocabulary lists, using both qualitative and quantitative evaluations (see Section 3.3 for evaluation criteria).

### 3.1 Data

As data for this investigation, we use the freely available CEFRLex<sup>1</sup> lists. These lists are derived

<sup>1</sup><https://cental.uclouvain.be/cefrlex>

from textbooks aimed at learners of different languages and contain among others for each lemma the frequencies at different textbook levels (see Figure 1) according to the Common European Framework of Reference for Languages (CEFR; Council of Europe 2018). We specifically use EFLLex (Dürlich and François, 2018) for English, ELELex (François and De Cock, 2018) for Spanish, FLELex<sup>2</sup> (François et al., 2014) for French, SVALex for Swedish (François et al., 2016) and NT2Lex (Tack et al., 2018) for Dutch<sup>3</sup>.

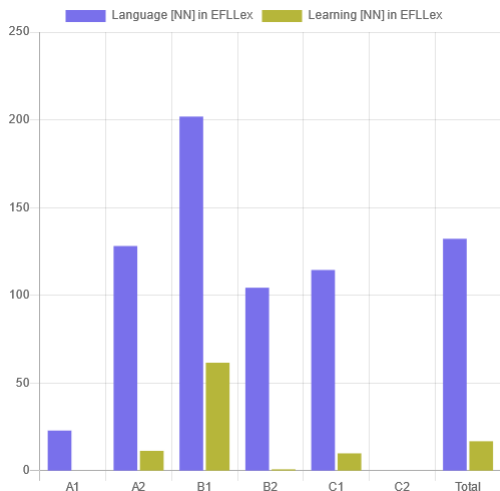


Figure 1: Frequencies across levels for the words ‘Language’ and ‘Learning’ in EFLLex

While the Cambridge English Vocabulary Profile (EVP; Capel 2015) or Pearson’s Global Scale of English (GSE; Pearson 2017) might potentially be more widely used, they are not available in a machine-readable form, being targeted at human end users. Furthermore, they only cover the English language. However, a study comparing these two resources between themselves and to EFLLex found moderate to high correlations both between EVP and GSE (0.85) and between EVP&GSE and EFLLex (0.70; Graën et al. 2020).

As the word lists contain some artifacts and word fragments (e.g., -hour\_day, bly7458/00578, flight\_kl0549), we perform some data cleaning. We only retain single words (excluding multi-word expressions), and exclude words that contain non-alphabetical characters such as digits or other sym-

<sup>2</sup>From the three available versions, for reasons of comparability, we chose the TreeTagger version without automatically assigned CEFR labels.

<sup>3</sup>We do not take into account the sense-disambiguated version of this list, as it mirrors the original list with additional sense labels

bols. We only retain nouns, verbs, adjectives, and adverbs.

Finally, we map each word to the level at which it is first observed (first-occurrence approach). While simple, this method has been shown to perform on-par with more complex level assignment methods (Gala et al., 2013; Alfter, 2021). We opt for a numerical scale rather than the CEFR scale that the word lists are derived from, mapping A1 to 0, A2 to 1, B1 to 2, B2 to 3, and C1 to 4. We disregard C2, which is only included in the French list, as the difference between C1 and C2 is difficult to assess (Springer, 2012; Sung et al., 2015; Isbell, 2017), and the focus of the study lies less in the discriminatory performance at the highest levels but rather a general ability to grade vocabulary from easiest to hardest.

Table 1 shows an overview over the final word lists used in the experiments.

List	WC	WC2
EFLLex (English)	29667	10295
ELELex (Spanish)	14290	13291
FLELex (French)	17237	13242
SVALex (Swedish)	15634	13662
NT2Lex (Dutch)	17743	13972

Table 1: Overview over word counts before (WC) and after (WC2) data cleaning

### 3.2 Experimental setup

We test five popular instruction-tuned model families: Google’s Gemma (Gemma Team et al., 2024), MistralAI’s Mistral (Jiang et al., 2023), Meta’s Llama (Touvron et al., 2023), Microsoft’s Phi3 (Abdin et al., 2024), and OpenAI’s GPT (OpenAI et al., 2024). Specifically, we use Gemma-1.1-2b-it, Gemma-1.1-7b-it, Mistral-7B-Instruct-v0.3, Llama3-8B-Instruct<sup>4</sup>, Phi-3-mini-4K-instruct, and GPT-4o. Gemma, Mistral, Llama, and Phi3 provide small versions of their models (2B to 8B) that do not necessitate massive servers to run, while GPT-4o potentially relies on multiple different models of larger size (cited as exceeding 200B; Ayub 2024) but can be queried programmatically, thus requiring only a paying account and access to the internet.

<sup>4</sup>Preliminary experiments with Llama2-7b-chat showed a strong underperformance in comparison to the other models, an “unwillingness” to follow instructions, and a tendency to mostly respond with a score of 3. As a result, the model was excluded from further experiments.

All models (except GPT4o) are loaded in 4bit quantized form, and GPT4o is queried through its API. All calculations were performed on a single high-end laptop computer with a 12th Gen Intel®Core i7 2.40Ghz processor, 32GB RAM and an NVIDIA GeForce TRX 3080 Ti Laptop GPU graphic card.

Parameters for the models were taken from their respective Huggingface pages with sample code, mirroring a ‘naive’ approach to using GLMs by simply copy-pasting their example code and running it. This means that some models use sampling or have a temperature parameter above zero, reducing the reproducibility of this study. All parameters can be found in Appendix B, Table 8.

### 3.2.1 Generating grades

For the first experiment, we use the word lists as basis and ask the generative language models to grade the vocabulary.

Similar to Enomoto et al. (2024) who prompt GPT4 with a single English prompt for lexical complexity values for different languages, we use a single English prompt for all languages and models, with the first part specified as *system* input if the model supports a *system* role, otherwise prepended to the *user* prompt. The full prompt is:

You are an experienced teacher of *language* as a second language. You can easily assess the difficulty of words in *language* for learners. You assess words on a scale from 0 (easiest) to 4 (hardest). You only answer with a number.

Assess: *word (part-of-speech)*

### 3.2.2 Generating vocabulary list

For the second experiment, we ask the generative language models to generate word lists from scratch.

Given the generation limit of GLMs and the associated cost, and the more qualitative evaluation of this experiment, we opt to prompt each model for a maximum of 100 words per level, using the following prompt. As the output may include repeated words, we take the set of unique words for each level and compare it to the word lists.

You are an experienced teacher of *language* as a second language. You can easily tell which words are suitable for learners of *language* at different levels.

You assess words on a scale from 0 (easiest) to 4 (hardest).

Generate 100 words for learners of level *level*.

## 3.3 Evaluation

First, we evaluate the models according to correctness in predicting grades in comparison to the textbook-derived grades assigned by the first-occurrence approach. For this quantitative evaluation, we use Pearson correlation, Jensen-Shannon distance, accuracy, adjacent accuracy (the prediction is considered correct if it deviates from the target level by at most one level), precision, recall, and F1 score.

Second, we evaluate the quality of the generated graded word lists. For this more qualitative evaluation, we consider coverage of generated vocabulary as the overlap with existing word lists and a more in-depth analysis and discussion.

We also investigate whether there is a link between frequency and discrepancy in prediction. A low frequency in the word list means that the level assignment will be less reliable; if we only observe one occurrence of a word, the level of the word will be the level where it was observed, by definition. If GLMs are *consistent* in grading, then we expect them to grade low-frequency words according to their own internal criteria (as opposed to observed frequency). Further, if GLMs are *consistent* and *correct* in grading vocabulary, then we expect that larger discrepancies are found in words with low frequency, and less discrepancy in high frequency words.

In addition, we explore the impact of the chosen grading scale, investigating whether prompting the models to grade vocabulary on the CEFR scale rather than a numerical scale might improve results. We have opted for a numerical scale because it might be a more generalizable concept for models to work with, rather than the CEFR scale, which the models might have limited knowledge of. For reasons of economy, we only perform this experiment using the best performing model and two word lists: the one it scored worst on, and the one it scored best on.

## 4 Results and Discussion

In this section, we report the results from the experiments and discuss the results. For space reasons, model names and word list names are abbrevi-



ated, with G2 and G7 standing for Gemma-2B and Gemma-7B respectively, GPT for GPT-4o, L8 for Llama3-8B, M7 for Mistral-7B, and P3 for Phi-3; EN for EFLLex, ES for ELELex, FR for FLELex, SV for SVALex, and NL to NT2Lex.

#### 4.1 Generating Grades

As a first measure of comparison, we compare the predicted label distributions to the original label distribution by normalizing the label counts by the total number of items and applying the Jensen-Shannon distance measure (Lin, 1991). This indicates how well the predictions follow the original label distributions, although it gives no indication of the *accuracy* of predicted labels. Table 2 shows the Jensen-Shannon distance between the original label distribution and the predictions for each model.

	G2	G7	GPT	L8	M7	P3
EN	0.30	0.40	<b>0.24</b>	0.43	0.32	0.46
ES	0.31	0.35	0.22	0.33	<b>0.20</b>	0.39
FR	0.48	0.51	0.32	0.40	<b>0.27</b>	0.40
SV	0.22	0.42	<b>0.12</b>	0.30	0.37	0.37
NL	0.47	0.43	0.16	0.32	<b>0.13</b>	0.17

Table 2: Jensen-Shannon distance between the original label distribution and the predicted label distributions by model. Results in bold indicate the best result per language.

In order to check for *accuracy*, we calculate accuracy, precision, recall, weighted F1 score, and adjacent accuracy. For reasons of space, we only report F1 scores in the main body of the paper. The full table including accuracy, adjacent accuracy, precision, and recall, can be found in Appendix A, Table 7. Table 3 shows the weighted F1 scores for each model and word list.

	G2	G7	GPT	L8	M7	P3
EN	0.17	0.18	<b>0.29</b>	0.16	0.24	0.15
ES	0.15	0.19	0.24	0.20	<b>0.28</b>	0.19
FR	0.15	0.12	0.21	0.19	<b>0.28</b>	0.22
SV	0.26	0.30	<b>0.33</b>	0.25	0.18	0.20
NL	0.18	0.19	0.35	0.35	0.36	<b>0.38</b>

Table 3: Results in terms of Weighted F1 score. Results in bold indicate the best result per language.

For comparability to lexical complexity predic-

	G2	G7	GPT	L8	M7	P3
EN	0.03	0.29	<b>0.48</b>	0.36	0.40	0.38
ES	-0.03	0.22	<b>0.42</b>	0.29	0.33	0.26
FR	0.03	0.29	<b>0.46</b>	0.33	0.39	0.37
SV	0.07	0.22	<b>0.39</b>	0.25	0.25	0.29
NL	0.07	0.24	<b>0.38</b>	0.26	0.27	0.32

Table 4: Results in terms of Spearman’s  $\rho$ . Results in bold indicate the best result per language.

tion, we also calculate Spearman’s  $\rho$ .<sup>5</sup> Table 4 shows the results.

Both tables 2 and 3 show a similar trend, with GPT-4o performing best on English and Swedish, Mistral performing best on Spanish and French, and Mistral performing best on Dutch in terms of predicted label distribution but outperformed by Phi-3 in terms of weighted F1 score. Table 4 shows that GPT-4o correlates most with the reference data in all cases, followed by Mistral-7B and Phi3.

Interestingly, although most models are exclusively meant for use with the English language, all models show a rather good cross-linguistic capacity. Further, none of the models performed particularly well in English, or remarkably better on English in comparison to the other languages.

Given possible fluctuations, it seems that both Mistral-7B and GPT-4o are performing similarly well on this task. Given that GPT-4o requires a paying subscription, Mistral-7B seems to be a viable free alternative. We can also observe that Mistral-7B performs quite well across languages, except for Swedish, where the Gemma models show surprisingly good performance, coming second (G7) and third (G2) after GPT-4o. We can also observe that all models except the Gemma family performed best on Dutch. Finally, we may see a language bias: Mistral-7B performed best on Romance languages, while GPT-4o performed best on Germanic languages, potentially reflecting a bias in training data.

#### 4.2 Frequency and Discrepancy

For this experiment, we order each list by total frequency as given in CEFlex and calculate the absolute difference in predicted level and assigned

<sup>5</sup>The lexical complexity prediction tasks indicate both Spearman’s  $\rho$  and Pearson’s correlation coefficient, since the numerical labels can be expressed as continuous numbers. However, we do not assume a normal distribution of the data, which is a prerequisite for Pearson’s correlation coefficient.

level. We then calculate the average discrepancy for the first and last  $x$  entries, varying  $x$  from 10 to 100 in steps of 10. Figure 2 shows the discrepancy for the different values of  $x$ .

As can be seen from the figure, Gemma-2B shows an opposite trend of what would be expected with higher discrepancies for high frequency words, and lower discrepancies for low frequency words. Gemma-7B shows a mixed picture, with the expected trend at  $x = 10, 20, 30$ , but an opposite trend from  $x = 40$ . GPT-4o, Llama3-8B, Mistral-7B, and Phi3-4K display a higher discrepancy for the lowest frequency words and a lower discrepancy for the most frequent words across all languages, following the expected pattern and confirming that GLMs may be useful for grading vocabulary items for which the total observed frequency is too low.

### 4.3 Impact of Grading Scale

As noted previously, we only investigate the impact of the grading scale using the best model and the word lists it performed best and worst on. Based on Table 2, we select Mistral-7B as model and Swedish and Dutch as word lists. For the two word lists, we proceed as described in Section 3.2.1, but we modify the prompt as follows:

You are an experienced teacher of *language* as a second language. You can easily assess the difficulty of words in *language* for learners. You assess words on **the CEFR scale ranging from A1 (easiest) via A2, B1, B2, to C1 (hardest). You only answer with a CEFR label.**

Assess: *word (part-of-speech)*

	Numerical scale	CEFR scale
SV	0.18	0.12
NL	0.36	0.20

Table 5: F1 scores (weighted) for numerical scale and CEFR scale

Table 5 shows the comparison between using a numerical scale versus using the CEFR scale. We can note a marked decrease in performance for both word lists, hinting at the possibility that the language model may not have come into contact with the CEFR in sufficient amounts to be able

to accurately apply it. We also notice a tendency towards predicting A1, which may be due to the problem of *primacy*, a tendency for the model to pick the first alternative from a list of alternatives, previously shown to exist in ChatGPT (Wang et al., 2023).

### 4.4 Generating Vocabulary Lists

In this section we present the results of the vocabulary generation task. During result examination, we noticed that Gemma-7B consistently output numbered lists that only list items 1-10 and 90-100, with ellipsis of the rest. We therefore opted to leave out the results for Gemma-7B in this section.

Table 6 shows an aggregated version over all languages and all levels for vocabulary generation. The table shows that we requested 2500 words from each GLM, with 100 words distributed over five levels for five languages ( $100 * 5 * 5$ ). We can see that only GPT-4o generated the exact number of requested words, Llama-3 generated almost the requested number of words, Gemma did not provide even half of the requested words, while Mistral-7B and Phi-3 overgenerated. However, the generated vocabulary lists contain duplicates. Based on the unique count of words, we can see that GPT-4o was closest to the target, followed by Llama-3 (who overgenerated).

When looking at the number of items generated at the requested level, we can again see that GPT-4o performed best, followed by Mistral-7B. However, Mistral-7B also shows the highest out-of-vocabulary rate, meaning that it generates words that are not present in the reference word list. In terms of overall coverage, we can see that GPT-4o performs best, followed by Mistral-7B and Llama-3-8B.

A detailed investigation of results reveals that Mistral-7B and Llama3-8B tend to group words by categories (numbers, days of the week, months of the year, greetings, travel, family, weather, . . .). Gemma often disregards the requested level and generates a list spanning all levels, grouped by level (easy, moderate, challenging, complex); this behavior is sometimes also observed for Mistral (French and Spanish). Phi3 does generate a list of at least 100 items, but starts repeating the same word after 20-30 words.

In the following, we examine each model language by language and investigate the causes for a low overlap by looking at words that the model generated that were not found in the reference list,

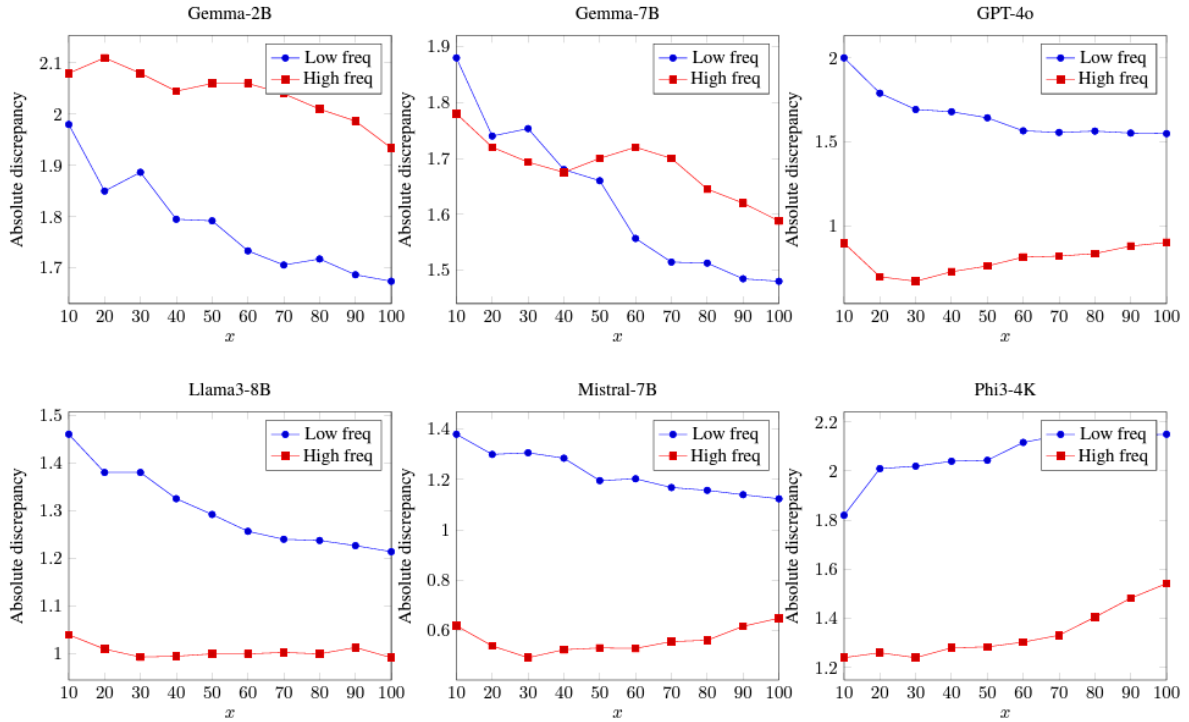


Figure 2: Discrepancies over different values of  $x$ . Each graph shows the average absolute discrepancy over all languages for the most frequent (High freq) and least frequent (Low freq) words when taking into account the  $x$  first and last words of the frequency-ranked list.

	R	G	U	L	D	OOV	OOVR (%)	LC (%)	OC (%)
Gemma-2B	2500	1204	935	156	542	237	25.35	6.24	27.92
Llama3-8B	2500	2433	2181	283	1148	750	34.39	11.32	57.24
GPT-4o	2500	2500	2460	487	1448	525	21.34	19.48	77.40
Mistral-7B	2500	2675	2574	355	1086	1133	44.02	14.20	57.64
Phi3-4K	2500	3073	1285	202	579	504	39.22	8.08	31.24

Table 6: Coverage of generated vocabulary lists aggregated over all languages and levels, with the requested number of words (100 per level per language; R), the number of generated words (G), the number of unique generated words (U), the number of words generated that correspond to the desired level (L), the number of generated words that are in the word list but at a different level (D), the number of generated words that are not in the reference word list (Out-of-vocabulary; OOV), the out-of-vocabulary rate (out of all generated unique words, how many are not in the reference word list; OOV), the level coverage (out of all generated words, how many are in the reference word list at the given level; LC), and the overall coverage (out of all generated words, how many are in the reference word list; OC)

then draw an overarching picture.

#### 4.4.1 English

**Gemma-2B** At level 0, the generated words include conjunctions (or) and prepositions (from) that were excluded from the reference word list due to their part-of-speech. At level 1, the model generates pronouns (I), interjections (hello) and multi-word expressions (family member, thank you) that were also excluded from the reference word list. At level 2, the model generates words that, according to the authors of the paper, are of arguably higher complexity than level 2 (e.g., accountability, resilience, transformative). At level 3, the model generates plausible candidates. At level 4, the model generates words that, again, are arguably above level 4, such as: superimpose, reticent, parsimonious, abrogate, concomitant, magnanimous, prevaricate, obsequious, iconoclast.

**GPT-4o** At level 0, the model includes personal pronouns (you, us) and numbers (zero) that were excluded based on part-of-speech. However, the model also generates some words that are suitable but missing from the reference list (lion, ant). At level 1, the model generates interjections (hello). At level 2, the model generates months of the year and prepositions (between, during). At level 3, the model generates plausible candidates. At level 4, the model generates words that are plausible but includes words of arguably higher complexity, such as: pernicious, surreptitious, vicissitude, obstreperous, prevaricate

**Llama3-8B** At level 0, the model generates numbers (four), multi-word expressions (thank you) and plural forms (socks). At level 1, the model generates words that would potentially be more appropriate at level 0 (lion, rectangle, triangle). At level 2, the model generates plural forms (nuts, pillows), easier words (lemon, omelette) but also plausible candidates. At level 3, the model generates words of a much higher level (inscrutable, garrulous, obfuscate, sagacious, jaded, callipygian). At level 4, the model generates even more complex words (abstruseness, papaphobia, mumpsimus, insouciant, tintinnabulation, perspicacious, zephyrine, gymnosophy).

**Mistral-7B** At level 0, the model generates numbers (zero), prepositions (among, between, from) and question particles (who, where, when, why, what) that were excluded from the reference list based on part-of-speech. At level 1, the model generates interjections (hello), multi-word expressions

(last week, thank you, next week, wake up), and conjugated verb forms (does, hasn't). At level 2, the model generates numbers (four, six) and multi-word expressions (I am fine, what time is it?, I do, you're welcome). At level 3, the model generates words that are of much higher complexity (jocund, aphorism, temerity, blithe, capricious, komphetamology). At level 4, the model also generates words of much higher complexity (obstreperous, sanctimonious, capacious, lachrymose).

**Phi3-4K** At level 0, the model generates numbers (seven, nine, four, six), multi-word expressions (a lot, thank you), but also some plausible missing words (giraffe, kangaroo, lion). At level 1, the model generates plural forms (shoes, socks, pants) and multi-word expressions (thank you, living room). At level 2, the model generates plural forms (shoes, socks, pants) and multi-word expressions (thank you). At level 3, the model generates plausible words but also words with arguably higher complexity (xylography, opulence). At level 4, the model generates plausible words but arguably of higher complexity (obfuscation, zephyr, rambunctious, nebulous, taciturn, dichotomy, ephemeral, ineffable, effulgent, limerence).

#### 4.4.2 Spanish

**Gemma-2B** At level 0, the model generates good candidates that simply are not in the reference word list (día 'day', gato 'cat', perro 'dog', casa 'home'). MWE: gracias de nuevo, por favor too high level: inspirador, felizmente number: uno At level 1, the model generates numbers (uno 'one'), multi-word expressions (gracias de nuevo 'thanks again', por favor 'please'), but also words of a higher complexity (inspirador 'inspiring/inspirer', felizmente 'happily'). At level 2, the model generates plausible candidates. At level 4, the model also generates plausible candidates, although one word seems to be misspelled (\*objetovo, probably objetivo '(an) objective').

**GPT-4o** At level 0, the model generates numbers (nueve 'nine', tres 'three'), feminine forms (hermana 'sister') and multi-word expressions (por favor 'please'). At level 1, the model generates interjections (hola 'hello', gracias 'thanks'), as well as *hermana* and *por favor* from the previous level. At levels 2 and 3, the model generates plausible words. At level 4, the model generates plausible words but also words with a higher complexity (caliginoso 'caliginous', inasible 'ungraspable', imperecedero 'imperishable', impertérito 'undaunted').

**Llama3-8B** At level 0, the model generates some good words (computadora ‘computer’, telefono ‘telephone’), but also some plural forms (animales ‘animals’, recursos ‘resources’, familias ‘families’) and words of higher complexity (pormenor ‘detail’). At level 1, the model generates plural forms (llaves ‘keys’, piernas ‘legs’, brazos ‘arms’, manos ‘hands’), *hermana*, interjections (hola ‘hello’), months of the year, days of the week and numbers. At level 2, the model still proposes *computadora*, *hermana*, and *esposa* ‘wife’. At level 3, the model generates multi-word expressions (lo siento ‘I’m sorry’, me gustaria ‘I would like to have’, hasta luego ‘see you soon’). At level 4, the model generates multi-word expressions and phrases, albeit with only few base constructions, such as *desarrollar* ‘develop’ (estrategias ‘strategies’, habilidades ‘habits’, ...), and *enfermedad de* ‘disease of’ (alzheimer ‘Alzheimer’, cuidados intensivos ‘intensive care’, ...).

**Mistral-7B** At level 0, the model generates a lot of words with articles (el choche ‘the car’, la nariz ‘the nose’, \*el nariz, el pantalón ‘the pants’, el sombrero ‘the hat’, el diente ‘the tooth’, la boca ‘the mouth’) and multi-word expressions (¿donde está el parque? ‘where is the park?’, ¿como se dice en español? ‘how do you say this in Spanish?’, me gusta ‘I like’). At level 1, the model generates numbers, interjections (hola ‘hello’), conjunctions (con ‘with’), conjugated verbs (ríe ‘laughs’, llora ‘cries’), and multi-word expressions (lo siento ‘I’m sorry’, no me gusta ‘I don’t like’). At level 2, the model generates numbers, plural forms (amigos ‘friends’, aguas ‘waters’) and multi-word expressions (buenas tardes ‘good evening’, buenas noches ‘good night’). At level 3, the model generates plausible words, but also plural forms (familiares ‘familiar-ADJ-PL’, misteriosas ‘mysterious-ADJ-PL’, hombres ‘men’, tiempos ‘times’, equipos ‘teams’, ventajas ‘advantages’) and conjugated verb forms (mantiene ‘maintains’, empieza ‘begins’, hablaste ‘you spoke’, cómprame ‘buy me!’). At level 4, the model generates mostly plausible words but also French words (flâner ‘stroll around’) and words with higher complexity (zozobrar ‘capsize’, cenotafio ‘cenotaph’, panoptico ‘panoptic’, acriminarse ‘incriminate oneself’).

**Phi3-4K** At level 0, the model generates interjections (hola ‘hello’, gracias ‘thanks’), multi-word expressions (a veces ‘sometimes’, por favor ‘please’) and plural forms (olas ‘waves’). At level 1, the model generates multi-word expres-

sions (manzana roja ‘red apple’, manzana amarilla ‘yellow apple’, manzana verde ‘green apple’), interjections (hola, gracias) and multi-word expressions (por favor). At level 2, the model generates interjections (hola, gracias). At level 3, the model generates personal pronouns (nosotros ‘us’), plural forms (olas ‘waves’, mesas ‘tables’, pájaros ‘birds’), conjugated verb forms (llegaron ‘they arrived’, llegaste ‘you arrived’, llego ‘(I) arrive’). At level 4, the model generates multi-word expressions (nave espacial ‘spacecraft’, cambio climático ‘climate change’, historia antigua ‘old history’, jardín botánico ‘botanical garden’, naturaleza muerta ‘still life’).

#### 4.4.3 French

**Gemma-2B** At level 0, the model generates feminine forms (grande ‘tall-FEM’, petite ‘small-FEM’), interjections (oui ‘yes’), plural forms (amis ‘friends’) and multi-word expressions (merci beaucoup ‘thank you very much’). At level 1, the model generates *grande* as on the previous level. At level 2, the model generates plausible words. At level 3, the model generates *oui* as on level 0. At level 4, the model generates feminine forms (ambigüe ‘ambiguous-FEM’) and apparently English words (incoherence, discreet).

**GPT-4o** At level 0, the model generates interjections (excusez-moi ‘excuse me’, oui ‘yes’), multi-word expressions (s’il vous plaît ‘please’, au revoir ‘goodbye’), feminine nouns (amie ‘friend-FEM’) but also slightly misspelled words (velo instead of vélo ‘bicycle’). At level 1, the model generates numbers, plural forms (amis ‘friends’), multi-word expressions (l’année \*derniere ‘last year’), but also words of lesser complexity (mois ‘month’). At level 2, there are no generated words not present in the reference word list. At level 3, the model generates multi-word expressions/reflexive verbs (se faufiler ‘sneak’). At level 4, the model generates plausible words, but possibly of too high complexity (prestidigitacion ‘sleight of hand’, pugnacité ‘pugnacity’, malversation ‘embezzlement’, acquiescer ‘acquiesce’).

**Llama3-8B** At level 0, the model generates mostly multi-word expressions and phrases or phrasal fragments (je suis impatient ‘I am impatient’, je voudrais ‘I would like’, c’est faux ‘that’s wrong’), but also some questionable phrases such as *ça est irraisonable*, which should be *c’est irraisonable*. At level 1, the model again mostly generates phrases, and again some questionable phrases such as *je suis*

*frère/femme* ‘I am brother/woman’. At level 2, the model generates plausible words and multi-word expressions (*réservation de taxi* ‘taxi reservation’, *transport en commun* ‘public transport’). At level 3, the model generates plausible words and plural forms, although these are generally encountered in the plural (*chaussures* ‘shoes’, *souliers* ‘shoes’, *épices* ‘spices’). At level 4, the model generates some questionable English words of high complexity as *mantic*, *catharsis*, and *kibosh*.

**Mistral-7B** At level 0, the model generates personal pronouns (*eux* ‘them’, *elle* ‘she’), multi-word expressions (*pommes frites* ‘French fries’), but also some clearly non-French words (*beef*, *chicken*, *vino*). At level 1, the model generates conjugated verb forms (*parlait* ‘(s/he) spoke’), plural forms (*doigts* ‘fingers’), multi-word expressions (*au revoir* ‘goodbye’), and some questionable or wrong forms such as *s’lever* (possible in slang but generally *se lever*), *ecouter* (*écouter*), *cafe* (*café*). At level 2, the model generates plural forms (*chiens* ‘dogs’) and some questionable words such as *\*prenon*, *coche* ‘car-SPANISH’, *milk*, *banana*, *egg*, *water*. At level 3, the model generates feminine forms (*délicieuse* ‘delicious-FEM’) and English words (*negociate*). At level 4, the model generates multi-word expressions (*une fois de plus* ‘once more’, *penser qu’il est possible* ‘think that it is possible’, *selon une étude* ‘according to a study’), feminine forms (*contemporaine* ‘contemporary-FEM’), English words (*idiosyncrasy*), and questionable so-called “multi-word expressions” (*trouver des choux de bruxelles sous les pierres* ‘finding brussels sprouts under stones’, *donner sa bague à quiconque veut l’attraper* ‘giving your ring to anyone who wants to grab it’, *s’asseoir sur \*un \*chais de poule* ‘sitting on a chicken chair(?)’).

**Phi3-4K** At level 0, the model generates multi-word expressions (*très bien* ‘very good’, *je n’ai pas* ‘I don’t have’, *je suis* ‘I am’, *je ne comprends pas* ‘I don’t understand’, *pas mal* ‘not bad’). At level 1, the model generates male/female alternatives (*apprenti(e)* ‘apprentice’, *professeur(e)* ‘professor’, *\*enfant(e)*<sup>6</sup>), multi-word expressions (*je suis* ‘I am’, *très bien*, *merci* ‘very good, thank you’, *je m’appelle* ‘my name is’, *s’il vous plaît* ‘please’). At level 2, the model generates multi-word expressions (*j’ai besoin de* ‘I need’, *un peu* ‘a bit’, *je voudrais* ‘I would like’). At level 3, the model gen-

<sup>6</sup>*Enfant* as a noun can take both the male and female article. *Enfante* exists as a conjugated form of *enfanter* ‘to give birth/bear fruit/bear a child’

erates plural forms (*conséquences* ‘consequences’, *héros* ‘heroes’), multi-word expressions (*justice sociale* ‘social justice’, *liberté individuelle* ‘individual liberty’), and English words (*warrant*). At level 4, the model generates plausible words.

#### 4.4.4 Swedish

**Gemma-2B** At level 0, the model generates superlative adjectives (*bästa* ‘best’, *högsta* ‘highest’), conjugated verb forms (*kom* ‘come-IMP/came’), alternatives separated by slash (*ja/nej* ‘yes/no’). At level 1, the model generates more alternatives separated by slash (*goddag/godnatt* ‘good day/night’, *skapar/tar* ‘create/take’, *jag/du/han/hon* ‘I/you/he/she’). At level 2, the model generates personal pronouns (*du* ‘you’, *vi* ‘we’), words of higher complexity (*semantik* ‘semantics’, *multipl* ‘multiple’, *konnotation* ‘connotation’) and conjunctions (*som* ‘as’). At level 3, the model generates non-Swedish words (*fyllek*, *inkluder*, *konsekvent*), fragments (*effektivitets*, *sammanfatt*) and plural forms (*distraktioner* ‘distractions’, *konditioner* ‘conditions’, *konflikter* ‘conflicts’). At level 4, the model generates plausible words.

**GPT-4o** At level 0, the model generates fragments (*gat*), interjections (*hej* ‘hello’), but also some valid forms that are simply not in the reference word list (*snart* ‘soon’, *idag* ‘today’, *snälla* ‘please’). At level 1, the model generates plural forms that are mostly encountered in the plural (*skor* ‘shoes’, *grönsaker* ‘vegetables’, *pengar* ‘money’, *byxor* ‘pants’). At level 2, the model generates valid forms that are not present in the reference word list (*plommon* ‘plum’, *citron* ‘lemon’, *fjärrkontroll* ‘remote control’, *körsbär* ‘cherry’, *fikon* ‘fig’). At levels 3 and 4, the model generates plausible words.

**Llama3-8B** At level 0, the model generates personal pronouns (*hon* ‘she’, *ni* ‘you-PL’), conjunctions (*om* ‘if’), and multi-word expressions (*du kan* ‘you can’, *ni är* ‘you-PL are’, *vi har* ‘we have’). At level 1, the model generates plural forms (*frukter* ‘fruits’, *händer* ‘hands’, *fötter* ‘feet’), genitive forms (*husdjurs* ‘of the pet(s)’), and non-Swedish words (*fartyk*, probably meant to be *fartyg* ‘vehicle’). At level 2, the model generates plural forms (*kängor* ‘boots’, *tänder* ‘teeth’, *fingerar* ‘fingers’) and definite forms (*landet* ‘the land’). At level 3, the model generates plausible words. At level 4, the model generates plausible words but also quite some plural/definite/genitive forms.

**Mistral-7B** At level 0, the model generates non-Swedish forms (*ananass*, *kokka*, *ingokt*), geni-

tive forms (köks ‘of the kitchen’), numbers, interjections (hej ‘hello’), personal pronouns (du ‘you’), and some questionable forms such as *man-nis(ka)* and *kvinn(a)* that cannot be decomposed as indicated in Swedish. The first word should be *människa* ‘human’, there is no such word as *männis*, and the second word should be *kvinna* ‘woman’, again there is no such word as *kvinn*. At level 1, the model generates plural forms (skor ‘shoes’, pengar ‘money’, kakor ‘cookies’, byxor ‘pants’), definite plural forms (äpplen ‘the apples’), and non-Swedish words (fräj). At level 2, the model generates clearly English words (autumn, january, march, august, winter), and the number one-hundred-eleven (hundraettioett). At level 3, the model generates non-Swedish words (hedervidy) and some misspelled words (heteronym, ockupation, perssonlighet). At level 4, the model generates plausible words.

**Phi3-4K** At level 0, the model generates noun phrases with articles (en liten flicka ‘a small girl’, en liten hund ‘a small dog’, en liten fisk ‘a small fish’, \*en liten hus ‘a small house’), multi-word expressions (jag har ‘I have’), definite forms (katten ‘the cat’), interjections (hej ‘hello’), articles (det ‘the’), comparative adjective forms (äldre ‘older’). At level 1, the model generates personal pronouns (du ‘you’, hon ‘she’, han ‘he’), nouns with article (en bilspår ‘a car track’, en bil ‘a car’), question particles (hur ‘how’) and small phrases (du/han/hon/jag/det är ‘you/he/she/I/it is/are’). At level 2, the model generates small phrases with *låt oss* ‘let’s’ (träffa ‘meet’, spela ‘play’, ...). At level 3, the model generates definite forms (skolan ‘the school’, dörren ‘the door’, gatan ‘the street’). At level 4, the model generates multi-word expressions (framtidens utveckling ‘future development’, \*kulturella identitet ‘cultural identity’).

#### 4.4.5 Dutch

**Gemma-2B** At level 0, the model generates articles (het ‘the’) and personal pronouns (ik ‘I’, jullie ‘you’, hij/zij ‘he/she’). At level 1, the model generates questionable words (\*esensieel, contextueel ‘contextual’, opwinding ‘excitement’, verenigt ‘united’, \*genuinen, opdrachten ‘commands’, \*overschrokken, onvoorspelbaar ‘unpredictable’). At level 2, the model also generates questionable words (oplossingen ‘solutions’, vervuld ‘fulfilled’, opwinding ‘excitement’, transformatie ‘transformation’, verhoogd ‘elevated’, liberaliseren ‘liberalize’, opvolging ‘succession’, mul-

tidimensionaal ‘multidimensional’) and English words (delicate, aromatic). At level 3, the model generates questionable words (\*exceptieel). At level 4, the model generates plausible words.

**GPT-4o** At level 0, the model generates days of the week (woensdag ‘Wednesday’, vrijdag ‘Friday’) and multi-word expressions (dank je ‘thank you’). At level 1 and 2, the model generates words with the diminutive *-je* ending (broodje ‘bread-DIM’, koekje ‘cake-DIM/cookie’). At level 3 and 4, the model generates plausible words.

**Llama3-8B** At level 0, the model generates questionable words related to games (spelletjeskistje ‘game box’, speelkaart ‘playing card’, spelletjesdoos ‘game box’, spelletjesbox ‘game box’, spelletje ‘game’, spelletjespak ‘game pack’) and words with the diminutive *-je* ending (hondje ‘dog-DIM’, huisje ‘house-DIM’, katje ‘cat-DIM’, autootje ‘car-DIM’). At level 1, the model generates plausible words, but also days of the week, numbers, multi-word expressions and diminutive expressions (broertje ‘brother-DIM’, zusje ‘sister-DIM’). At level 2, the model generates more diminutive forms (liedje ‘song-DIM’, broertje ‘brother-DIM’, koekje ‘cake-DIM/cookie’, muziekje ‘music-DIM’, broodje ‘bread-DIM’, pakketje ‘package-DIM’, zusje ‘sister-DIM’, spelletje ‘game-DIM’, briefje ‘letter-DIM’). At level 3 and 4, the model generates plausible words.

**Mistral-7B** At level 0, the model generates days of the week (donderdag ‘Thursday’, woensdag ‘Wednesday’), numbers (vier ‘four’, vijf ‘five’), months of the year (augustus ‘August’, oktober ‘October’), personal pronouns (jullie ‘you’, hij ‘he’) and multi-word expressions (hoe zoekt u? ‘how do you search?’, met vriendelijke groet ‘yours sincerely’). At level 1, the model generates diminutive forms (vierkantje ‘square-DIM’, bankje ‘bench-DIM’, tabletje ‘tablet-DIM’, hakje ‘heel-DIM’, bootje ‘boat-DIM’, klusje ‘chore-DIM’). At level 2, the model also generates diminutive forms (appeltje ‘apple-DIM’, dagje ‘day-DIM’). At level 3 and 4, the model generates plausible words.

**Phi3-4K** At level 0, the model generates diminutive forms (appeltje ‘apple-DIM’). At level 1, the model generates plural forms (dieren ‘animals’, rozen ‘roses’), superlative adjective forms (oudste ‘oldest’), personal pronouns (ik ‘I’), conjugated verb forms (eet<sup>7</sup> ‘eats/eat-IMP’) and days of the week (maandag ‘Monday’). At level 2, the model

<sup>7</sup>In Afrikaans, *eet* is the infinitive form of the verb ‘to eat’

generates plural forms (boodschappen ‘groceries’, vrienden ‘friends’, autos ‘cars’). At level 3, the model generates all days of the week and *kledingstukken* ‘garments’. At level 4, the model generates multi-word expressions (regionale economie ‘regional economy’, sociale kwesties ‘social issues’).

#### 4.4.6 General Remarks

Overall, we see a common pattern in the generated graded word lists, namely a propensity to generate personal pronouns (you, he, it), days of the week, months of the year, and numbers. All those categories were excluded from the reference word list based on part-of-speech filtering. A common motive also seems to be food and animals.

The models also tend to generate phrases rather than single words at times; phrases and multi-word expressions are undeniable useful for language learners, but the models do not adhere to the prompt.

In contrast to the grading task, which does not require models to output any language, the vocabulary generation tasks shows some shortcomings of the models when it comes to *producing* language other than English. This is noticeable for Spanish (Gemma-2B, Mistral-7B), French (GPT-4o, Llama3-8B, Mistral-7B, Phi3-4K), Swedish (Gemma-2B, Llama3-8B, Mistral-7B), and Dutch (Gemma-2B).

Finally, especially for English, all models generate words of the highest complexity when prompted for words of level 4. This may well be a phrasing problem in the prompt, as we explicitly state 4 as the *highest* level, albeit for language learners.

One general problem that we noticed is that if the word to assess is (or could be interpreted as) an English word, apparently mostly related to computer programming (by, blank, score, index, column, sample, type), the model fails to recognize the word to assess. We also notice that sometimes the models score outside of the given range (5,6,7), repeats the input prompt, or generates additional explanations even though it was asked not to. This is especially true for Gemma-7B.

## 5 Conclusion

In this paper, we presented experiments of using small versions of large generative language models out-of-the-box for (1) grading vocabulary lists and (2) generating graded vocabulary lists. Results show that while most of the models may only be

targeted at English, they perform quite well cross-linguistically at the task of *grading* vocabulary. However, when it comes to *producing* vocabulary, the quality suffers.

One key finding is that GLMs that perform well on the task of grading vocabulary can be used to grade vocabulary items with low observed frequency. This use case uses the strength of graded word lists and GLMs for synergy effects.

We have also shown that using a numerical scale rather than the CEFR scale yields better results. This may be because the language models have not had enough contact with CEFR material to learn and “understand” what the different levels mean. A numerical scale may be more generalizable in this case.

To answer the hypothesis put forward at the beginning of the paper: “GLMs are effective at grading vocabulary”, we can conclude that all tested models exhibited some form of grading ability, although the predicted scores do not exactly match the textbook-derived scores, leading to low accuracy, precision and recall. However, when taking into account adjacent accuracy (the prediction is considered correct if it is at most one level from the target level), we can see values up to 99% (see Table 7 in the Appendix A).

When it comes to generating vocabulary from scratch, GLMs can be a starting point, although their potential for generating large graded vocabulary lists seem limited and needs further investigation. The inclusion of inflected forms (plural forms, conjugated verb forms) is undesired for most purposes.

One (maybe unsurprising) finding is that the much larger base-model GPT-4o performed best on average, indicating the larger GLMs may be more accurate in grading and generating vocabulary lists. However, Mistral-7B showed promising performance at second place and thus might be a viable free option.

Overall, while generative language models show promise in grading vocabulary across languages, continued research and development are needed to enhance their performance and applicability in language learning contexts.

In the hopes that the data may be of use to other researchers in the field, we make the data available at [https://github.com/daalft/cefrlex\\_llm](https://github.com/daalft/cefrlex_llm).



## 6 Future Work

We noticed that all models show a general tendency towards the middle levels. Previous research on feature-based classifiers shows that these classifiers tend to perform well on the extremes of the scale, and tend to mix up the middle levels (Pilán et al., 2016; Alfter and Volodina, 2018). Hence, we could potentially use feature-based classifiers to confidently identify items at the extremes of the scale, and GLMs to classify the middle levels.

Prompt engineering would also be a possible avenue for future work. A chain-of-thought prompt as used by Enomoto et al. (2024) may be more effective at eliciting not only a grade but also the decision process for arriving at that grade, allowing for greater transparency. As LLMs are sensitive to prompt formulation (Sclar et al., 2024), experimenting with different prompt wording may also prove beneficial.

Finally, it would be interesting to investigate how fine-tuning models impacts performance. We suggest a scenario where fine-tuning is done on one language family (e.g., Romance) and tested on a different language family (e.g., Germanic), to check for language-agnostic transferability of the graded vocabulary concept.

## Limitations

In this work, we investigate only European languages, giving the work a strong Eurocentric focus. It would be beneficial to extend the investigation to more non-European languages.

In this work, we only tested small models. It is highly possible that the larger models may yield better results. However, such models also require significantly more power, both computational and financial.

Finally, we only generate up to 100 words for each level for each language. The generation limit of the GLM can be circumvented through a chat with history by repeatedly asking for *another* set of 100 words and passing the previously generated answer as history. Alternatively, the LLM can be prompted to generate *texts* of a certain proficiency level, based on which frequency-level information about words can be extracted, simulating a learner-oriented textbook (comprehension) corpus.

## References

- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacrose, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. [Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone](#). Preprint, arXiv:2404.14219.
- David Alfter. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. Ph.D. thesis, University of Gothenburg, Sweden.
- David Alfter, Lars Borin, Ildikó Pilán, Therese Lindström Tiedemann, and Elena Volodina. 2019. Lärka: From Language Learning Platform to Infrastructure for Research on Language Learning. In *Selected papers from the CLARIN Annual Conference 2018*, pages 1–14. Linköping University Electronic Press.
- David Alfter and Johannes Graën. 2019. Interconnecting lexical resources and word alignment: How do learners get on with particle verbs? In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 321–326.
- David Alfter and Elena Volodina. 2018. Towards Single Word Lexical Complexity Prediction. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88.
- Yigal Attali, Andrew Runge, Geoffrey T LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A Von Davier. 2022. The interactive reading task:

- Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5:903077.
- Drilon Avdiu, Vanessa Bui, Klára Ptacinová Klimci, et al. 2019. Predicting learner knowledge of individual words using machine learning. In *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019)*, September 30, Turku Finland, 164, pages 1–9. Linköping University Electronic Press.
- Hamid Ayub. 2024. GPT-4o: Successor of GPT-4? <https://hamidayub.medium.com/gpt-4o-successor-of-gpt-4-8207acf9104e>. Accessed: June 12, 2024.
- Anthony Baez and Horacio Saggion. 2023. **LSLlama: Fine-tuned LLaMA for lexical simplification**. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 102–108, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Stefano Bannò, Hari Krishna Vydana, Kate M Knill, and Mark JF Gales. 2024. Can GPT-4 do L2 analytic assessment? *arXiv preprint arXiv:2404.18557*.
- Ummugul Bezirhan and Matthias von Davier. 2023. **Automated Reading Passage Generation with OpenAI’s Large Language Model**. *Computers and Education: Artificial Intelligence*, 5:100161.
- Ummugul Bezirhan and Matthias von Davier. 2023. Automated reading passage generation with OpenAI’s large language model. *Computers and Education: Artificial Intelligence*, 5:100161.
- B Boras, E Smolić, G Gledec, and T Jagušt. 2024. Exploring the Educational Potential of Generative AI: An Application for Spelling Practice. In *INTED2024 Proceedings*, pages 6945–6952. IATED.
- Vaclav Brezina and Dana Gablasova. 2015. Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, 36(1):1–22.
- Annette Capel. 2015. The English Vocabulary Profile. In Julia Harrison and Fiona Barker, editors, *English Profile in Practice*, chapter 2, pages 9–27. Cambridge University Press.
- Council of Europe. 2018. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors. Accessed 09.03.2019 from [www.coe.int/lang-cefr](http://www.coe.int/lang-cefr).
- Averil Coxhead. 1998. *An academic word list*, volume 18. School of Linguistics and Applied Language Studies.
- Averil Coxhead. 2011. The academic word list 10 years on: Research and teaching implications. *Tesol Quarterly*, 45(2):355–362.
- Edwin Creely. 2024. Exploring the Role of Generative AI in Enhancing Language Learning: Opportunities and Challenges. *International Journal of Changes in Education*.
- Luise Dürlich and Thomas François. 2018. EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. In *11th International Conference on Language Resources and Evaluation (LREC 2018)*.
- Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2018. Mining Words in the Minds of Second Language Learners for Learner-specific Word Difficulty. *Journal of Information Processing*, 26:267–275.
- Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. TMU-HIT at MLSP 2024: How Well Can GPT-4 Tackle Multilingual Lexical Simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598.
- Chloe Farr. 2024. *Unmasking ChatGPT: The Challenges of Using Artificial Intelligence for Learning Vocabulary in English as an Additional Language*. Ph.D. thesis.
- Thomas François and Barbara De Cock. 2018. ELELex: a CEFR-graded lexical resource for Spanish as a foreign language. In *PLIN Linguistic Day 2018: Technological innovation in language learning and teaching*.
- Thomas François, Núria Gala, Patrick Watrin, and Cédric Fairon. 2014. FLELex: a graded lexical resource for French foreign learners. In *LREC*, pages 3766–3773.
- Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded Lexical Resource for Swedish Foreign and Second Language Learners. In *LREC*.
- Núria Gala, Thomas François, and Cédric Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *E-lexicography in the 21st century: thinking outside the paper*, Tallin, Estonia.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy,

- Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Milligan, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. [Gemma: Open Models Based on Gemini Research and Technology](#). *Preprint*, arXiv:2403.08295.
- Robert Godwin-Jones. 2024. [Distributed agency in second language learning and teaching through generative AI](#). *Preprint*, arXiv:2403.20216.
- Johannes Graën, David Alfter, and Gerold Schneider. 2020. [Using Multilingual Resources to Evaluate CE-FRLEX for Learner Applications](#). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 346–355, Marseille, France. European Language Resources Association.
- Daniel R Isbell. 2017. Assessing C2 writing ability on the Certificate of English Language Proficiency: Rater and examinee age effects. *Assessing Writing*, 34:37–49.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7B](#). *Preprint*, arXiv:2310.06825.
- Abdelhak Kelious, Matthieu Constant, and Christophe Coeur. 2024. Complex Word Identification: A Comparative Study between ChatGPT and a Dedicated Model for This Task. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3645–3653.
- Reza Kianian, Deyu Sun, Eric L. Crowell, and Edmund Tsui. 2024. [The Use of Large Language Models to Generate Education Materials about Uveitis](#). *Ophthalmology Retina*, 8(2):195–201.
- Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavriliidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48(1):121–163.
- Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99:101861.
- Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022. [Multitasking Framework for Unsupervised Simple Definition Generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5934–5943, Dublin, Ireland. Association for Computational Linguistics.
- Richard LaBontee. 2019. *Strategic Vocabulary Learning in the Swedish Second Language Context*. Ph.D. thesis, University of Gothenburg.
- Batia Laufer. 2006. Comparing focus on form and focus on forms in second-language vocabulary learning. *Canadian Modern Language Review*, 63(1):149–166.
- John Lee and Chak Yan Yeung. 2018. Automatic prediction of vocabulary knowledge for learners of Chinese as a foreign language. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–4. IEEE.
- Joanne Leong, Pat Pataranutaporn, Valdemar Danry, Florian Perteneder, Yaoli Mao, and Pattie Maes. 2024. Putting Things into Context: Generative AI-Enabled Context Personalization for Vocabulary Learning Improves Learning Motivation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–15.
- Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.
- Mounica Maddela and Wei Xu. 2018. A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760.
- I. Miralpeix and C. Muñoz. 2018. [Receptive vocabulary size and its relationship to EFL language skills](#). *IRAL-International Review of Applied Linguistics in Language Teaching*, 56:1–24.
- Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023a. Deep learning approaches to lexical simplification: A survey. *arXiv preprint arXiv:2305.12000*.
- Kai North, Marcos Zampieri, and Matthew Shardlow. 2023b. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeesh Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. *GPT-4 Technical Report*. Preprint, arXiv:2303.08774.
- Jenny A Ortiz-Zambrano and Arturo Montejo-Ráez. 2020. Overview of ALexS 2020: First Workshop on Lexical Analysis at SEPLN.
- Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *SemEval at NAACL-HLT*, pages 560–569.
- Pearson. 2017. GSE Teacher Toolkit. <https://www.english.com/gse/teacher-toolkit/user/vocabulary>. Accessed: 2024-06-10.
- Zhenhui Peng, Xingbo Wang, Qiushi Han, Junkai Zhu, Xiaojuan Ma, and Huamin Qu. 2023. Storyfier: Exploring Vocabulary Learning Support with Text Generation Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–16.
- Ivica Pesovski, Ricardo Santos, Roberto Henriques, and Vladimir Trajkovik. 2024. *Generative ai for customizable learning experiences*. *Sustainability*, 16(7).
- Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111.
- Ildikó Pilán, David Alfter, and Elena Volodina. 2016. Coursebook texts as a helping hand for classifying linguistic complexity in language learners' writings. In *Proceedings of the workshop on Computational Linguistics for Linguistic Complexity (CLALC)*. COLING 2016. Osaka, Japan.

- Adrielli Rego Lopes, Joshua Snell, and Martijn Meeter. 2024. [Language Models Outperform Cloze Predictability in a Cognitive Model of Reading](#).
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting](#). *Preprint*, arXiv:2310.11324.
- Matthew Shardlow. 2013. A Comparison of Techniques to Automatically Identify Complex Words. In *51st annual meeting of the association for computational linguistics proceedings of the student research workshop*, pages 103–109.
- Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, and Anna Huelsing. 2024a. [The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.
- Matthew Shardlow, Kai North, and Marcos Zampieri. 2024b. A Multilingual Survey of Recent Lexical Complexity Prediction Resources through the Recommendations of the Complex 2.0 Framework. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context@ LREC-COLING 2024*, pages 51–59.
- Philip Ernest Springer. 2012. *Advanced learner writing: A corpus-based study of the discourse competence of Dutch writers of English in the light of the C1/C2 levels of the CEFR*. Ph.D. thesis, University of Amsterdam, Netherlands.
- Yao-Ting Sung, Wei-Chun Lin, Scott Benjamin Dyson, Kuo-En Chang, and Yu-Chia Chen. 2015. Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, 99(2):371–391.
- Anaïs Tack, Thomas François, Piet Desmet, and Cédric Fairon. 2018. NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 137–146.
- Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédric Fairon. 2016. Evaluating Lexical Simplification and Vocabulary Knowledge for Learners of French: Possibilities of Using the FLELex Resource. In *LREC*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [LLaMA: Open and Efficient Foundation Language Models](#). *Preprint*, arXiv:2302.13971.
- Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. 2023. Primacy Effect of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 108–115.
- Stuart Webb and Paul Nation. 2017. *How vocabulary is learned*. Oxford University Press.
- Rodrigo Wilkens, David Alfter, Xiaou Wang, Alice Pintard, Anaïs Tack, Kevin P Yancey, and Thomas François. 2022. Fabra: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233.
- Kevin Yancey and Yves Lepage. 2018. Korean L2 Vocabulary Prediction: Can a Large Annotated Corpus be Used to Train Better Models for Predicting Unknown Words? In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.
- Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, United States. Association for Computational Linguistics.
- Jiaxin Yuan, Cunliang Kong, Chenhui Xie, Liner Yang, and Erhong Yang. 2022. [COMPILING: A Benchmark Dataset for Chinese Complexity Controllable Definition Generation](#). In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 921–931, Nanchang, China. Chinese Information Processing Society of China.
- Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024. [GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection](#). *Preprint*, arXiv:2403.03507.
- Imran Zuolkernan and Salsabeel Shapsough. 2024. Towards Using Large Language Models to Automatically Generate Reading Comprehension Assessments for Early Grade Reading Assessment. In *INTED2024 Proceedings*, pages 3772–3782. IATED.
- Anika Milena Żerkowska. 2024. Personalized Language Learning in the Age of AI: Leveraging Large Language Models for Optimal Learning Outcomes. Master's thesis.

## A Generating grades: Full result table

	Gemma-2B					Gemma-7B					GPT-4o				
	Acc	AAcc	P	R	F1	Acc	AAcc	P	R	F1	Acc	AAcc	P	R	F1
EFLLex	0.20	0.84	0.21	0.20	0.18	0.24	0.93	0.32	0.24	0.18	<b>0.30</b>	<b>0.94</b>	<b>0.39</b>	<b>0.30</b>	<b>0.29</b>
ELELex	0.18	0.77	0.23	0.18	0.16	0.23	0.93	0.30	0.23	0.19	0.25	0.91	<b>0.33</b>	0.26	0.25
FLELex	0.21	0.85	0.36	0.22	0.15	0.17	0.89	0.39	0.18	0.12	0.21	0.91	0.36	0.22	0.22
SVALex	0.26	0.94	0.25	0.27	0.24	0.30	<b>0.96</b>	0.31	0.30	0.21	<b>0.33</b>	<b>0.96</b>	<b>0.34</b>	<b>0.34</b>	<b>0.33</b>
NT2Lex	0.24	0.90	0.30	0.24	0.18	0.27	0.97	0.28	0.27	0.19	0.33	0.97	0.39	0.34	0.35
	LLaMA3-8B					Mistral-7B					Phi3-4K				
	Acc	AAcc	P	R	F1	Acc	AAcc	P	R	F1	Acc	AAcc	P	R	F1
EFLLex	0.21	0.93	0.35	0.22	0.16	0.26	0.87	0.35	0.27	0.24	0.22	0.91	0.33	0.22	0.15
ELELex	0.24	<b>0.95</b>	0.29	0.24	0.20	<b>0.30</b>	0.90	0.31	<b>0.30</b>	<b>0.28</b>	0.25	0.94	0.35	0.25	0.19
FLELex	0.25	<b>0.96</b>	0.40	0.26	0.19	0.28	0.92	0.39	0.28	<b>0.28</b>	<b>0.29</b>	<b>0.96</b>	<b>0.42</b>	<b>0.29</b>	0.22
SVALex	0.28	<b>0.96</b>	0.31	0.29	0.25	0.21	0.87	0.28	0.21	0.18	0.26	0.95	0.28	0.26	0.20
NT2Lex	0.37	<b>0.99</b>	0.38	0.38	0.35	0.37	0.97	0.36	0.37	0.36	<b>0.40</b>	<b>0.99</b>	<b>0.40</b>	<b>0.40</b>	<b>0.38</b>

Table 7: Results in terms of Accuracy (Acc), Adjacent accuracy (AAcc), Precision (P), Recall (R), F1 score (F1), all weighted by label. Results in bold indicate the best result per category (Acc, AAcc, P, R, F1)

## B Model parameters

Model parameters for generation. For the Gemma models and GPT-4o, no additional parameters were passed. For Mistral-7B and Llama3-8B, sampling was enabled, for Llama3-8B the temperature and top\_p parameters were set, and for Phi-3, temperature was explicitly set to zero. The example code for Phi-3 additionally includes do\_sample=False, which has no effect when temperature is zero, thus we excluded this parameter.

Model	Generation parameters
Gemma-2B	None
Gemma-7B	None
Mistral-7B	do_sample=True
Llama3-8B	do_sample=True, temperature=0.6, top_p=0.9
Phi-3	temperature= 0.0
GPT-4o	None

Table 8: Model generation parameters

# Investigating Acoustic Correlates of Whisper Scoring for L2 Speech Using Forced Alignment with the Italian Component of the ISLE corpus

Nicolas Ballier

LLF & CLILLAC-ARP

Université Paris Cité

rue Thomas Mann

75013 PARIS, FRANCE

nicolas.ballier@u-paris.fr

Adrien Méli

CLILLAC-ARP

Université Paris Cité

rue Thomas Mann

75013 PARIS, FRANCE

adrienmeli@gmail.com

## Abstract

Automatic Speech Recognition (ASR) can be used to analyse L2 speech but researchers cannot be sure that the ASR transcriptions accurately represent learner speech. We aim to confront the ASR outputs with the acoustic analysis of learner speech. Whisper (Radford, 2023) provides transcriptions and probabilities associated to the predicted transcriptions. This paper analyses how global phonetic analyses of learner data can be used to potentially confirm these Whisper probability scores assigned to learner transcriptions. We tested the Italian component of the ISLE corpus with phonetic analyses of 23 learners of English. We compared the levels assigned to these speakers by the corpus experts to the outputs of Whisper’s `tiny` model. We discuss the phonetic features that may account for these Whisper predictions using acoustic data extracted from forced alignment. We try to correlate the levels assigned to the speakers in the ISLE corpus with the quality of the phonetic realisation, using global vocalic measurements such as the convex hull or Euclidian distances between monophthongs. We show that Levenshtein distance to the reference transcription of the Whisper `tiny` model (measured using Levenshtein distance to the read text) correlates with the grades assigned by the annotators.

## 1 Introduction

Learner speech has mostly been recently researched with Automatic Speech Recognition (ASR) system and the focus has been on phone substitution (Chanethom and Henderson, 2023). These analyses presuppose time-consuming manual checking of the transcriptions against the recordings. We would like to explore acoustic correlates of ASR transcriptions and investigate

whether phonetic data extracted from the transcriptions could be used to confirm the ASR diagnoses. Our Research question is thus: ‘Can Whisper’s automatically generated transcriptions be used to assess a non-native speaker’s pronunciation?’ OpenAI’s Whisper (Radford et al., 2023) generates time-stamped transcriptions of recorded speech from simple audio files. When mapping the signal to the best candidates for transcription, Whisper ascribes a probability score to each subtoken, which evaluates the likelihood that the transcription that was selected is correct. With non-native speakers, one potential issue is that mispronunciations, especially when systematic or when pertaining to phonemic sequences with dense phonological neighbourhoods, may lead to transcription errors in spite of high probability scores. The purpose of this study is to find out whether vocalic analyses based on force-aligning Whisper’s transcriptions provide reliable, usable acoustic information about speakers’ characteristics in pronouncing English; 1) to find out whether Whisper’s scores correlate with speakers’ proficiencies in pronouncing English; 2) to find out whether vocalic data collected from force-aligning Whisper’s transcriptions provides reliable information regarding the speakers’ performances.

We focus on vowels as they are notoriously difficult (Ballier and Martin, 2015) for learners. We explore several holistic representations of vowels: the acoustic (F1 and F2) formants, the global vowel trapezium plots and the corresponding convex hull as they are likely to be indicative of any actual phonological or phonetic phenomena underlying non-native speakers’ pronunciations. Using the recordings of 23 Italian speakers from the ISLE corpus, this study investigates the linguistic significance of Whisper’s probability scores, *i.e.*, whether they are indicative of the non-native speakers’ proficiencies in pronouncing English. It also explores whether vocalic analyses based

---

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.



on force-aligning Whisper’s transcriptions provide reliable, usable information about the speakers’ performances.

Whisper (Radford et al., 2023) is a multilingual audio model trained to do language detection, voice activity detection, transcription translation, and up to a point, diarisation. It was trained on 680k hours of labelled speech data and reports state-of-the-art performance for transcription (Radford et al., 2023). Among these functionalities, the language detection task has not really been used for second language acquisition analysis. The analysis of the probability assigned to the sub-token predicted is still in its early stage (Ballier et al., accepted(a),a). With the C++ implementation of Whisper, we produced the transcriptions and the probability assigned to the sub-tokens. When accessing the internal representations of Whisper like the probability, linguists do not deal with tokens but with subtokens, which are the results of a byte-pair-encoding process designed to eliminate out-of-vocabulary tokens (Sennrich et al., 2016). This very sub-tokenisation also varies across models, even though Whisper uses the same dictionary of sub-tokens for the different models.

We focused on the Italian component of the ISLE corpus, because the level of the corpus is not homogeneous between the Italian component and the German one. The ISLE corpus derives from a European project aiming at analysing non-native speech, notably English spoken by German and Italian learners. The quality of the English spoken by each speaker was graded (from 1 to 5) and the raters reach a good agreement (Atwell et al., 2003).

The remaining sections of the paper are organised as follows: Section 2 mentions previous research, Section 3 outlines our method, Section 4 delves into our results, and Section 5 provides a discussion of these results.

## 2 Background Research

Automatic speech recognition (ASR) has been frequently used for the automatic analysis of learner speech (Dalby et al., 1998; Inceoglu et al., 2020; Tejedor-García et al., 2021; Ando and Zhang, 2005), compared to audio Large Language Models (LLMs). The number of papers using Whisper for the investigation of L2 speech is, for the time being, limited, but previous research suggests

that the probability assigned to the sub-token can be used as a proxy for the prediction of the levels of the learners (Ballier et al., 2023). Speech recognition is typically used to compute deviations from reference texts in read speech and investigate phone substitutions (McCrocklin et al., 2019; McCrocklin and Edalatishams, 2020; Chanethom and Henderson, 2023). An important contribution is the paper that uses the Otter system to try to measure the shortcomings of the models in relation to the vowel system on a very limited set constraints (Chan et al., 2022). And using the ISLE corpus data, (Arora et al., 2018) try to interpret the mis-transcriptions in terms of phonological features, thus focusing more on consonants.

## 3 Material and methods

In this section, we describe the ISLE corpus data and the pipeline utilised to annotate the data and the main phonetic representations. We analyzed the convex hull representing the trapezium of vowels, the number of vertices produced by the vowel trapezium representation, and then we present the Whisper output.

### 3.1 The ISLE Corpus Data

The corpus was collected to analyse non-native speech and is available from ELRA. The sections of the ISLE corpus correspond to phonological targets that were tested, with the exception of the read speech task (block A) which contained them all. We re-organised the ELRA data compiled in 1999 in a unique dataset gathering metadata, prompts, objectives and expert annotations. Table 1 illustrates the types of prompts that learners had to read.

The material used in this study comes from the ISLE corpus (Menzel et al., 2000). The recordings of 23 Italian speakers reading 180 blocks of text were analyzed in the fashion described in the following paragraph. The ISLE corpus is particularly interesting to study as it provides standardised recordings of a sizeable sample of speakers, whose performances were evaluated by trained annotators. These features make it possible to obtain two baselines, the script to read and the human evaluations, against which Whisper’s performances can be compared.

Block	# Sents.	Linguistic Issue	Exercise Type	Examples
A B C	27 33 22	Wide vocabulary coverage (410)	Adaptation/ Reading	“In 1952 a Swiss expedition was sent and two of the men reached a point only three hundred metres from the top before they had to turn back.”
D	81	Problem phones Weak Forms	Minimal Pair Item selection/ combination	“I said bad not bed.” “She’s wearing a brown wooly hat and a red scarf.”
E	63	Stress Weak Forms Problem Phones Consonant clusters	Reading	“The convict expressed anger at the sentence.” “The jury took two days to convict him.”
F	10	Weak Forms Problem Phones	Description/ Item selection/ combination	“I would like chicken with fried potatoes, broccoli, peas and a glass of water.”
G	11	Weak Forms Problem Phones	Item selection/ Combination	“This year I’d like to visit Rome.

Table 1: Typology of prompts in the ISLE data (after Menzel et al., 2000)

### 3.2 Whisper outputs

We have used the C++ implementation (Gerganov, 2023) of Whisper and the `tiny` model, more likely to be sensitive to non-native deviations from the training model realisations (Ballier et al., 2023). Whisper transcribes speech and the C++ implementation also allows researchers to extract the probability assigned by the Whisper model to the predicted subtokens. Figure 1 gives an example of the probabilities assigned to the predicted subtoken. The lowest probability score here corresponds to a mispronunciation of learner #134 who realises “*weather*” with a long vowel [wi:]. As this example shows, “*weather*” is transcribed as “*weeder*” in the transcription but corresponds internally to two subtokens (*we—eder*) in the Whisper representations. It is very difficult to re-align subtokens (*we—eder*) to tokens transcribed by Whisper (*weeder*) and to map these outputs to the reference (“*weather*”), so that we did not exploit probabilities at the subtoken level but only globally. When modelling data, we only considered the mean value of Whisper’s probability scores as a unique datapoint per speaker).

### 3.3 Whisper Scoring

We extracted the probability assigned to each subtoken and to the language assigned by the lan-

[_TT_460]	0.747888
The	0.992373
second	0.995847
difficulty	0.996018
about	0.956371
climbing	0.998327
Everest	0.962417
is	0.991093
the	0.986653
we	0.332225
eder	0.876064
.	0.970952

Figure 1: Example of the C++ Whisper output. The subtokens of the Whisper transcriptions are associated to a probability. `[_TT_460]` is a special subtoken corresponding to temporal value. The mistranscribed “*weeder*” (corresponding to “*weather*”) is split into two subtokens *we—eder*. The realisation of the first syllable by the learner is phonetically [wi].

guage prediction functionality. Whisper’s probability scores are generated in a file with a subtoken and a score per line. Subtokens often correspond to words, but are sometimes made up of syllables, silences, or punctuation marks such as commas or periods. “*Expedition*”, for instance, constitutes a token, but its plural, “*expeditions*”, is split into “*exped*” and “*itions*”, each with their

respective probability scores. Unfortunately, this feature makes it non-trivial to match the scores with the alignment, so that per-speaker probability scores were simply calculated by averaging over each token’s score. Figure 2 shows a visualisation of the different levels of probability assigned to the subtokens by the tiny model. A transcription like “wee—der” (corresponding to “*weather*”) shows low probabilities that are consistent with misrealisations of the vowel quality and of the interdental fricative.

### 3.4 Data Processing

For each speaker, the original short sound files were concatenated into a main audio file and input into Whisper, which in turn generated time-stamped `.srt` subtitles and a `.txt` file listing the probability scores for each token. The time-stamps from the subtitles were then used to split the main audio files into short ones. These short audio files and their matching Whisper transcription from the subtitles were fed into forced-aligner P2FA (Yuan and Liberman, 2008), which generated Praat (Boersma and Weenink, 2019) TextGrids with alignments at the segmental and lexical levels. The reason underlying this seemingly tedious procedure is the contention that feeding the forced aligner with short audio recordings will prevent cascading alignment errors. A syllabic tier and another segmental tier based on the British pronunciations of the *Longman Pronunciation Dictionary* (Wells, 2000) were also added. Finally, all these short TextGrids were merged into one main TextGrid. Vocalic data was then collected by parsing the LPD segmental tier of each speaker’s main TextGrid and storing relevant information, such as formant values and duration, when the segment was a vowel.

Figure 3 recaps the different alignments produced with our pipeline. We used the P2FA aligner to process the recordings. The aligner is fed with the CMU phonetic dictionary, one of the rare open source available for English, but which assumes an American pronunciation. We then used the PEASYV pipeline (Méli and Ballier, 2023) to generate the reports on the phone inventories of the different learners. Figure 7 sums up the vowel inventories corresponding to the transcriptions, with the proviso that some learners misread sentences or that some sentences for some speakers are not actually present in the ELRA data.

### 3.5 Evaluation Metrics

We wanted to correlate the mean probability scoring assigned by Whisper, the grades assigned by the annotators of the corpus (ranging from 1 to 5 for the 23 Italian speakers) and acoustic properties extracted from our forced alignment of the learner recordings with the Whisper transcription.

#### 3.5.1 Levenshtein distance

One metric instrumental to this study is the Levenshtein distance, which calculates the number of edits needed to change one string of characters into another. The systematic comparison of each speaker’s Whisper-generated transcription with the original ISLE script to read, provided by the designers of the corpus, was made after taking the following steps: the script to read was stripped of capital letters, blank spaces and punctuation marks. Measurements written in full letters were converted to numbers, in keeping with Whisper, which transcribes most numbers in Arabic. Subtleties such as “3 meters”, transcribed by Whisper in Arabic numbers, but “three mountains” transcribed in full letters, were accounted for. Each speaker’s Whisper-generated transcription underwent the same treatment: blank spaces and punctuation marks were removed, and capital letters were converted to lower-case.

#### 3.5.2 Main Acoustic Correlates

The next step was to determine whether correlations existed between the two baselines of the ISLE corpus, *i.e.* the annotators’ grades (from 1 to 4) and the Levenshtein distance to the original script (formatted in the way described in the previous section). In order to do so, several phonetic metrics for each speaker were computed with the formant values extracted at mid-temporal values from our forced alignment:

1. the Euclidean distances of each monophthong to all other 11 monophthongs in the F1/F2 space using mid-temporal values (66 datapoints per speaker);
2. the Vowel Inherent Spectral Change (Nearey and Assmann, 1986; Nearey, 2012; Morrison and Nearey, 2007; Morrison, 2012) of each vowel, *i.e.*, both monophthongs and diphthongs, using the mean formant values at 20% and 80% of the vowels’ durations (19 datapoints per speaker);

There are three main difficulties facing any part in attempting to climb Everest. The first of these difficulties is that of altitude. At great age the air is very thin. The air contains so little oxygen that climbers can only move very slowly. They also think more slowly and these make-outs then to make mistakes. There is one way of reducing this difficulty. If I climb Everest 6,000 meters or so and stays at the altitude for a few days, he will become a user to the fin air. This process of getting user to the altitude is called acclimatization. However, once the climber reaches an altitude of 8,000 meters, acclimatization is not longer possible. Instead of acclimatization, the body suffers damage. The musculus becomes smaller and loads their straight and within a few days the climber is no longer able to move. The summit of Everest is over 8,800 meters. It would be best to climb the last 2,500 meters in only one or two days but it is not possible. A second way of reducing the difficulty of altitude is for each climber to carry oxygen in a bottle on this bag. This oxygen equipment we knew from earlier expedition must be light in weight. The second difficulty about climbing Everest is the weeder.

Figure 2: Probability Scoring of Whisper’s Tiny model predictions for the subtokens of the transcription of (male) speaker 134. Purple corresponds to high probability, cyan to low probability

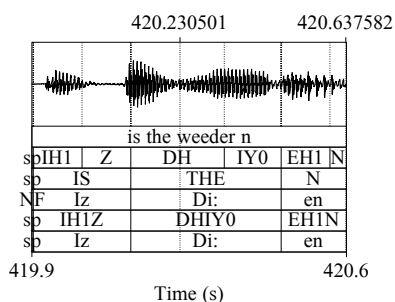


Figure 3: Fragment of a TextGrid corresponding to “is the weather n”. Under the waveform are the five tiers that correspond to the Whisper output transcript (“is the weeder n”), to the phoneme (CMU Arpabet transcription), to the words (“weeder” was missed by the aligner), to the British transcription (SAMPA), to the syllables of the CMU Arpabet transcription and to the British transcription (SAMPA) of the syllables.

3. the area of the speaker’s convex hull and its number of vertices (2 datapoints).

Pearson correlations with the Levenshtein distance and the annotators’ marks were then systematically computed.  $p$ -values above the 0.05 conventional threshold were rejected, along with absolute  $R$ -values inferior to 0.55, in order to exclude weak correlations.

### 3.5.3 Probability Density and Kernel Density Estimates (KDE), Convex Hulls and Number of Vertices

We wanted to test several global validation procedures based on acoustic correlates of vowels, investigating the convex hull and number of vertices as a representation of the trapezium produced by the different learners as compared to potential British models (the pronunciation norm indicated

for the ISLE data). We used the British pronunciation norm, reported as being the one used by the learners in the corpus (Atwell et al., 2003). We computed the convex hull and the number of vertices needed to represent the trapezium of vowels for the speakers. Figure 11 illustrates the trapezoids of the Italian male and female speakers, the vertices connecting the means of the F1/F2 values for vowels. The reference trapezium corresponds to the values reported in one of the reference studies for British English (Deterding, 1997). Because the formant extractions were based on lab speech (vowels in the /hVd/ context), these means correspond to hyperarticulated values. Our last attempt at exploiting the area of the vowel space is the number of vertices associated to the different vowel trapezia representing the vowel plots. The mean of each vowel distribution serves as an edge for the vowel space trapezoid and we reported the number of vertices. The hypothesis in terms of the number of vertices was “the higher the number of vertices, the bigger the vocalic space”, and then the clearer or the more separate the various vocalic realisations are and therefore the better the overall pronunciation might be.

## 4 Results

In this section, we present the different results from individual realisations of vowels to more global comparisons.

### 4.1 KDE of Vowel Realisations

We used kernel density estimates (KDE) to represent in three dimensions, the F1 and F2 probability density. We are using this visual representation as a cue for the separability of the different

vowels. We would expect the properly realised phonetic minimal pairs to be realised as two distinct cones. Conversely, when only one vortex or pyramid can be observed, we assume that the distinction between the two phonemes is not realised. We computed these KDE for the vowels shown in Figure 7, and we only show here the most relevant pair of confusing vowels (KIT vs. FLEECE) illustrated by two speakers for our learners.

## 4.2 Number of Vertices

Table 2 reports the number of vertices that is associated with the different vowel trapezia representing the vowel plots. Our hypothesis was “the higher the number of vertices, the bigger the vocalic space” and then the clearer or the more separate the various vocalic realisations were and therefore the better the overall pronunciation might be. This hypothesis is not verified, at least with our data.

Level	mean of vertices	support
1	6.71	7
2	6.73	11
3	6.50	4
4	6.00	1

Table 2: Mean of complex hull vertices per level for Italian speakers

## 4.3 Reference Vowel Inventories

For a global analysis, we tried to come up with a reference inventory of the phoneme systems, the vocalic system, because most of the subjects were assumed to have British pronunciation. We used the British transcription from the Longman Pronouncing Dictionary to try to estimate the reference vowel inventory. Such an undertaking is challenging because we need to eliminate the variants that are automatically assigned by the phonetic aligner. The variants, when available in the dictionary of the aligner, are selected on the basis of the acoustic signal. We systematically took the first variant when several were present.

The distribution of the vowel inventories that we would expect varies across speakers but we do not report phoneme error rates, we are trying to offer a global appreciation. This is based on the transcription of the target, the text that needed to be read by the different learners following the different tasks of the ISLE data. A total of 30,032 vowels

across the 23 Italian speakers were collected and analyzed. No filters, such as removing function words or focusing on stressed syllables only, were applied. The per-vowel distributions can be found in Figure 7. Monophthongs account for 79% of all collected phonemes, with /ə/ amounting to 19.2% of all vocalic occurrences with 5,757 tokens.

## 4.4 Correlations to Levenshtein Distance

The Levenshtein distance to the reference text read by the ISLE speakers (the smaller the distance, the better the pronunciation) proved to be robustly correlated to per-speaker mean of the probability scores ( $R=-0.94$ ), to the ability to classify monophthongs ( $R=-0.7$ ), and partially correlated to the learner grades ( $R=-0.57$ ) assigned by ISLE annotators.

The main result is a strong correlation ( $R = -0.94, p < 0.005$ ) of Whisper’s probability scores with the Levenshtein distance separating the transcriptions from the script of the reading assignment. Figure 8 confirms the hypothesis that higher probability scores in the Whisper prediction corresponds to a better pronunciation (lesser deviation from the expected realisations). Speakers whose automatic transcriptions have a higher Levenshtein distance are more likely to have lower Whisper probability scores.

The second observed correlation is with the acoustic data. The Levenshtein distance is partially correlated to the ability to classify monophthongs for each speaker on the basis of their formant values. We extracted the formant values from the forced aligned data and used the  $k$  nearest neighbour ( $k$ -NN) algorithm (Deng et al., 2016) to classify the monophthongs on the basis of their F1/F2 formant values<sup>1</sup>. We computed the accuracy for this classification task. The scatter plot on Figure 9 illustrates the relationship between the Levenshtein distance to the original text string and the accuracy reported for the per-speaker classification of Italian speaker’s monophthongs using the  $k$ -NN algorithm ( $k$ -NN Accuracy) represented on the  $y$ -axis. One can see a clear negative correlation between the Levenshtein distance and this  $k$ -NN Accuracy, as indicated by the downward sloping trend line and Pearson’s correlation coefficient  $R = -0.7$ . The relationship is statistically significant, with  $p < 0.001$ . The data points are some-

<sup>1</sup>Vowel discrimination between native and non-native realisations have already been tested with this type of clustering (Méli and Ballier, 2019).

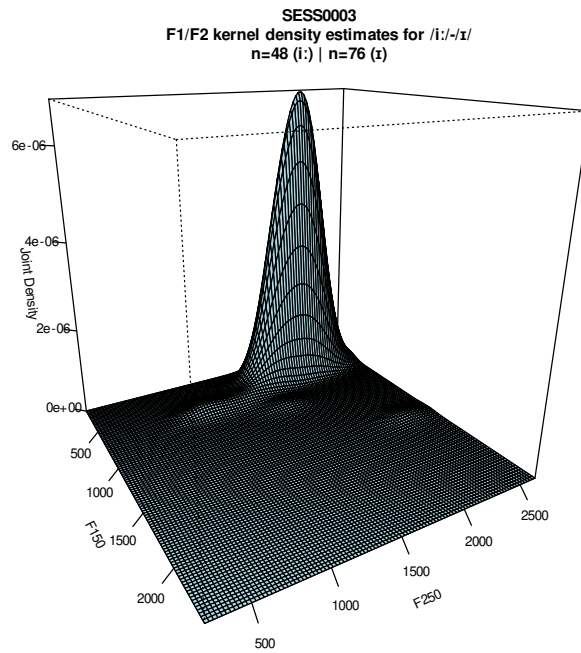


Figure 4: KDE estimate for F1 / F2 probability density for Speaker #S003

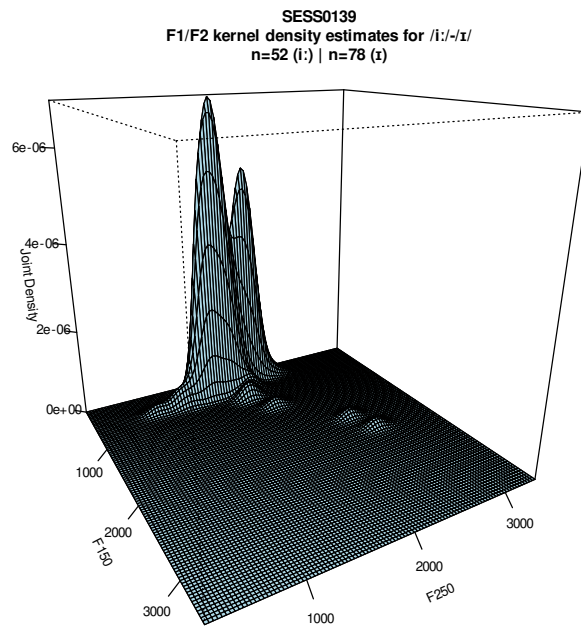


Figure 5: KDE estimate for F1 / F2 probability density for Speaker #139

Figure 6: Comparison of the two Kernel Density Estimates (KDE) for the KIT vs. FLEECE vowels for two speakers. The unimodal distribution of the acoustic realisations (one peak) suggests that speaker #3 does not properly categorise the two vowels (top), whereas speaker #139 produces two distinct series of realisations (bottom) for the KIT vs. FLEECE vowels, suggesting that the vowel categorisation has been properly acquired.

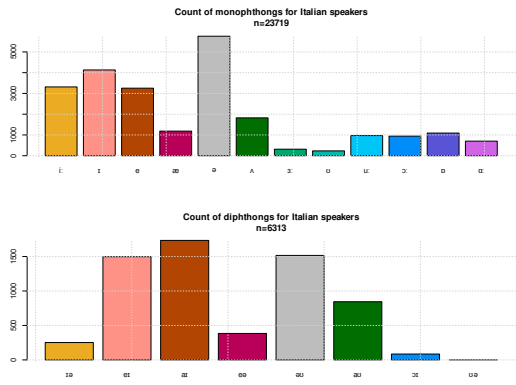


Figure 7: Vowel inventories aggregated on the 23 Italian Speakers, monophthongs (top) and diphthongs (bottom), based on the forced alignment of the tiny Whisper transcriptions

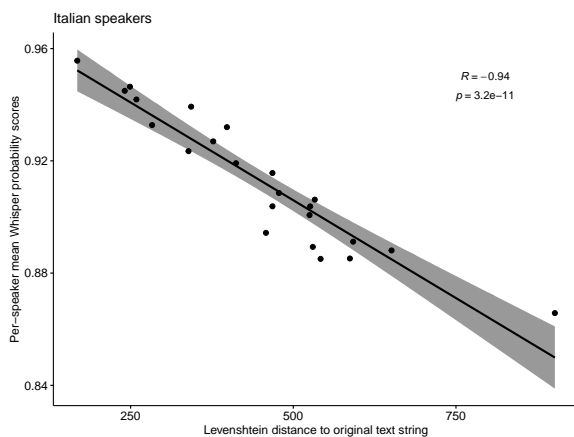


Figure 8: Negative correlation between the Levenshtein distance to the original text string and the per-speaker mean of the Whisper probability scores. The grey shaded area represents the confidence interval, which widens at the extremes of the x-axis, indicating less certainty in the prediction at these points.

what scattered around the trend line, but generally follow the negative trend. The grey shaded area represents the confidence interval around the regression line, which widens at the extremes of the x-axis, indicating less certainty in the prediction at these points. This plot confirms that our use of the Levenshtein distance is a sensible correlate to the assessment of phonetic quality: the more a text is altered from its original form, the harder it becomes for the k-NN algorithm to accurately classify the monophthongs of a given speaker. Admittedly, the accuracy reported is far from perfect, as the accuracy of the prediction (with 70% train, 30% test) ranged from about 0.35 to 0.55, but it should be borne in mind that vowel data points

for monophthongs partially overlap, so that accuracy for native speakers’ monophthong classification would also be limited. With a skewed distribution and 12 classes to predict, this is no easy task. Nevertheless, this accuracy of the classification of the monophthongs on the basis of their formant values correlates with the Levenshtein distance.

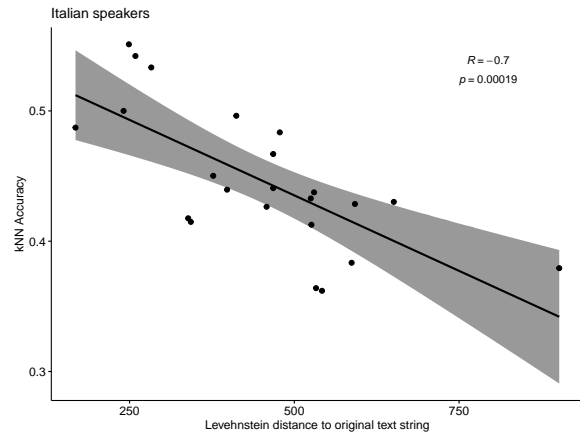


Figure 9: Negative correlation between the Levenshtein distance and the accuracy of the prediction of the monophthongs using k-NN. The grey shaded area represents the confidence interval, which widens at the extremes of the x-axis, indicating less certainty in the prediction at these points.

Finally, the correlation of the Levenshtein distance with the grades assigned by the ISLE annotators are weaker but the correlation remains statistically significant ( $R = -0.57$ ,  $p < 0.005$ ). Figure 10 suggests that as the Levenshtein distance increases (indicating greater difference from the original text), the annotators’ marks tend to decrease. This means that the annotators’ grading of the Italian speakers does decrease when the Whisper `tiny` model transcriptions deviate more from the original text. The Levenshtein distance is therefore a metric consistent with the grades assigned to the Italian learners in the ISLE corpus.

#### 4.5 Absence of Global Correlations

However, the analyses of 88 parameters related to vocalic data (*e.g.*, the Euclidean distances between each monophthong in the F1/F2 vocalic space) return no, or very weak, correlations with the Levenshtein distance. One exception may be found in the  $/i:/-/ɪ/$  distance ( $R = -0.56$ ,  $p < 0.005$ ). This validates our hypothesis that visual inspection of the KDE density of these two vowels is

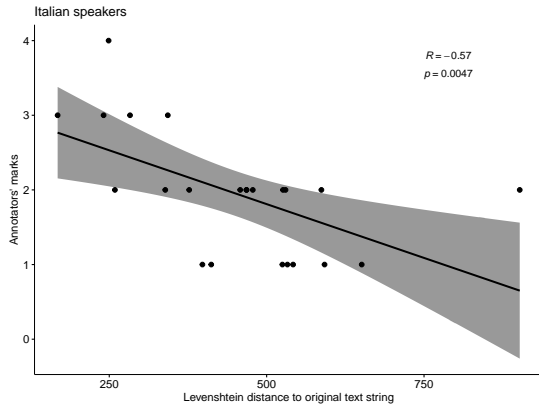


Figure 10: Negative correlation between the Levenshtein distance and the corpus annotator’s grades. The grey shaded area represents the confidence interval, which widens at the extremes of the x-axis, indicating less certainty in the prediction at these points.

a valid cue for the interpretation of the quality of realisations. This distinction between the two phonemes is noted in the papers describing the ISLE corpus (Atwell et al., 2003) and in the L2 phonetic literature (Kenworthy, 1987). This global representation of the probability density for F1 and F2 for these two vowels show distinct visual representations. We assume that phonetic realisations are distinctive if two peaks can be distinguished. Conversely, learners failing to mark F1/F2 differences for these vowels have a unimodal distribution. As can be seen on Figure 6, speaker 003 has a unimodal distribution for the F1/F2 realisations of the FLEECE and KIT vowels.

## 5 Discussion

To the best of our knowledge, this is the first paper that tries to correlate the grades assigned to taylor-made spoken corpora to Whisper outputs (transcriptions and internal representations of their probabilities) and phonetic correlates extracted from forced alignment of the Whisper transcriptions. Assuming we take the ISLE grades as golden reference taken for granted, the discussion bears on how we collected the phonetic data points (subsection 5.1), aggregated the Levenshtein distance neglecting task effects (subsection 5.2), compared scores of linguistic units varying in size and scope (subsection 5.3), measured the correlations (subsection 5.4) and on the Whisper outputs we have not investigated yet (subsection 5.5).

### 5.1 Precision of the Aligners

The first point to discuss is the precision of the aligner, the tool that automatically aligns the Whisper transcription to the signal. As shown in Figure 3, there may be errors in the forced alignment. We have used the P2FA aligner whose performances may be lower than more recent ones. The Montreal Forced aligner (McAuliffe et al., 2017) may produce better results, but is not that easy to integrate into our annotating pipeline. Previous research suggests that the precision of our pipeline may be lower than more recent ones (Méli et al., 2023). One key question is therefore that of the accuracy of the forced alignment performed by P2FA. A hopefully convincing way to answer this question is to plot the means of the mid-temporal F1 and F2 formant values and to compare them to established values in the literature. Figure 11 shows that the vocalic trapezoids thus obtained for per-sex average values are consistent with those listed in Deterding (1997). The lines trace the convex hulls of the sets of average F1/F2 values. Unfortunately, the number of vertices required to represent the trapezium did not present a consistent pattern correlating with Levenshtein distance or learner grades.

### 5.2 Task Effects: the Different Prompts

We merged the different sound files corresponding to the subtasks (see Table 1) to analyze the ISLE data and reported global results, in line with the global grading of the sound files by the annotators. We do not report the probability scores or the Levenshtein distance per type of prompts (see Table 1) and do not investigate whether some task effects could be measured, looking at the language prediction and the average probability assigned to the subtokens of the different group. A related research question is the need to estimate what would be the optimal duration of the data to be used by automatic systems when predicting the level of the learners.

### 5.3 Granularity and Scope of Scoring

Our analysis focuses on vowels, but the papers presenting the ISLE corpus also insisted on phone substitutions for consonants (Atwell et al., 2003). One of the difficulties of using Whisper scoring is that probability scores correspond to subtokens, which do not exactly correspond to syllables and rarely match phonemic representations.



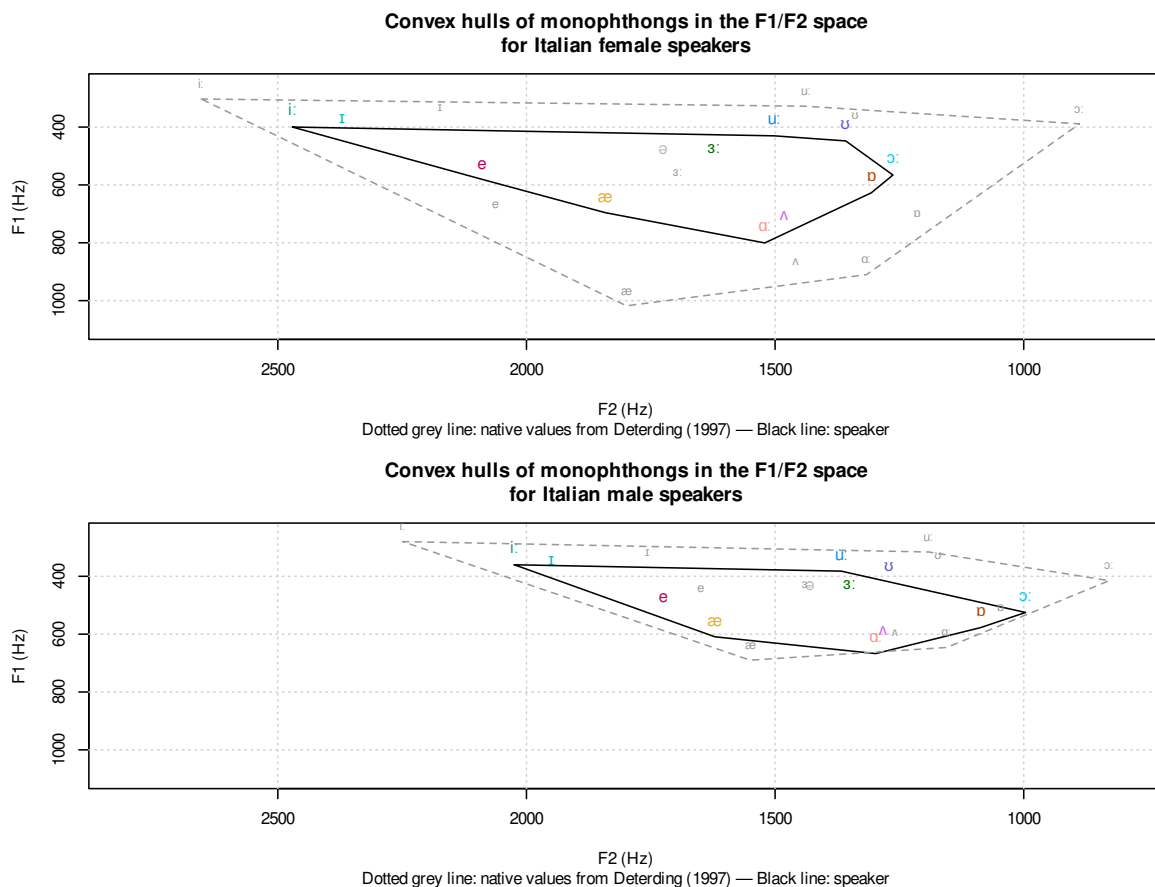


Figure 11: Convex Hulls of female (top, 6 vertices) and male speakers (bottom, 5 vertices) for all the Italian speakers. The dashed trapezium corresponds to the British reference means in lab conditions (hyperarticulation) reported in (Deterding, 1997)

We could at best report the confidence of the system (the Whisper probabilities) at the subtoken level, but in this paper, we mostly consider holistic estimations: Levenshtein distance (the transcription of the full recording) or mean probability of all the predicted subtokens. These 'textual' Whisper predictions can only be partially mapped to speech units of a different granularity. A phonemic transcription could parallel the Levenshtein distance and include phonological consonants. The PEASYV pipeline extracts formants and focuses on vowels. Vowel plots and their structures (numbers of vertices) are holistic representations, and so are vowel inventories. Most other metrics are between two vowels (distances or kernel density estimates) and may be used to monitor whether vowel distinctions have been acquired.

#### 5.4 Alternative Methods

For our 'kitchen sink method', the phonetic variables related to vowel plots reveal little correla-

tion to the levels assigned by the annotators or to the Levenshtein distance. In particular, the number of vertices (at least based on our forced alignment) does not seem to be a plausible correlate for the level of the learners. An ablation method for a multinomial ordinal regression may highlight other variables. Another approach of the vowel realisations is based on Pillai scores. For an intrinsic measure of the dispersion of L2 speech, we may use Pillai scores applied for L2 speech for vowels (Mairano et al., 2019, 2023). Additionally, we have not explored clustering techniques that would try to investigate if the grouped phonetic data-points corresponding to the reference grades of the corpus would produce consistent results. Assuming there are actually four levels to be considered for the Italian ISLE speakers, what would be the confusion matrices of these four levels using k-means (with  $k$  equals 4 for the four levels) for the vowel acoustic correlates we examined? Would the four clusters produced by the k-means correspond to the four levels of the corpus grades?

## 5.5 Whisper Scoring of Language Detection

Another feature is worth investigation. Whisper has been trained to recognise the language spoken as an identification task (Radford et al., 2023). This language identification (and its associated probability) could be potentially used to analyse learner data, to discriminate speakers predicted to be English or Italian. For example, using Whisper’s `large` model to predict the language spoken by the Italian speaker, we observed that more advanced learners (level 3 or 4) were labelled as English, whereas the learners graded as level 2 were either predicted as being English or Italian. With presumably the most robust Whisper model, there seems to be a threshold between less advanced learners whose first language is predicted (Italian) and more advanced learners that are detected as being English. The most interesting case study is the intermediate group of Italian learners labelled “2” in the ISLE corpus that is sometimes predicted as English or as Italian. We want to analyse the potential phonetic correlation that may account for this judgement, therefore potentially validating the idea of a threshold detecting between less advanced learners and more advanced learners with Whisper. We intend to compare these Whisper predictions with the phonetic realisations (including consonants) using the P2FA aligner to compare the various phonetic realisations with the Whisper predictions, trying to account for that difference in the system.

## 6 Conclusion

In this paper, we have tried to correlate Whisper’s transcriptions with the levels assigned to the Italian learners in the ISLE corpus and with acoustic correlates of vowels. We used the Levenshtein distance to measure deviation from the read texts for each speaker based on Whisper’s ASR transcriptions and we used forced alignment and the PEASYV annotation pipeline (Méli and Bal-lier, 2023) to produce our vowel-based acoustic data (vowel formants), phone reports and phonetic measurements. Levenshtein distance does correlate with the levels, but the acoustic correlates we analysed are not convincing. The assumption that Whisper scoring could be a good cue to the quality of the phonetic realisation is validated because it is negatively correlated to the deviation from the reference read text measured with the Levenshtein distance. Our explorations of

the holistic phonetic correlates is less successful. Holistic representations like vowel plots apparently fail to be correlated to the grades attributed to the Italian learners in the ISLE corpus. Nevertheless, the type of trapezoids we produced with the PEASYV pipeline could be used in Computer-Assisted Pronunciation Training (CAPT) systems (Rogerson-Revell, 2021) as actionable visualisations for teachers and expert users.

## Limitations

The first limitation is the number of speakers and languages for our analysis. Because graded spoken learner corpora are not that frequent, we focused on the ISLE data, and only on the Italian component, since the German component has a different level distribution. Only 11 male speakers were analysed, which also introduces a gender limitation to our work. A second limitation is the focus on segmental errors, like many studies based on Automatic Speech Recognition. The analysis of L2 speech should also account for suprasegmental features. Last, our metrics, visualisation and k-NN analysis of the vowels mostly tackled monophthongs and not diphthongs and these techniques in investigating vowel separability may be contradicted by perception tests.

## Ethics Statement

It is important to note that the Whisper scoring should not be used as a substitute for human feedback. Whisper does not explicitly monitor suprasegmental features. As noted during our analyses, the probabilities associated with the Whisper transcriptions do not necessarily guarantee that the word transcribed by Whisper is the most accurate rendition of what the learner actually pronounced. As a consequence, we endorse a human-in-the loop approach to this kind of technology.

## Acknowledgments

We thank the NLP4CALL anonymous reviewers, Sara Ng and Alice Henderson for their comments on a previous version of this paper. We also thank Jean-Baptiste Yunès for his implementation of the extraction of the Whisper probabilities.

## References

- Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition*, 6:1817–1853.
- Vipul Arora, Aditi Lahiri, and Henning Reetz. 2018. Phonological feature-based speech recognition system for pronunciation training in non-native language learning. *The Journal of the Acoustical Society of America*, 143(1):98–108.
- Eric Atwell, Paul Howarth, and Clive Souter. 2003. The isle corpus: Italian and German spoken learner’s english. *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*, 27:5–18.
- Nicolas Ballier, Taylor Arnold, Adrien Méli, Tori Thurston, and Jean-Baptiste Yunès. accepted(a). Whisper for I2 speech scoring. *International Journal of Speech Technology*, 27(4):–.
- Nicolas Ballier, Léa Burin, Behnoosh Namdarzadeh, Sara Ng, and Jean-Baptiste Yunès. accepted(b). Probing Whisper Predictions for French, English and Persian Transcriptions. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, Online. Association for Computational Linguistics.
- Nicolas Ballier and Philippe Martin. 2015. Speech annotation of learner corpora. *The Cambridge handbook of learner corpus research*, pages 107–134.
- Nicolas Ballier, Adrien Méli, Maelle Amand, and Jean-Baptiste Yunès. 2023. Using whisper LLM for automatic phonetic diagnosis of L2 speech, a case study with French learners of English. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 282–292, Online. Association for Computational Linguistics.
- Paul Boersma and David Weenink. 2019. Praat: doing phonetics by computer [computer program]. version 6.1.07, retrieved 26 november 2019 from <http://www.praat.org/>.
- May Pik Yu Chan, June Choe, Aini Li, Yiran Chen, Xin Gao, and Nicole Holliday. 2022. Training and typological bias in ASR performance for world Englishes. In *Proc. Interspeech 2022*, pages 1273–1277.
- Vincent Chanethom and Alice Henderson. 2023. Alignment in ASR and L1 Listeners’ Recognition of L2 Learner Speech: French EFL Learners & Dictation.Io. *Research in Language*, 21(3):245–266.
- Jonathan Dalby, Diane Kewley-Port, and Roy Sillings. 1998. Language-specific pronunciation training using the hearsay system. In *Speech Technology in Language Learning*, pages 25–28.
- Zhenyun Deng, Xiaoshu Zhu, Debo Cheng, Ming Zong, and Shichao Zhang. 2016. Efficient knn classification algorithm for big data. *Neurocomputing*, 195:143–148.
- David Deterding. 1997. The Formants of Monophthong Vowels in Standard Southern British English Pronunciation. *Journal of the International Phonetic Association*, 27(1-2):47–55.
- Georgi Gerganov. 2023. whisper.cpp : A high-performance inference of OpenAI’s whisper automatic speech recognition (ASR) model.
- S. Inceoglu, Hyojung Lim, and Wen-Hsin Chen. 2020. ASR for EFL pronunciation practice: Segmental development and learners’ beliefs. *The Journal of Asia TEFL*, 17(3):824–840.
- Joanne Kenworthy. 1987. *Teaching English pronunciation*. Longman.
- Paolo Mairano, Caroline Bouzon, Marc Capliez, and Valentina De Iacovo. 2019. Acoustic distances, pillai scores and lda classification scores as metrics of I2 comprehensibility and nativelikeness. In *ICPhS2019*, pages 1104–1108.
- Paolo Mairano, Fabián Santiago, and Leonardo Contreras Roa. 2023. Can L2 Pronunciation Be Evaluated without Reference to a Native Model? Pillai Scores for the Intrinsic Evaluation of L2 Vowels. *Languages*, 8(4):280.
- Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.
- Shannon McCrocklin and Idée Edalatishams. 2020. Revisiting popular speech recognition software for ESL speech. *Tesol Quarterly*, 54(4):1086–1097.
- Shannon McCrocklin, Abdulsamad Humaidan, and Idée e Edalatishams. 2019. ASR dictation program accuracy: Have current programs improved? *Pronunciation in Second Language Learning and Teaching Proceedings*, 10(1):191–200.
- Adrien Méli and Nicolas Ballier. 2019. Analyse de la production de voyelles anglaises par des apprenants francophones, l’acquisition du contraste /ɪ/-/i:/ à la lumière des k-nn. *Anglophonia / Caliban-French Journal of English Linguistics*, 27.
- Wolfgang Menzel, Eric Atwell, Patrizia Bonaventura, Daniel Herron, Peter Howarth, Rachel Morton, and Clive Souter. 2000. The ISLE corpus of non-native spoken English. In *Proceedings of LREC 2000: Language Resources and Evaluation Conference, vol. 2*, pages 957–964. European Language Resources Association.

- Geoffrey Stewart Morrison. 2012. [Theories of Vowel Inherent Spectral Change](#). In *Vowel Inherent Spectral Change*, pages 31–47. Springer Science + Business Media.
- Geoffrey Stewart Morrison and Terrance M. Nearey. 2007. [Testing theories of vowel inherent spectral change](#). *J. Acoust. Soc. Am.*, 122(1):EL15–EL22.
- Adrien Méli and Nicolas Ballier. 2023. PEASYV: A procedure to obtain phonetic data from subtitled videos. *Proceedings of the International Congress of Phonetic Sciences*, pages 3211 – 3215.
- Adrien Méli, Steven Coats, and Nicolas Ballier. 2023. Methods for phonetic scraping of youtube videos. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 244–249.
- Terrance M. Nearey. 2012. [Vowel Inherent Spectral Change in the Vowels of North American English](#). In *Vowel Inherent Spectral Change*, pages 49–85. Springer.
- Terrance M. Nearey and Peter .F. Assmann. 1986. Modeling the role of vowel inherent spectral change in vowel identification. *JASA*, 125:2387-97.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.
- Pamela M Rogerson-Revell. 2021. Computer-assisted pronunciation training (CAPT): Current issues and future directions. *Relc Journal*, 52(1):189–205.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Cristian Tejedor-García, Valentín Cardeñoso-Payo, and David Escudero-Mancebo. 2021. Automatic speech recognition (ASR) systems applied to pronunciation assessment of L2 Spanish for Japanese speakers. *Applied Sciences*, 11(15):6695.
- John C. Wells. 2000. *Longman Pronunciation Dictionary*. Pearson Longman, London.
- Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, 123(5):5687-5690.

# Leading by Example: The Use of Generative Artificial Intelligence to Create Pedagogically Suitable Example Sentences

Jasper Degraeuwe and Patrick Goethals

LT<sup>3</sup> / MULTIPLES

Ghent University

Belgium

Jasper.Degraeuwe@UGent.be, Patrick.Goethals@UGent.be

## Abstract

Studies on second language acquisition have argued in favour of practising vocabulary in authentic contexts. After the tradition of obtaining these usage examples by “invention” (i.e. language experts creating examples based on their intuitions) was superseded by corpus-based approaches (i.e. using dedicated tools to select examples from corpora), the rise of large language models led to a third possible “data source”: Generative Artificial Intelligence (GenAI). This paper aims to assess GenAI-based examples in terms of their pedagogical suitability by conducting an experiment in which second language (L2) learners compare GenAI-based examples to corpus-based ones, for L2 Spanish. The study shows that L2 learners find GenAI-based sentences more suitable than corpus-based sentences, with – on a total of 400 pairwise comparisons – 265 artificial examples being found most suitable by all learners (compared to 10 corpus-based examples). The prompt type (different zero-shot and few-shot prompts were designed) did not have a noticeable impact on the results. Importantly, the GenAI approach also yielded a number of unsuitable example sentences, leading us to conclude that a “hybrid” method which takes authentic corpus-based examples as its starting point and employs GenAI models to rewrite the examples might combine the best of both worlds.

## 1 Introduction

Although vocabulary items can be learnt in isolation (e.g., through flash cards; Nation, 2022), providing in-context usage examples of vocabulary items strengthens word form - word meaning associations (Laufer and Shmueli, 1997) and has shown to foster both language comprehension and production (Frankenberg-Garcia, 2012, 2014). As

a result, example sentences are often used in vocabulary lists, learners’ dictionaries, and grammar sections as a means to illustrate the usage(s) of vocabulary items and grammatical patterns. Some types of materials even depend entirely on the presence of example sentences, such as fill-in-the-blanks and in-context translation exercises.

To obtain example sentences, linguistic disciplines have a long tradition of using intuited/invented examples (IEs) created by language experts such as lexicographers and teachers (Cook, 2001; Laufer, 1992; Stefanowitsch, 2020). The underlying idea is that their advanced linguistic competence allows them to formulate well-formed, relevant, and grammatically correct sentences. However, the last decades witnessed an increased interest in the selection of example sentences from digital(ised) native (L1) corpora, first manually and later following (semi-)automatic selection procedures (Frankenberg-Garcia et al., 2021). Even though well-designed IEs can have pedagogical value (Cook, 2001), carefully selected corpus examples can be considered more authentic, reliable, and valid expressions of language (Firth, 1968; Stefanowitsch, 2020). Moreover, thanks to the continued improvements made to the tools and techniques used for corpus processing and consultation, performing corpus queries to extract sentences that should meet a given set of criteria has become highly efficient.

Recently, major developments in the field of Generative Artificial Intelligence (GenAI) uncovered another pathway to obtain example sentences: based on a prompt specifying the desired criteria, GenAI systems can be asked to output a series of – according to the model – suitable usage examples. Although the artificial way in which they are conceived bears some resemblance with IEs, these examples can also be said to have a corpus-based touch, since the GenAI tools that produce them are trained on (extremely large) collections of text.

---

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

In the present paper, we present an experiment in which second language (L2) learners of Spanish compare example sentences selected following a *corpus-based method* to examples created following a *GenAI-based method*. In doing so, we aim to make a contribution to the assessment of the pedagogical usability and validity of artificially generated learning materials. The paper is structured as follows: after providing an overview of the related research in Section 2, we describe the methodology (Section 3) and elaborate on the results (Section 4). A discussion of those results is presented in Section 5. Finally, Section 6 includes a conclusion together with possible directions for future research.

## 2 Related Research

Broadly speaking, the criteria which define a “good” example can be categorised as either form-related or content-related. The former type refers to grammatical soundness and straightforward superficial properties such as a capitalised first letter and a punctuation mark at the end of the sentence. Content-related criteria, on the other hand, encompass features such as naturalness (i.e. containing formulations which can also be encountered in real-life language use), context independence, intelligibility (often captured in terms of sentence length and number of difficult words), typicality (i.e. containing collocations or colligations), and informativeness (i.e. containing clues which help understand the meaning of the target item).

The definition of sentence selection criteria has been considered from both a pedagogical (Pilán et al., 2016) and a lexicographic point of view (Atkins and Rundell, 2008). Although many criteria apply to both of them, the two perspectives also exhibit differences. With regard to the intelligibility criterion, lexicographic resources tend to prefer short sentences, while language learning resources are considerably more tolerant towards long sentences, as exposing learners to more (relevant) context can be beneficial for the learning process (Kosem et al., 2019). Secondly, in a language learning setting, the criteria of informativeness and typicality are often isolated and linked to, respectively, the concepts of “decoding” (i.e. aimed at fostering comprehension) and “encoding” (i.e. aimed at fostering production). As these concepts reflect two very distinct aspects of language learning, the example selec-

tion methods used to create language learning resources often focus on only one of these two criteria, instead of looking for sentences incorporating both (Frankenberg-Garcia, 2014). Finally, selecting sentences for pedagogical purposes also requires assessing a sentence’s complexity in terms of learner proficiency levels and adapting the selection accordingly, as there exist considerable differences between the language knowledge of beginning, intermediate, and advanced learners.

### 2.1 Corpus-based Examples

Finding its origins in the grammar-translation method of the mid-19<sup>th</sup> century, invented examples (IEs) have long been the primary source for presenting new words or exemplifying linguistic phenomena of a lexical (i.e. collocations) or grammatical (i.e. colligations) nature (Cook, 2001). In essence, IEs are concocted by experts (e.g., L2 teachers or lexicographers) and rely on the intuitions these experts have about the usage of the word/pattern to be presented/exemplified. Towards the end of the 20<sup>th</sup> century, however, the rise of online accessible corpora together with advances in the technological means to process and consult them opened new horizons in the selection/creation of examples. The COBUILD initiative (Sinclair, 1987), for example, radically rejected the use of IEs and only used unaltered corpus examples in its resources.

Importantly, much of this research into corpus-based example selection methods originated from lexicographic motives, which – as mentioned earlier – do not necessarily include pedagogical considerations. Yet, many lexicographic methods were (and still are) also used for pedagogical purposes (Kosem et al., 2019). One of those methods is GDEX (Good Dictionary EXamples; Kilgariff et al., 2008), which marked a major milestone in the field of corpus-based example selection. In brief, the method takes as input a list of corpus concordances for a given target item and returns a ranked version of that list. The main particularities of GDEX are the overall scoring algorithm with adjustable parameters (a so-called “GDEX configuration”) and the “second collocate” classifier that prioritises sentences containing the most typical collocates of a given collocation. Moreover, as the adjustable parameters allow users to tailor the sentence selection criteria to their specific needs, the need for posterior manual revisions also decreases.

As mentioned above, GDEX is – despite its lexicographic origins – widely applied in language learning contexts as well (Kallas et al., 2015; Smith et al., 2010). The SKELL tool, for example, employs GDEX to retrieve the most useful examples for language learners from large corpora and return them as a ranked list (see Figure 1). Nevertheless, extra curation is still required when selecting examples from GDEX-based concordances, particularly when priority has to be given to specific collocation or colligation patterns (Frankenberg-Garcia et al., 2021).

Regarding the (limited) research dedicated to corpus-based sentence selection specifically for language learning purposes, a first important study to highlight is that on HitEx (Pilán et al., 2016), a sentence selection framework for L2 Swedish. Combining both rule-based and machine learning-based components, the HitEx framework pays special attention to linguistic complexity and independence from the surrounding corpus sentences, but also takes into account well-formedness and a series of structural criteria (e.g., presence of modal verbs and sentence length) and lexical criteria (e.g., word frequency and presence of proper names). Next, Heck and Meurers (2022) developed an algorithm which can select suitable examples to be used as input for L2 English grammar exercises. Apart from offering different data sources to choose from (the web, precompiled corpora, or custom texts), the method also includes tailor-made selection criteria such as the presence of relative pronouns, extraposition, and preposition stranding.

## 2.2 GenAI-based Examples

The process to obtain artificially generated example sentences is very straightforward: based on a

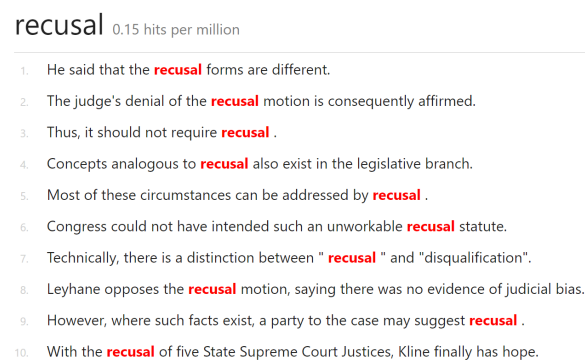


Figure 1: SKELL output for *recusal*. Date of performing query: 6 June 2024.

natural language prompt as input, a GenAI model can be asked to return a series of sentences, without any specific prior training. Depending on the model's architecture, the prompt can be formulated as a *zero-shot learning* or *few-shot learning* phrase. As shown in Figure 2, zero-shot prompts can be written as if one is making a request/asking a question to a fellow human being. In this case, we simply ask the model for three sentences that have to meet a set of criteria (sentences cannot be longer than 20 words and have to clarify the meaning and usage of the target item). With few-shot prompting, the request/question is complemented by (or sometimes even replaced by) a limited number of examples the model can learn from, as illustrated in Figure 3. In this case, we just take the three sentences returned by the model for the zero-shot query, convert them into a structured format, and prompt the model to return the corresponding information for a new item. The underlying idea is that the model will “deduce” the desired characteristics from the examples (e.g., the sentence length) and use this information when generating the response for the new items.

The GenAI-driven creation of example sen-

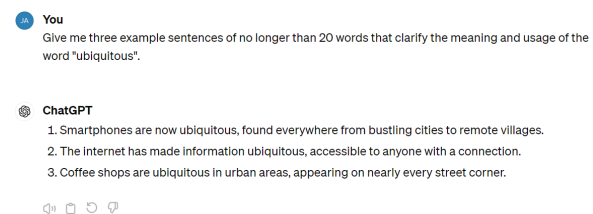


Figure 2: Artificially generated example sentences by means of zero-shot learning (i.e. a simple instruction/question) as prompting technique. Model: OpenAI's GPT-3.5 (accessed through ChatGPT interface). Date of performing prompt: 6 May 2024.

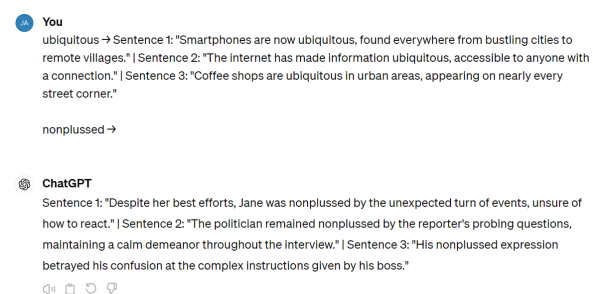


Figure 3: Artificially generated example sentences by means of few-shot learning as prompting technique. Model: OpenAI's GPT-3.5 (accessed through ChatGPT interface). Date of performing prompt: 6 May 2024.

tences has been explored in the context of data-driven learning (Crosthwaite and Baisa, 2023) and as a means to clarify difficult words (Kohnke et al., 2023). A large-scale study which specifically assesses the pedagogical suitability of artificially generated example sentences has not yet been performed, a gap we aim to fill with this study. However, an important observation to make in the context of GenAI research is that the non-deterministic nature of the (online) models makes the research, per definition, irreproducible. Due to randomness being included in the generation process, GenAI models can produce different outputs at different times for the same input prompts<sup>1</sup>. Regular updates to existing models (e.g., of OpenAI’s proprietary GPT-3.5) and launches of new models (e.g., OpenAI’s GPT-4 and GPT-4o or Google’s Gemini models) further complicate adequately assessing the pedagogical value of artificially generated sentences. Nevertheless, even given these methodological drawbacks, there is a growing consensus that scientific research is needed to explore the use of GenAI models for the creation of all kinds of L2 learning materials and to help shed light on the pedagogical suitability of this approach (Crosthwaite and Baisa, 2023; Caines et al., 2023).

### 3 Methodology

As mentioned in the introduction, our aim is to evaluate the pedagogical usability of artificially generated example sentences by comparing them to corpus-based sentences, which have become the standard approach for obtaining pedagogically suitable example sentences. To this end, we organise an experiment in which L2 Spanish learners compare corpus examples selected according to a dedicated sentence selection framework (Section 3.2) with examples generated by means of OpenAI’s GPT-3.5 Turbo model, using different types of prompts (Section 3.3). In total, we recruit seven students from both beginner and advanced proficiency levels, all with Dutch as their L1 (see Section 3.4 for more details). For the former group, we envisage a general vocabulary learning course as the target setting; for the latter,

<sup>1</sup>Recently, many large language model providers added a “seed” parameter to their (API) interface, allowing developers to receive (mostly) consistent outputs. Yet, due to the inherent non-determinism of GenAI models, there will always be a small chance that responses differ even when a seeding parameter is specified.

we take a language for specific purposes course on legal vocabulary as our anchor point. The research questions we aim to answer are defined as follows:

1. Which source of example sentences is found most suitable by L2 learners: corpus-based or GenAI-based?
2. Which type of prompt used to query the GenAI model is found most suitable by L2 learners: zero shot (with varying degrees of specificity) or few shot?

#### 3.1 Dataset

For each of the two target groups (beginner and advanced), we collect a set of 250 target items, which are selected based on their relevance and representativeness for the target setting defined above. For the beginner group, we take the first 150 nouns, 50 verbs, and 50 adjectives from the 1,001-2,000 frequency range in the Davies and Hayward Davies (2018) word list, excluding Spanish-Dutch cognates (e.g., ES *proyecto* - NL *project* - EN *project*). For the advanced group, we take a 25M specialised corpus containing newspaper articles on legal topics<sup>2</sup> as our starting point, rank all words in the corpus based on Odds Ratio as the keyness metric (Pojanapunya and Watson Todd, 2018; Gabrielatos, 2018) and select the first 150 nouns, 50 verbs, and 50 adjectives from the resulting list. Apart from cognates, we also exclude region-specific eponyms (e.g., *baltarismo*, which refers to the political movement named after the Galician politician José Manuel Baltar) and derivations with *ex*, *sub*, and *vice* as the prefixes (e.g., *exdiputado*: ‘former MP’; *subgobernador*: ‘vice governor’; *vicepresidente*: ‘vice president’).

#### 3.2 Corpus-based Examples

To obtain corpus-based sentences for the 500 target items, we develop a dedicated framework to select examples from corpora. Our framework – named SeleJemCor (Selección de Ejemplos de Corpus) – builds on the work of Pilán et al. (2016), who developed the HitEx sentence selection framework for L2 Swedish (see also Section 2.1). In comparison to HitEx, our framework – the first of its kind for L2 Spanish – includes the integration of a tailor-made word difficulty classifier and the promotion of *typicality*

<sup>2</sup>The corpus is available within the Spanish Corpus Annotation Project (SCAP; Goethals, 2018).



to being a main selection criterion as novel aspects. A comprehensive overview of all selection criteria included in the framework is presented in Appendix A. The Python implementation of the framework is made publicly available in a [GitHub repository](#). To obtain the morphosyntactic information required for certain selection criteria (e.g., on part-of-speech tags, morphosyntactic features, and dependency relations), the Python module makes use of `spaCy`'s automatic morphosyntactic analysis pipeline<sup>3</sup>. To render our framework as language-independent as possible, we use the morphosyntactic categories and labels proposed by the [Universal Dependencies](#) initiative (Nivre et al., 2016).

All Boolean criteria in `SeEjemCor` function as *filters* (i.e. if the criterion is not met, the sentence will be excluded from the selection), whereas all numerical criteria function as *rankers* (i.e. the closer the numerical value lies to the desired value, the higher the sentence will be ranked). For filters, criterion values can be set to either *True* (filter active, all sentences which do not pass the filter are excluded) or *None* (filter inactive). For rankers, values can be set to any numerical value (in which case the criterion will act as a threshold-based ranker, with all sentences obtaining a better value than the threshold being considered equally suitable), to *all* (in which case the selection algorithm will simply rank all sentences from highest to lowest value), or to *None* (ranker inactive). In the end, all sentences which have not been filtered out receive one single overall “goodness score”, which corresponds to the average of all individual ranking positions.

We apply the `SeEjemCor` framework to a 7.5M corpus containing accessible reportages about tourist destinations<sup>4</sup> (for the 250 items in the beginner group) and to the abovementioned 25M specialised corpus containing newspaper articles on legal topics (for the 250 items in the advanced group). For each target item, we select the top-ranked sentence according to the selection algorithm explained above. The values set for the different selection criteria are included in Appendix A. For the advanced group, we make

<sup>3</sup>Even though other NLP toolkits such as `UDPipe` and `Stanza` tend to perform (slightly) better at tagging and parsing natural text, `spaCy`'s built-in large and Transformer-based models have shown to achieve near state-of-the-art performance with a significantly higher processing speed.

<sup>4</sup>Also compiled within SCAP.

the values slightly more tolerant in terms of non-lemmatised tokens, modal verbs, word frequency, and out-of-vocabulary words.

### 3.3 GenAI-based Examples

To obtain the artificially generated sentences, we use OpenAI's GPT-3.5 Turbo model. We define four different prompt types and corresponding prompt texts to access the model, with the prompt texts also varying according to the target group. The prompts define both a *system role* (which specifies the way in which the model answers questions) and a *user role* (which specifies the *output* that should be returned). A short description of the prompt types is provided below (see Appendix B for the full overview):

1. **ZS-GEN** (zero-shot general): only the broad context (L2 learning setting; Spanish as target language; desired sentence length; sentence has to be usage example) is included in the prompt.
2. **ZS-GEN+AUD** (ZS-GEN plus target audience): apart from the broad context, also the target audience is specified in the prompt.
3. **ZS-GEN+AUD+CRIT** (ZS-GEN+AUD plus criteria): next to the broad context and the target audience, the prompt also includes the specific “goodness” criteria the output sentence should adhere to.
4. **FEWSHOT**: a limited number of suitable sentences (one sentence for each part of speech; with target words that do not occur in dataset) are provided in the prompt for the model to learn. The example words are selected from the 2,001-3,000 frequency range in [Davies and Hayward Davies \(2018\)](#) and the sentences are extracted from the Spanish Clave dictionary ([González, 2012](#)). The prompt also presents the broad context and differentiates between the two target audiences (see ZS-GEN and ZS-GEN+AUD above).

To enable the analysis at the layer of the prompt type, we randomly subdivide the 250 items in each group (beginner and advanced) into five subsets of 50 items (30 nouns, 10 verbs, and 10 adjectives). For the 50 items in the first subset (IDs 1 and 6), we generate an example sentence based on the ZS-GEN prompt type; for the second set (IDs 2 and

	A	B	C	D	E
1	word	1	2	best	comment
2	<b>cultural</b> (cultureel)	Desde el Aeropuerto Internacional Augusto C. Sandino se llega con facilidad al centro de la ciudad, donde se conserva gran parte de la riqueza histórica y <b>cultural</b> .	Es importante tener sensibilidad <b>cultural</b> al viajar a un país extranjero para evitar malentendidos.		
3	<b>numeroso</b> (talrijk)	El concierto fue un éxito gracias a la <b>numerosa</b> asistencia de fans entusiastas.	<b>Numerosos</b> grupos venidos de todos los rincones de China se arremolinan cada día frente a las obras.		
4	<b>intenso</b> (intens, intensief, zwaar)	La luz es más <b>intensa</b> que la que ilumina la costa norte.	La tormenta de verano trajo consigo lluvias <b>intensas</b> que inundaron las calles del pueblo.		

Figure 4: Example of pairwise comparison between corpus-based and GenAI-based example sentences (in subset 1 to 4 and subset 6 to 9). The order in which the sentences are presented is randomised.

	A	B	C	D	E	F	G	H	I
1	word	1	2	3	4	5	best	worst	comment
11	<b>inmenso</b> (enorm, gigantisch, immens, kolossaal, onmeteljk)	El lago Titicaca es un cuerpo de agua <b>inmenso</b> que comparten Perú y Bolivia, rodeado de una belleza natural impresionante.	Leyendas aparte, Uyuni es un <b>inmenso</b> océano mineral que ocupa una superficie de 12.000 kilómetros cuadrados en la región boliviana de	El océano era <b>inmenso</b> y azul, extendiéndose hasta donde alcanzaba la vista desde la costa.	El amor que siento por mi familia es tan <b>inmenso</b> que no cabe en palabras para expresarlo completamente.	El amor de una madre por su hijo es <b>inmenso</b> y siempre está presente en todos los momentos de la vida.			
12	<b>el pelo</b> (haar)	Amanda tiene el <b>pelo</b> largo y rubio, le encanta peinarse con trenzas y coletas para clases de yoga.	Ella tiene el <b>pelo</b> largo y rizado, le queda muy bonito.	Mi hermana tiene el <b>pelo</b> largo y rizado, siempre lo lleva recogido en una cola de caballo.	A Marta le encanta cambiar de peinado y color de <b>pelo</b> cada vez que inicia una nueva estación.	Un joven europeo con largo <b>pelo</b> rizado me recibe con una sonrisa.			
13	<b>el acontecimiento</b> (evenement, gebeurtenis)	El <b>acontecimiento</b> cultural más importante del año será la inauguración de la exposición de arte contemporáneo en el museo	El <b>acontecimiento</b> más importante del año será la celebración del bicentenario de la independencia de nuestro país.	El <b>acontecimiento</b> más importante del año fue la visita del presidente extranjero a nuestra ciudad.	Uno de los <b>acontecimientos</b> más importantes es la exposición Brel, el derecho a soñar.	El concierto de anoche fue un emocionante <b>acontecimiento</b> cultural que disfrutamos juntos.			

Figure 5: Example of BWS comparison between corpus-based and GenAI-based example sentences (in subset 5 and 10). The order in which the sentences are presented is randomised.

Prompt type	Subset ID	
	BEG	ADV
ZS-GEN	1	6
ZS-GEN+AUD	2	7
ZS-GEN+AUD+CRIT	3	8
FEWSHOT	4	9
ALL	5	10

Table 1: Overview of prompt types used to generate artificial example sentences. “BEG” stands for beginner, “ADV” for advanced.

7) based on ZS-GEN+AUD; for the third set (IDs 3 and 8) based on ZS-GEN+AUD+CRIT; and for the fourth set (IDs 4 and 9) based on FEWSHOT (see Table 1). For the 50 items in the fifth subset (IDs 5 and 10), we generate an artificial example sentence based on all four prompt types. Finally, a Dutch translation is added for all 500 target items in the dataset (see Table 2 for a dataset sample).

### 3.4 Evaluation Procedure

For each of the two target audiences (beginner and advanced), the first four subsets are used to perform pairwise comparisons between corpus-based sentences and artificially generated ones. As the artificial sentences are generated based on different prompts, comparing the results at subset level will also enable us to gain insights into the per-

formance of each prompt type. The fifth subset is used to compare all five possible sentence sources (i.e. corpus-based and the four different GenAI prompts) at once in a best-worst scale (BWS) setup. The 250 beginner items are evaluated by three L2 Spanish learners ( $\approx$  B1 proficiency level, 19 years old, L1 Dutch), the 250 advanced items are assessed by four learners ( $\approx$  C1 proficiency level, 22-24 years old, L1 Dutch)<sup>5</sup>.

Prior to starting the experiment, participants were given a written document including the instructions, which were discussed orally with one of the researchers involved in the study. In the pairwise comparisons, participants were asked to indicate the best sentence, as illustrated in Figure 4; in the BWS comparisons, they were asked to indicate both the best and the worst one, as illustrated in Figure 5. To make the term “best” as concrete as possible, the instructions stipulated that the participants should first check if the sentences complied with a series of criteria, which are explained below. Together, these descriptions reflect how the term “pedagogical suitability” as used in this paper should be interpreted.

- The sentence is not a definition. If it is, the participant should write “definition” in the

<sup>5</sup>All students are enrolled in the Applied Linguistics study career at Ghent University, Belgium. The Applied Linguistics curriculum stipulates – based on the CEFR scale – the minimal linguistic competences students should gain before they are admitted to the next year of the career. As a result, we can estimate the proficiency level of the learners based on the year they are enrolled in.

Item	POS	Value	ID	Corpus-based	GenAI-based
<i>enemigo</i> (‘vijand’)	NOUN	1,024	1	El coche es el <b>enemigo</b> público número uno: en Londres se aplica una tasa ambiental a los vehículos más contaminantes.	Durante la guerra, es importante reconocer quién es tu verdadero <b>enemigo</b> para poder luchar de manera estratégica y efectiva.
<i>causar</i> (‘veroorzaken’)	VERB	1,007	3	Los bares y restaurantes de madera <b>causan</b> una impresión de poblado tradicional.	El exceso de velocidad puede <b>causar</b> accidentes graves en la carretera.
<i>político</i> (‘politiek’)	ADJ	1,237	5	Los de los partidos <b>políticos</b> acompañan a sus votantes en la otra vida.	<ul style="list-style-type: none"> <li>• Es importante estar informado sobre la situación político-social de un país para comprender su realidad y desarrollo.</li> <li>• El discurso <b>político</b> del presidente generó opiniones divididas entre la población.</li> <li>• La situación <b>política</b> en América Latina es muy complicada debido a diversos factores económicos y sociales.</li> <li>• El discurso <b>político</b> del presidente fue muy persuasivo y tuvo gran impacto en la opinión pública.</li> </ul>
<i>exacción</i> (‘heffing’)	NOUN	75.8	7	La investigación le atribuye presuntos delitos de cohecho, prevaricación, blanqueo de capitales y fraude y <b>exacciones</b> ilegales.	La <b>exacción</b> de impuestos a menudo genera debate y controversia en la sociedad.
<i>deslegitimar</i> (‘delegitimeren’)	VERB	206	9	Los independientes, a su modo de ver, <b>deslegitiman</b> y desnaturalizan la participación de los partidos”.	El periódico publicó un artículo que intentó <b>deslegitimar</b> las acusaciones contra el político.

Table 2: Dataset sample. “ID” refers to the subset ID. Values for the beginner group (subset 1-5) refer to the rank in Davies and Hayward Davies (2018); values for the advanced group (subset 6-10) refer to the Odds Ratio value.

“comment” column and annotate the other sentence as “best”.

- The sentence can be understood without any additional context (i.e. it is context-independent). If not, the participant should write “context-dependent” in the “comment” column and annotate the other sentence as “best”.
- The sentence does not contain words that are too difficult. If it does, the participant should write “too difficult” in the “comment” column and annotate the other sentence as “best”.

In case the example sentences adhered to all criteria, participants were instructed to indicate which sentence they found best (and worst in case of the BWS setup) based on their intuitions and needs as L2 learners. Regarding measures taken to arrive at qualitative annotations, we organised the first batch of ten annotations as an on-site session without any time constraints, allowing us to

provide guidance and answer questions whenever necessary. The remaining annotations could be completed at home. For their annotation work, the participants also received a financial compensation, serving as an additional incentive for them to complete the classification task diligently.

Finally, we checked if the sentences complied with the following formal criteria:

- The target item occurs in the sentence. If not, we label the other sentence as “best”<sup>6</sup>.
- The target item has the correct part of speech (POS). If not, we label the other sentence as “best”.
- The sentence is complete (i.e. it starts with capital letter and ends with punctuation mark). If not, we label the other sentence as “best”.

<sup>6</sup>Unless the target item does also not occur in that sentence, in which case we label both sentences as “N/A”.

## 4 Results

The results of the experiment have been summarised into a series of tables, listed below. The tables will be extensively referred to in our two main analyses: the comparison between corpus-based and GenAI-based as the source of the sentence (RQ1; Section 4.1) and the comparison between the different prompt types to generate the artificial example sentences (RQ2; Section 4.2).

- Table 3: results for pairwise comparisons (statistics)
- Table 4: results for pairwise comparisons (compliance with criteria)
- Table 5: results for BWS comparisons (statistics)
- Table 6: inter-annotator agreement (IAA) scores per subset

### 4.1 Comparison between GenAI-based and Corpus-based

As appears from Table 3, GenAI-based sentences are more frequently being found suitable than corpus-based sentences, with learners unanimously choosing the artificially generated sentence over the corpus-based one in 265 of the 400 pairwise comparisons (148/200 for the beginner group and 117/200 for the advanced group). In comparison, where the source is corpus-based, this value only amounts to 10/400. The moderate to substantial IAA scores (Table 6) for the corresponding subsets (between 0.62 and 0.72 for beginner and 0.55 and 0.65 for advanced) indicate that these annotations can be considered reliable, especially for the beginner group.

When looking at why corpus-based sentences are found less suitable than their GenAI-based counterparts, Table 4 reveals that – apart from a few cases where they contain the target item in a wrong POS (Example 1) – the corpus examples are less preferred mainly because they are (1) more context-dependent (Example 2, with *la otra* [‘the other’] being dependent on the preceding context) and (2) too difficult (e.g., *rugen* and *se abalanzan* in Example 3). In other words, the selection algorithm based on the SelEjemCor framework sometimes fails to meet the main criterion of *context independence* and the specific criterion of *difficult*

*vocabulary* (see Appendix A). Especially the context dependence of the corpus-based sentences (in 120 of the 400 sentences, i.e. 30%) can be considered an indication that selecting suitable examples from corpora at sentence level is a challenging task. Working at paragraph level might reduce this risk at context dependence (as paragraphs should constitute a more coherent unit of text), but will at the same time also increase the cognitive load and response time of the learning materials based on the examples.

1. Un parlamentario del **tripartito** puso como ejemplo de “buen funcionamiento” y “discreción” la comisión de investigación foral sobre el fraude de la Hacienda de Irún. (‘An MP of the tripartite gave as an example of “good functioning” and “discretion” the foral commission of enquiry into the fraud of the Irún Treasury.’) – Example taken from subset 6 for the adjective *tripartito*
2. Mercedes Alaya instruye ahora además la otra gran **macrocausa** andaluza: el fraude en los cursos de formación. (‘Mercedes Alaya is now also investigating the other big Andalusian mega lawsuit: the fraud in the training courses.’) – Example taken from subset 7 for the noun *macrocausa*
3. En invierno rugen los torrentes que se abalanzan montaña abajo, y el aire fresco agita las **ramas** de los robles. (‘In winter the torrents roar and rush down the mountain, and the fresh air stirs the branches of the oak trees.’) – Example taken from subset 1 for the noun *rama*

In the subsets with BWS evaluations (Table 5), we observe a similar trend: corpus-based examples are more frequently annotated as “worst” (28/50 times by all participants in the beginner group, 26/50 times in advanced) compared to artificially generated examples (2/50 in total for all GenAI prompt types in both beginner and advanced groups). Yet, even though the GenAI approach outperforms the corpus-driven approach by a large margin, Table 4 highlights that there is a non-negligible number of cases where the artificially generated sentences contain the target item in a wrong POS (3 instances in the beginner group, 7 in the advanced group; Example 4), consist of a definition (12 instances in the advanced

Subset	GenAI-based   Corpus-based			
	NOUN (/ 30)	VERB (/ 10)	ADJ (/ 10)	Total (/ 50)
1	23   0	7   0	8   1	38   1
2	25   0	6   1	8   0	39   1
3	24   0	6   0	6   0	36   0
4	22   1	7   1	6   0	35   2
Total	94   1	26   2	28   1	148   4
6	16   3	6   1	6   0	28   4
7	19   0	6   0	3   1	28   1
8	19   0	8   0	6   0	33   0
9	15   1	7   0	6   0	28   1
Total	69   4	27   1	21   1	117   6

Table 3: Statistics on example sentences annotated as “best” by all participants ( $N = 3$  for subsets 1-4 and  $N = 4$  for subsets 6-9) in pairwise comparison format. Results for the artificially generated sentences appear before the vertical line, results for corpus-based appear after.

	GenAI-based   Corpus-based			
	ZS-G	ZS-G+A	ZS-G+A+C	FEWSH
Beginner				
Definition	0   0	0   0	0   0	0   0
Context-dependent	0   16	0   13	0   15	0   17
Too difficult	1   11	1   12	0   11	0   9
No target item	0   0	1   0	0   0	0   0
Wrong POS	1   1	1   1	0   0	1   0
Incomplete	0   2	0   2	0   0	0   1
Advanced				
Definition	3   0	1   1	5   0	3   0
Context-dependent	1   14	0   16	1   15	1   14
Too difficult	3   23	0   21	0   19	2   23
No target item	0   0	0   0	0   0	0   0
Wrong POS	2   3	2   1	3   1	0   0
Incomplete	0   0	0   0	0   0	0   0

Table 4: Details on sentences that did not meet the suitability criteria defined in the annotation instructions, for the pairwise comparison subsets (see also Section 3.4). The number of sentences for GenAI-based appear before the vertical line, the number for corpus-based after the vertical line (on a total of 50, i.e. the number of sentences in a subset). “ZS-G” stands for the ZS-GEN prompt type, “ZS-G+A” for ZS-GEN+AUD, “ZS-G+A+C” for ZS-GEN+AUD+CRIT, and “FEWSH” for FEWSHOT.

		Full agreement   $\geq 1$ agreement				
		CORP	ZS-G	ZS-G+A	ZS-G+A+C	FEWSH
5 <sub>best</sub>	NOUN	0   2	0   19	4   16	0   16	0   13
	VERB	0   1	0   3	2   6	1   3	0   5
	ADJ	0   0	0   1	0   2	2   5	4   6
	Total	0   3	0   23	6   24	3   24	4   24
5 <sub>worst</sub>	NOUN	17   29	0   1	0   2	0   4	1   7
	VERB	7   9	0   2	0   1	0   0	0   1
	ADJ	4   9	0   2	0   3	0   2	1   2
	Total	28   47	0   5	0   6	0   6	2   10
10 <sub>best</sub>	NOUN	0   3	0   20	0   16	1   20	2   18
	VERB	1   1	1   7	0   5	1   4	0   4
	ADJ	0   1	1   5	0   5	1   7	0   6
	Total	1   4	2   32	0   26	3   31	2   28
10 <sub>worst</sub>	NOUN	17   27	0   3	1   4	0   1	0   3
	VERB	3   8	0   1	0   3	0   1	1   2
	ADJ	6   10	0   2	0   2	0   0	0   3
	Total	26   45	0   6	1   9	0   2	1   8

Table 5: Statistics on example sentences annotated as “best” and “worst” in subsets 5 (beginner target group) and 10 (advanced). “CORP” stands for corpus-based. The value before the vertical line refers to the sentences for which all of the participants ( $N = 3$  for subset 5 and  $N = 4$  for subset 10) agreed, the value after the vertical line reports the number of sentences for which at least one of the participants chose the sentence. The values in the “Total” rows are on a total of 50 (i.e. the number of sentences in a subset). “ZS-G” stands for the ZS-GEN prompt type, “ZS-G+A” for ZS-GEN+AUD, “ZS-G+A+C” for ZS-GEN+AUD+CRIT, and “FEWSH” for FEWSHOT.

Subset	IAA ( $\alpha$ )	ZS-G	ZS-G+A	ZS-G+A+C	FEWSH	ALL
1	0.7	✓				
2	0.72		✓			
3	0.66			✓		
4	0.62				✓	
5 <sub>best</sub>	0.29					✓
5 <sub>worst</sub>	0.61					✓
Avg	0.6	✓	✓	✓	✓	✓
6	0.6	✓				
7	0.58		✓			
8	0.55			✓		
9	0.65				✓	
10 <sub>best</sub>	0.22					✓
10 <sub>worst</sub>	0.71					✓
Avg	0.55	✓	✓	✓	✓	✓

Table 6: IAA scores – as computed by Krippendorff’s alpha ( $\alpha$ ) – for the annotation task in which L2 learners compare corpus-based sentences to artificially generated ones. “ALL” refers to subsets for which an example sentence based on each of the four different input prompts is generated. “Avg” rows report the average IAA value per target group. “ZS-G” stands for the ZS-GEN prompt type, “ZS-G+A” for ZS-GEN+AUD, “ZS-G+A+C” for ZS-GEN+AUD+CRIT, and “FEWSH” for FEWSHOT.

group; Example 5), or are found to be too difficult (2 instances in the beginner group, 5 in the advanced group; Example 6, with the word *desencadenó* being considered difficult by some of the advanced learners). This finding is also backed by the BWS evaluation results in Table 5, which show that there are 27/50 (beginner) and 25/50 (advanced) artificially generated examples annotated as “worse” by at least one of the learners (“ $\geq 1$  agreement”) in total across the four prompt types.

4. **Mañana** vamos a visitar el museo de arte moderno en el centro de la ciudad. (‘Tomorrow we are going to visit the museum for modern art in the city centre.’) – Example taken from subset 2 for the noun *mañana*
5. El **blanqueo** de dinero es un delito grave que involucra la transformación de dinero de origen ilícito en apariencia lícita. (‘Money laundering is a serious crime involving the conversion of money of an illegal nature into a lawful form.’) – Example taken from subset 6 for the noun *blanqueo*
6. La **destitución** del director desencadenó una crisis en la empresa que aún no se ha resuelto. (‘The dismissal of the director triggered a crisis in the company that has not yet been resolved.’) – Example taken from subset 9 for the noun *destitución*

#### 4.2 Comparison between Different Prompt Types

When comparing the full agreement results for the different GenAI prompts in Table 3, there is no noticeable difference (total scores range between 35/50 and 39/50 for the beginner group and between 28/50 and 33/50 for the advanced group). The only values which are slightly out of the ordinary are those for the adjectives in the advanced group: for ZS-GEN, ZS-GEN+AUD+CRIT, and FEWSHOT 6/10 sentences are annotated as “best” by all of the learners, while for the ZS-GEN+AUD prompt type this value only amounts to 3/10. Yet, this evidence is not substantial enough from which to draw conclusions, particularly because ZS-GEN+AUD obtains the top value (8/10) in the corresponding subset for the beginner group (subset 2, ADJ).

The results of the BWS evaluations (Table 5), however, paint a somewhat different picture. For

the beginner group, the full agreement scores show that specifying the target audience (ZS-G+A, 6/50 chosen as “best”) and the criteria (ZS-G+A+C, 3/50) has an added value compared to the broad context description (ZS-G, 0/50), just as providing the GenAI model with a few examples (FEWSH, 4/50). Nevertheless, when looking at the “ $\geq 1$  agreement” results, this difference disappears: 23/50 for ZS-GEN and 24/50 for the other three prompt types. Moreover, for the advanced group the ZS-GEN prompt type actually comes out as the arguably second-best prompt type with 2/50 full agreement and 32/50  $\geq 1$  agreement (compared to 0 and 26/50 for ZS-GEN+AUD, 3 and 31/50 for ZS-GEN+AUD+CRIT, and 2 and 28/50 for FEWSHOT). In other words, even though the BWS evaluations reveal somewhat more outspoken differences, these differences do not follow any clear pattern. This observation is also corroborated by the IAA scores, which are fairly low for the “best” annotations in subset 5 ( $\alpha = 0.29$ ; beginner group) and 10 ( $\alpha = 0.22$ ; advanced group).

## 5 Discussion

Regarding RQ1 (corpus-based versus GenAI as sentence source), the experiment has shown that, overall, L2 Spanish learners find artificially generated example sentences considerably more suitable than corpus-based sentences. The evaluation by the learners revealed that 30% of the corpus sentences were not fully comprehensible without further context. Put otherwise, GenAI methods seem most sensible to use for examples at sentence level, while corpus-based methods might be more suitable to retrieve items in a broader context, for example at paragraph level. However, the results also showed that in a number of cases the L2 learners did prefer the corpus-based example at sentence level, implying that exclusive reliance on GenAI to create sentence-level example sentences is not to be recommended. Moreover, even though the large language models used to generate the artificial examples are trained on large corpora, it is highly questionable if these sentences can be said to represent an authentic expression of language. Therefore, a third method which combines the best of both worlds might be worth considering: starting from a corpus-based example and using a GenAI model to rewrite it.

As for RQ2 (comparison between GenAI prompt types), the results were inconclusive:

adding a higher degree of specificity (by describing the target audience and the criteria the sentence should meet) did not result in any observable improvement compared to using a zero-shot prompt that only sketched the broad context. Opting for a few-shot prompt (i.e. providing a few examples the model can learn from) instead of a zero-shot prompt did not have any noticeable impact on the results either.

A first limitation of the study is that both the dataset size and the number of L2 learners evaluating the example sentences should be increased to arrive at more substantiated conclusions. Furthermore, even though the four different prompts provided considerable variation, more extensive prompt engineering could constitute a valuable avenue for further research, as would the comparison between different large language models for generating the artificial examples. Especially the choice between open-source (e.g., [Meta's Llama models](#)) and proprietary/closed-source models (e.g., OpenAI's GPT models) will become one of the most crucial methodological decisions, with the possibility to have a “peek under the hood” being weighed against performance levels and ease of use.

A third potential limitation is that – in the current setup – the target words may appear in a different linguistic construction (e.g., as a part of a collocation/colligation or not), meaning (e.g., literal versus metaphorical sense), or syntactic role (e.g., subject versus object position). It might be argued that differences in these aspects should be limited as much as possible, as they could have an impact on how easy or difficult it is for learners to understand the example sentences. Finally, the role of the texts from which the corpus-based sentences are chosen should also be analysed in further detail, for example by studying if compiling a specific corpus consisting exclusively of texts that have been written for users with a lower proficiency (e.g., from newspapers for children or adolescents) has a positive impact on the corpus-based scores for the beginner group.

## 6 Conclusion and Future Work

In this paper we compared corpus-based sentences to artificially generated sentences in terms of pedagogical suitability. We constructed a dataset containing 500 target items (250 vocabulary items to be taught to beginner learners and 250 to

advanced learners), for which we selected corpus examples according to a dedicated selection algorithm based on the [SelEjemCor](#) framework (Appendix A) and generated artificial examples by querying the GPT-3.5 Turbo large language model. The comparative evaluation of the sentences was performed by means of an experiment with seven students of L2 Spanish. The results of the experiment can be summarised into three main takeaways:

1. L2 learners find GenAI-based sentences considerably more suitable than corpus-based sentences. Of the 400 pairwise comparisons between corpus-based and GenAI-based sentences, 265 artificially generated examples were found suitable by all learners, compared to only 10 corpus-based examples.
2. Despite their excellent performance, the use of GenAI models has also shown to yield a number of unsuitable example sentences (with the target word in a wrong POS, the sentence being a definition instead of a usage example, or the sentence containing words that are too difficult).
3. A general zero-shot prompt describing the broad context of the task (i.e. the creation of example sentences for language learning purposes) provides enough information to create suitable example sentences. More specific prompts (describing the target audience and the criteria the sentence should meet) do not lead to better results, nor does formulating the prompt in a few-shot format (i.e. containing a few examples the model can learn from).

In potential follow-up experiments, the limitations discussed in Section 5 should be addressed, starting with increasing the number of target items and participants, evaluating the impact of using different corpora, and applying more extensive prompt engineering based on techniques for educational purposes in general ([Cain, 2024](#)) and for L2 learning purposes in particular ([Isemonger, 2023](#)). To convert the experimental design adopted in the current study into a more “controlled environment”, testing different GenAI models with the same prompts or using designated platforms such as [LMStudio](#) are options worth considering. Additionally, fine-tuning the annotation instructions



(e.g., by adding an explicit evaluation of the grammatical soundness and syntactic properties of the sentence) would allow us to gain more in-depth insights into the exact reasons why one example sentence is preferred over another.

Furthermore, as hinted at in the discussion (Section 5) as well, developing a new method that combines a corpus-based and GenAI-based approach constitutes another important topic for future research. In such a “hybrid” method, authentic corpus-based examples can be taken as the starting point and GenAI models can be used as the means to rewrite the examples in order to make them meet the required criteria, especially regarding context independence and difficulty. Different types of rewriting prompts could be compared, from zero shot over few shot to retrieval-augmented generation (in which we let the model “look for” the most relevant information in large set of corpus examples and then prompt it to generate new examples based on this information). Yet, our (preliminary) finding that the corpus-based method (yielding *authentic* example sentences) is being outperformed by the GenAI-based one (yielding *artificial* examples) can also be considered a reason to bring that other source of non-authentic examples, the invented example (IE; Section 2.1), back into the equation. Conducting an experiment in which IEs are compared to artificially generated sentences could shed renewed light on the role IEs can play in an L2 setting.

## 7 Acknowledgements

This research has been carried out as part of a PhD fellowship on the IVESSE project (file number 11D3921N), funded by the Research Foundation - Flanders (FWO). Additionally, we want to express our sincere gratitude to the reviewers for their valuable feedback and suggestions.

## References

- Beryl T. S. Atkins and Michael Rundell. 2008. *The Oxford guide to practical lexicography*, 1 edition. Oxford linguistics. Oxford University Press, Oxford.
- Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.
- William Cain. 2024. [Prompting Change: Exploring Prompt Engineering in Large Language Model AI and Its Potential to Transform Education](#). *TechTrends*, 68(1):47–57.
- Andrew Caines, Luca Benedetto, Shiva Taslimipour, Christopher Davis, Yuan Gao, Oeistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, Marek Rei, Helen Yannakoudakis, Andrew Mullooly, Diane Nicholls, and Paula Buttery. 2023. [On the application of Large Language Models for language teaching and assessment technology](#).
- G. Cook. 2001. [‘The philosopher pulled the lower jaw of the hen’](#). Ludicrous invented sentences in language teaching. *Applied Linguistics*, 22(3):366–387.
- Peter Crosthwaite and Vit Baisa. 2023. [Generative AI and the end of corpus-assisted data-driven learning? Not so fast!](#) *Applied Corpus Linguistics*, 3(3):100066.
- Mark Davies and Kathy Hayward Davies. 2018. *A frequency dictionary of Spanish: Core vocabulary for learners*, 2 edition. Routledge frequency dictionaries. Routledge, London ; New York.
- Nick C. Ellis. 2006. [Language Acquisition as Rational Contingency Learning](#). *Applied Linguistics*, 27(1):1–24.
- J.R. Firth. 1968. *Selected Papers of J.R. Firth*. Longman, London; Harlow.
- A. Frankenberg-Garcia. 2012. [Learners’ Use of Corpus Examples](#). *International Journal of Lexicography*, 25(3):273–296.
- Ana Frankenberg-Garcia. 2014. [The use of corpus examples for language comprehension and production](#). *ReCALL*, 26(2):128–146.
- Ana Frankenberg-Garcia, Geraint Paul Rees, and Robert Lew. 2021. [Slipping Through the Cracks in e-Lexicography](#). *International Journal of Lexicography*, 34(2):206–234.
- Costas Gabrielatos. 2018. Keyness analysis: Nature, metrics and techniques. In C. Taylor and A. Marchi, editors, *Corpus Approaches To Discourse: A critical review*, pages 225–258. Routledge, Oxford.
- Patrick Goethals. 2018. Customizing vocabulary learning for advanced learners of Spanish. In *Technological innovation for specialized linguistic domains : languages for digital lives and cultures, proceedings of TISLID’18*, pages 229–240, Gent, Belgium. Éditions Universitaires Européennes.
- Maldonado González, editor. 2012. *Diccionario Clave: diccionario de uso del español actual*, 9 edition. SM, Boadilla del Monte (Madrid).
- Stefan Th. Gries. 2013. 50-something years of work on collocations: What is or should be next . . . . *International Journal of Corpus Linguistics*, 18(1):137–166.

- Tanja Heck and Detmar Meurers. 2022. [Generating and authoring high-variability exercises from authentic texts](#). In *Proceedings of the 11th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL 2022)*, pages 61–71.
- Ian Isemonger. 2023. [Generative Language Models in Education: Foreign Language Learning and the Teacher as Prompt Engineer](#). *TEFL Praxis Journal*, 2:3–17.
- Jelena Kallas, Adam Kilgarriff, Kristina Koppel, Elgar Kudritski, Margit Langemets, Jan Michelfeit, Maria Tuulik, and Ülle Viks. 2015. Automatic generation of the Estonian Collocations Dictionary database. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, pages 11–13.
- Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, volume 1, pages 425–432. Universitat Pompeu Fabra Barcelona.
- Lucas Kohnke, Benjamin Luke Moorhouse, and Di Zou. 2023. [ChatGPT for Language Teaching and Learning](#). *RELC Journal*, 54(2):537–550.
- Iztok Kosem, Kristina Koppel, Tanara Zingano Kuhn, Jan Michelfeit, and Carole Tiberius. 2019. [Identification and automatic extraction of good dictionary examples: the case\(s\) of GDEX](#). *International Journal of Lexicography*, 32(2):119–137.
- Batia Laufer. 1992. Corpus-based versus lexicographer examples in comprehension and production of new words. In *Proceedings of the Fifth Euralex International Congress*, pages 4–9. University of Tampere.
- Batia Laufer and Karen Shmueli. 1997. [Memorizing New Words: Does Teaching Have Anything To Do With It?](#) *RELC Journal*, 28(1):89–108.
- I.S.P. Nation. 2022. *Learning Vocabulary in Another Language*, 3 edition. Cambridge University Press.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Žeman. 2016. [Universal Dependencies v1: A Multilingual Treebank Collection](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).
- Ildikó Pilán, Elena Volodina, and Lars Borin. 2016. Candidate sentence selection for language learning exercises: From a comprehensive framework to an empirical evaluation. *Revue Traitement Automatique Des Langues*, 57(3):67–91.
- Punjaborn Pojanapunya and Richard Watson Todd. 2018. [Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis](#). *Corpus Linguistics and Linguistic Theory*, 14(1):133–167.
- John McHardy Sinclair. 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. Collins ELT, London.
- Simon Smith, P.V.S. Avinesh, and Adam Kilgarriff. 2010. Gap-fill tests for language learners: Corpus-driven item generation. In *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*, pages 1–6. Macmillan Publishers India.
- Anatol Stefanowitsch. 2020. *Corpus linguistics: A guide to the methodology*. Number 7 in Textbooks in language sciences. Language Science Press, Berlin.

## Appendices

### Appendix A. SelEjemCor framework

The criteria included in the SelEjemCor are presented in Table 7. The values set for obtaining the example sentences in the experiment are included in the “V<sub>set</sub> BEG” and “V<sub>set</sub> ADV” columns.

### Appendix B. Prompt types

The different prompt types used in the experiment are presented in Table 8.

Nr	Criterion	V <sub>set</sub> <b>BEG</b>	V <sub>set</sub> <b>ADV</b>
/	Proficiency level target audience.	B1	C1
/	Number of years experience target audience.	1	3
<b>1</b>	Boolean value indicating if search term has to occur in sentence.	<i>True</i>	<i>True</i>
<b>2</b>	Numerical value indicating maximum number of times search term can occur in sentence.	1	1
<b>3</b>	Numerical value between 0 and 1 indicating at which position search term has to occur.	<i>None</i>	<i>None</i>
<b>4</b>	Boolean value indicating if sentence has to contain dependency root.	<i>True</i>	<i>True</i>
<b>5</b>	Boolean value indicating if sentence has to contain subject or finite verb.	<i>True</i>	<i>True</i>
<b>6</b>	Boolean value indicating if sentence has to contain explicit subject.	<i>True</i>	<i>True</i>
<b>7</b>	Boolean value indicating if sentence has to start with capital letter and end with punctuation mark.	<i>True</i>	<i>True</i>
<b>8</b>	Numerical value indicating maximum number of tokens which do not occur in <b>SCAP-based lemma lexicon</b> .	0	1
<b>9</b>	Numerical value indicating maximum number of non-alphabetical tokens (e.g., mark-up traces in web materials).	0	0
<b>10</b>	Boolean value indicating that no conjunction or subjunction can appear in sentence-initial position.	<i>True</i>	<i>True</i>
<b>11</b>	Numerical value indicating maximum number of demonstrative pronouns (e.g., <i>este esta</i> : ‘this’; <i>ese esa</i> : ‘that’).	0	0
<b>12</b>	Numerical value indicating maximum number of words/phrases which occur in precompiled list of anaphoric expressions (e.g., <i>allí</i> : ‘there’; <i>aquí</i> : ‘here’; <i>entonces</i> : ‘then’).	0	0
<b>13</b>	Numerical value indicating maximum number of negation adverbials (e.g., <i>no</i> : ‘no’; <i>nadie</i> : ‘nobody’; <i>nada</i> : ‘nothing’).	0	0
<b>14</b>	Boolean value indicating that sentence cannot represent direct question.	<i>True</i>	<i>True</i>
<b>15</b>	Boolean value indicating that sentence cannot represent direct speech (i.e. speaking verb combined with delimiters such as quotation marks).	<i>True</i>	<i>True</i>
<b>16</b>	Boolean value indicating that sentence cannot represent answer to closed question (i.e. sentence-initial adverb of affirmation or negation followed by delimiter).	<i>True</i>	<i>True</i>
<b>17</b>	Numerical value indicating maximum number of tokens which occur in precompiled list of modal verbs (when functioning as an auxiliary verb).	1	3
<b>18</b>	Numerical value indicating maximum number of tokens in the sentence (including punctuation).	10-30	10-30
<b>19</b>	Numerical value indicating maximum number of words above the proficiency level of the target audience according to a personalised machine learning classifier.	0	0
<b>20</b>	Numerical value indicating minimum frequency of words in <b>SCAP lemma frequency dictionary</b> (expressed in percentiles).	P90	P75
<b>21</b>	Numerical value indicating maximum number of words not included in <b>SCAP token lexicon</b> .	0	1
<b>22</b>	Boolean value indicating that sentence cannot contain words which occur in <b>precompiled list of potentially sensitive words</b> related to PARSNIP topics.	<i>True</i>	<i>True</i>
<b>23</b>	Numerical value indicating maximum number of proper names.	2	2
<b>24</b>	Numerical value indicating minimum average normalised Lexicographer’s Mutual Information (Bouma, 2009) score for verb-noun pairs (in subject, object, and oblique relation) and all noun-adjective pairs (in attributive or predicative relation) in the sentence. The scores are retrieved from a <b>SCAP-based resource</b> .	<i>all</i>	<i>all</i>
<b>25</b>	Numerical value indicating minimum average $\Delta P$ score (Ellis, 2006; Gries, 2013) for verb-noun pairs (in subject, object, and oblique relation) and all noun-adjective pairs (in attributive or predicative relation) that include the search term. The scores are retrieved from a <b>SCAP-based resource</b> .	<i>all</i>	<i>all</i>
<b>26</b>	Numerical value indicating minimum average cosine similarity of search term with head and dependents (both static and contextualised word embeddings).	<i>all</i>	<i>all</i>
<b>27</b>	Numerical value indicating minimum average $n$ -gram frequency of the sentence (excluding $n$ -grams with punctuation marks). Frequencies are retrieved from <b>SCAP dictionary</b> containing lemma-based $n$ -grams.	<i>all</i>	<i>all</i>

Table 7: Criterion descriptions and Values set for SelEjemCor criteria. Filters are put in bold, rankers in plain text. “BEG” and “ADV” refer to the beginner and advanced target groups respectively.

Prompt ID	Prompt text
SYS1	You are a teacher of Spanish as a foreign language.
SYS2	You are a teacher of Spanish as a foreign language to a beginner/lower-intermediate group of university students who have been studying Spanish for one year.
SYS3	You are a teacher of Spanish as a foreign language to an upper-intermediate/advanced group of university students who have been studying Spanish for three years.
USR1	Write a sentence between 10 and 30 words in Spanish that presents an authentic usage of the Spanish [POS] '[WORD]', a vocabulary item that has to be learnt by your students. The sentence should not be a definition of the word.
USR2	Write a sentence between 10 and 30 words in Spanish that presents an authentic usage of the Spanish [POS] '[WORD]', a vocabulary item that has to be learnt by your students. The sentence should not be a definition of the word. The sentence should be well-formed and context-independent, it should be tailored to the proficiency level of your students, and it should contain phrases that frequently co-occur with the target item '[WORD]'.
USR3	Write a sentence between 10 and 30 words in Spanish that presents an authentic usage of a Spanish vocabulary item that has to be learnt by your students: word=diseño; part of speech=noun; sentence=Para hacer un buen diseño de un mueble hay que pensar en su utilidad. ### word=comprometer; part of speech=verb; sentence=Sus revelaciones comprometían en el caso de corrupción a otras dos organizaciones. ### word=dramático; part of speech=adjective; sentence=Toda la prensa se hace eco del dramático caso de la niña desaparecida. ### word=[WORD]; part of speech=[POS]; sentence=

Prompt type	System role	User role	Subset
Beginner			
ZS-GEN	SYS1	USR1	1
ZS-GEN+AUD	SYS2	USR1	2
ZS-GEN+AUD+CRIT	SYS2	USR2	3
FEWSHOT	SYS2	USR3	4
Advanced			
ZS-GEN	SYS1	USR1	6
ZS-GEN+AUD	SYS3	USR1	7
ZS-GEN+AUD+CRIT	SYS3	USR2	8
FEWSHOT	SYS3	USR3	9

Table 8: Detailed overview of prompt types used to generate artificial example sentences.

# Potential of ASR for the study of L2 learner corpora

**Sarra El Ayari**

Structures Formelles du Langage  
CNRS & Paris 8 University  
sarra.elayari@cnrs.fr

**Zhongjie Li**

Structures Formelles du Langage  
CNRS & Paris 8 University  
lzh44010@gmail.com

## Abstract

This study is at the crossroads of Natural Language Processing (NLP) and Second Language Acquisition (SLA). We used Whisper’s speech recognition on a French L2 learner corpus to get automatic transcripts, and compared them with pre-existing manual transcripts. We then conducted quantitative and qualitative analysis of the issues which are inherent to the specificities of interlanguage for any automatic tool. We will discuss the different issues encountered by Whisper that are specific to learner corpora.

## 1 Introduction

The TranSLA project aims at analyzing to which extent Automatic Speech Recognition systems (ASR) can provide useful information on the distance between interlanguage and the internalized norm of those systems. Providing tools for corpus linguistics is an essential part of the research carried out in Second Language Acquisition (SLA). Recent technological advances raise new methodological questions. The act of transcribing involves an initial task of interpreting the discourse in L2, which is particularly delicate since it can influence the researcher’s subsequent analysis (Benazzo and Watorek, 2021).

If the results obtained for speech recognition in general are very encouraging (Radford et al., 2023), we still need to be able to evaluate precisely their performance on non-standard languages, such as interlanguage of foreign learners (Selinker, 1972). Interlanguage is the idiolect developed by second language learners and it refers to the mental grammar constructed by a learner at a specific stage of the learning process (Ellis and Barkhuizen, 2005). It is therefore intrinsically

subject to variation and evolution simultaneously and possesses a unique linguistic organization.

This study aims firstly at measuring the performance of an ASR system on a L2 French learner corpora, and secondly to observe if ASR systems could be used as a tool to evaluate how close or distant learners speech productions can be from the language model that is used, and therefore to correlate it with learners’ acquisition levels. We will discuss the discrepancies linked to SLA issues as well.

## 2 Transcription of learner corpora

The transcription process is a time-consuming phase for any researcher who wants to work on audio or multimodal data. It is also a very precise work that requires already to have clear thoughts about which linguistics phenomena will be analyzed, and therefore which elements have to be transcribed and how.

In Second Language Acquisition, this process is even more important because fine-grained access to information is crucial. To transcribe exactly what the learners are actually saying and pronouncing is the goal - even if it is not always attainable. In that way, how to transcribe is already a choice. It is even more complicated when the language has a wide gap between oral and written modalities, like French (Blanche-Benveniste, 2000). Thus choosing one form over the others carries the risk of over-estimate or under-estimating the knowledge of the learner (Benazzo and Watorek, 2021).

We present a few examples from the ESF (*European Science Foundation Second Language*) corpus (Perdue, 1993) which shows the problems of choosing a specific form for transcription in Figure 1.

The transcriptions presented (El Ayari and Wa-

---

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

- /ʒəparle/ vs. /ʒeparle/ => 'je parlais' vs. 'j'ai parlé'  
 - /ʒəse/ vs. /ʒeseje/ => 'j'essaie' vs. 'j'ai essayé'  
 - /ilese/' vs. /illese/ => 'il essaie' vs. 'il l'essaie' ou 'il les sait'  
 - /ilavole/ vs. /illavole/ => 'il a volé' vs. 'il l'a volé'

Figure 1: Examples of transcriptions

torek, 2021) show that transcribing a corpus is already choosing which linguistic form the learner has pronounced, even though we do not always have enough knowledge to decide.

### 3 State of art

Not many studies focused on ASR systems' performances on non-native languages. It is a very important matter as those systems have been trained predominantly on standard varieties (Graham and Roll, 2023). Studies focusing on learner corpora and ASR mostly focus on the global evaluation of ASR performances: (Graham and Roll, 2023; Cumbal et al., 2021) on Swedish or on providing pronunciation feedback with a focus on phonetic features: (Wei et al., 2022) on Dutch, (Ballier et al., 2023; Chanethom and Henderson, 2022) on French. We did not find any studies intertwining ASR performances and learners' proficiency.

### 4 The corpus

The LANGSNAP corpus<sup>1</sup> is based on the study abroad of 29 advanced learners of L2 French (mit) (eleven years length of French study). The learners are L1 English speakers and Anglophone university students, learning French over a 21-month period, including a 9-month stay abroad. This analysis is based on 14 participants. The audio data as well as the transcriptions are freely available on Talkbank<sup>2</sup> (CLARIN Knowledge Centre), an open access integrated repository for spoken language data.

The LANGSNAP corpus is longitudinal and therefore offers a good basis to compare the oral productions of the learners at different times. There are different linguistic tasks available: oral interviews (where participants took part in a semi-structured interview led by a member of the research team); story retelling (where participants retold a story guided by a sequence

<sup>1</sup>LANGSNAP: <https://web-archiver.southampton.ac.uk>

<sup>2</sup>Talkbank: <https://www.talkbank.org>

of pictures); argumentative writing (where participants wrote a timed 200-words response to a stimulus question). We chose to analyze the oral interviews where participants took part in a semi-structured interview led by a member of the research team, which have been conducted regularly through the project. We analyzed data at different times: October 2011 in stay abroad (T1), May 2012 in stay abroad (T2) and October 2012 post stay abroad (T3). The interviews have already been manually transcribed in chat format (MacWhinney, 2000), with speech alignment. The corpus contains also the same oral interviews performed by French native speakers manually transcribed too, which we will use as a baseline for the ASR performances on native French.

Examples of utterances:

- (1) alors pour commencer décris moi où tu habites et les gens avec qui tu habites ?
- (2) &-euh j'habite à City donc c'est une ville vers &-euh l'ouest <de la France>[/] &-euh de Paris.

This corpus is ideal for looking at the evolution of the interlanguage of the learners (Corder, 1980), as they have produced the same tasks at different times and as the data have been transcribed and analyzed beforehand. The data have a good audio quality without background noises, which is also something important to take into consideration for an automatic analysis.

### 5 Methodology

Our methodology consists in comparing the transcriptions obtained automatically by Whisper<sup>3</sup> to the ones produced manually to see precisely the differences and to evaluate the performance of the ASR system in general on this corpus.

#### 5.1 ASR system

We used the ASR system Whisper, created by OpenAI. Our choice of ASR is a pragmatic one as Whisper is the only one freely available on governmental servers by the IR Huma-Num<sup>4</sup> and the CINES<sup>5</sup> in France (release 20231117). It has been

<sup>3</sup>Whisper: <https://github.com/openai/whisper>

<sup>4</sup>Huma-Num: <https://www.huma-num.fr>

<sup>5</sup>CINES: <https://www.cines.fr>

trained on French dataset, and therefore can produce speech recognition task and automatic transcriptions of oral data.

“Whisper architecture is a simple end-to-end approach, implemented as an encoder-decoder Transformer. [...]. A decoder is trained to predict the corresponding text caption, intermixed with special tokens that direct the single model to perform tasks such as language identification, phrase-level timestamps, multilingual speech transcription” (Radford et al., 2023).

The challenge here is to see how well the system performs on the particular oral data that are learner oral productions. As different linguistics levels are in the process of being acquired, the transitional aspect of interlanguage offers difficulties for any type of automatic process. Pronunciation, vocabulary, morphology and syntax will not be standard. As such, learner corpora can be considered as one type of less-resourced language, and specific resources might be needed to process them accurately.

## 5.2 Evaluation metrics

Different metrics can be used to evaluate ASR systems. WER (*Word Error Rate*) evaluates the proportion of correct words compared to manual transcripts, while the CER (*Character Error Rate*) measures the proportion of correct characters. Both metrics are commonly used to quantify ASR performance. We are aware that those metrics have limitations such as only taking into account the word level and therefore not pondering the results linked to semantic similarity. Nevertheless they offer us a global metric to evaluate Whisper’s performances despite the evolving nature of L2 data and interlanguage. We wanted to get a global overview of the results across time for a semi-control task. Nevertheless, we will deepen the analyse by looking closely at Whisper’s corrections: insertions, substitutions and deletions in order to get a better understanding of the results. We did not look into Part-Of-Speech Error Rate because of the nature of the data, and particularly the bias created by the meaning idiosyncrasy where a form used by a learner does not imply that its linguistics function is also mastered (*proximity fallacy* (Perdue, 1993)).

## 5.3 Data processing

Our goal is to provide parallel corpora in order to compare manual and automatic transcripts. Figure 2 shows the pipeline for files normalization, in order to be able to compare the transcriptions.

The manual transcripts are in chat format, which belongs to the CLAN program (*Computerized Language Analysis*) (MacWhinney, 2000). The speakers are introduced by a code and an asterisk and a pos-tagging has been automatically generated (line %mor) as shown on Figure 3.

We have encountered different issues during the process of the data. The first difficulty encountered when processing the transcriptions is the turn-taking. Long turns of speech are cut into several lines so it was difficult to combine the lines together in order to compare them. Secondly, as manual transcriptions have been done by different transcribers at different times, human errors and changes in the transcription guide had to be taken into consideration and were lacking regularities.

Another issue is linked to Whisper itself which creates bugs increasing the WER score of automatic transcriptions. The first bug relates to language changes detected in the middle of a French transcription. In the examples below, the transcription alternates between several languages and is not pronounced by the speaker at all (extra-hallucinated errors).

- (3) Euh... Ça lui soulage la sensation. J is subi au passé. tardised bear. Hope you are okay now. Jean Apple. Brooke de Ney sur une locale.

The second type of bug concerns the repetition of a word or words in several lines. Here, as illustrated in the following example, the word “oui” is reproduced in several lines, which is not the case in the audio file. Word repetition degrades the quality of the automatic transcription by adding non-existent words or replacing several turns of speech.

- (4) où il y a des élèves un peu difficiles. Oui. Il n’y a pas beaucoup... Une prof m’a dit qu’elle a... Oui. Oui. Oui. Oui. Oui. Oui.

The third type of bug is that Whisper fails to detect speech for certain audio sequences, leaving

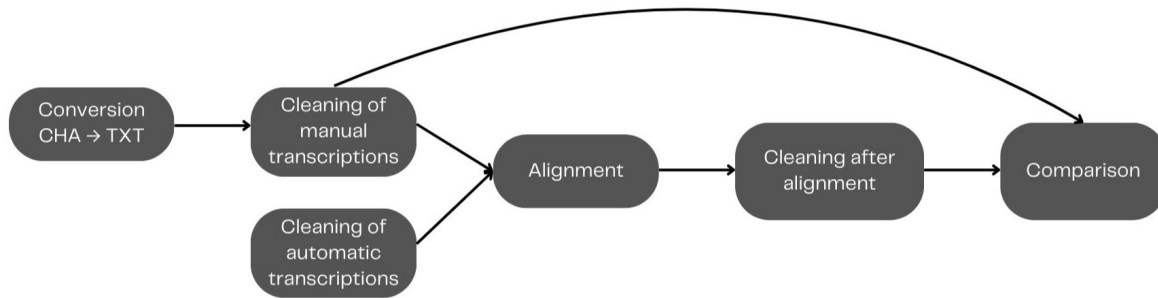


Figure 2: Pipeline

```

*100: donc c'est une ville vers &-euh l'ouest <de la France> [/ /]
&-euh de Paris . •12032_16266•
%mor: adv|donc pro:dem|ce$V:aux|être&PRES&3s det:art|une&f&sg
n|ville&f prep|vers det:art|le$N|ouest&m prep|de n:prop|Paris .
*KMCM: hum hum . •16266_17279•
%mor: co|hun co|hun .
*100: j'habite avec ma tante et &-euh voilà . •17279_20299•
%mor: pro:sub|je$V|habiter-IMP&2s prep|avec det:poss|ma&sg
n|tante&f conj|et adv:place|voilà .
  
```

Figure 3: Manual transcription format

white spaces instead - which requires also manual intervention in order to preserve the alignment between the two transcripts.

Most of the normalization process has been done automatically with python scripts, after analyzing the data. Nevertheless because of Whisper's irregularities, we had to manually check and sometimes manually correct the utterances to guarantee the accuracy of the alignment.

#### 5.4 Harmonization of the data

The preparation of the data for metrics calculation has been done in two steps.

Manual transcriptions have been done by different transcribers and therefore do not always follow the same conventions. We had to address this, especially as oral conventions in Clan can have different formats. Different elements had to be removed, such a speaker codes, timestamps, punctuation and characters used for idioms. A conversion to lowercase has been done. A specific treatment on the numbers to convert them into words, as well as for the time. Table 1 illustrated the problems encountered to be able to compare the transcripts as accurately as possible.

#### 5.5 Transcriptions' comparison

Transcription is the first stage in the study of any oral corpus and, as such, it implies theoretical

choices. We assume that "learner varieties are not imperfect imitations of a 'real language' - the target language - but systems in their own right, error-free by definition" (Klein and Perdue, 1997). Indeed, there are notable differences, both quantitative and qualitative linguistic behavior between native and non-native languages (Dekydtspotter et al., 2006). For those reasons, two important points need to be kept in mind:

- **comparative fallacy** (Bley-Vroman, 1983): learners' language explained by reference to the target language system rather than as set of rules and performance characteristics ;
- **closeness fallacy** (Perdue, 1993) : learners' language explained by attributing references of the target language on the basis of their formal resemblance.

Those two elements are likely to cause difficulties for automatic tools processing on learner corpora.

We developed a framework to visualize both results at the same time, and automatically highlight and categorize the differences: elements inserted, replaced or deleted, and to be able to check the audio file for each utterance. The Figure 4 shows an excerpt of the interface.

o102 LANGSNAP/ abroad1	5	reponse [elle] - [is]	elle: sont très sympas	is: sont très sympas	0.25	Play
o102 LANGSNAP/ abroad1	6	déline [enb] - [ ] [enb] - [ ] [enb] - [ ]	et enb ou j'ai je partage enb mon douche avec une autre	et ou je partage mon douche avec une autre	0.25	Play
o102 LANGSNAP/ abroad1	7	reponse [ex douche] - [coudouche] [sonsi] - [sonsi]	donc j'ai une ex douche qui est très sonsi et très sympa aussi	donc j'ai une coudouche qui est très sonsi et très sympa aussi	0.23076923076923078	Play

Figure 4: Interface of the comparison framework



Issues	Manual transcripts	Correction
Speakers code	<b>*109:</b> mais je suis pas sûre	mais je suis pas sûre
Timestamps	je suis pas sûre . <b>137805_139862</b>	je suis pas sûre
Compound words	rez + de +chaussé	rez de chaussé
Type case	j'habite au <b>City</b>	j'habite au <b>city</b>
Disfluencies	je suis content <b>&amp;-euh</b> ici	je suis content ici
Punctuations	Oui , et où ?	Oui et où
Numbers	environ <b>3</b> minutes	environ <b>trois</b> minutes
Time	à <b>1h30</b> du matin	à <b>une heure et demie</b> du matin

Table 1: Transcripts' harmonization

## 5.6 Evaluation measures

WER metric, derived from Levenshtein's distance, provides a score based on the number of incorrectly transcribed words. The higher the score, the lower the similarity between the documents being compared similarity. CER metric indicates the percentage of characters that were incorrectly predicted. They are defined by the ratio between the number of incorrectly aligned words/characters and the total number of words/characters in the reference transcript:

$$WER|CER = \frac{s + i + d}{n}$$

where s, i and d are the number of substitutions, insertions and deletions and n is the total number of words/characters in the reference transcript. They both measure the overall word/character recognition performance without distinguishing between fluent and disfluent words (Lou and Johnson, 2020). Both calculations have been done with Python and the JiWER package<sup>6</sup>.

## 5.7 Speech disfluencies

Speech disfluencies are non-pathological hesitations happening during speaking, like the use of fillers ("like" or "uh") or the repetition of a word or phrase. Unfortunately, "for faithful transcription of conversational speech, there remain challenges both in terms of the content predicted by [transformer based] models (hallucinations, unintended normalization of disfluencies and transcriptions of background noises) and in terms of alignment accuracy" (Yamasaki et al., 2023). The main reason being that the models of ASR systems are trained on fluent (and native) speech, the mismatch between training data and other

types of corpora decreases their performance (Lou and Johnson, 2020).

## 6 Results

In this section, we will be comparing the two types transcriptions: manual (MT) versus automatic (AT). Results are better for natives, a type of speech closer to the ones Whisper has been trained on - especially if we remove the speech disfluencies. Taking those into account make a real difference in the calculation of WER and CER for audio corpora, in tasks such as interviews where speakers are speaking freely and answering questions.

	+ disfluencies		- disfluencies	
Corpus	WER	CER	WER	CER
L-T1	0.31	0.25	0.25	0.19
L-T2	0.35	0.26	0.29	0.23
L-T3	0.28	0.21	0.22	0.18
Natives	0.36	0.28	0.23	0.17

Table 2: WER and CER measurements

A WER score between 0.1 and 0.2 is considered as good. The results without disfluencies, especially for natives and learners after stay abroad are good for that kind of corpora. We can conclude that Whisper's performances on the LANGSNAP corpus, for native speakers and advance learners are very decent.

### 6.1 Longitudinal scores

Our second research question concerns the hypothesis that ASR evaluation metrics can be correlated with learners' proficiency and should therefore decrease as learners get closer to the French speech Whisper has been trained on.

<sup>6</sup>JiWER: <https://pypi.org/project/jiwer>

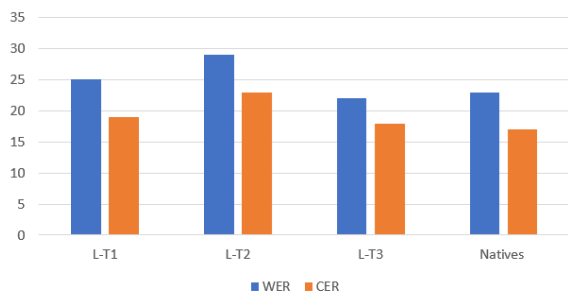


Figure 5: Longitudinal WER & CER metrics

As shows Figure 5, both WER and CER metrics get lower as the learners improve their knowledge of French, and get closer to the results obtained on the natives speakers. This result is consistent with the improvement of the learners and their acquisition level in general.

If the results between T1 and T3 are decreasing (WER at T1: 0.25 / WER at T3: 0.22), we can also see that they are increasing at T2. To explain this phenomena from an acquisitional point of view, we can point out the critical rule hypothesis stated by W. Klein (Klein, 1989). The idea is that a linguistic rule inside interlanguage is not definitive and therefore is subject to change and evolve. So it could be possible that T2 would represent a specific acquisitional time where rules acquired by learners during language courses would evolve through the stay abroad, because of direct input from native speakers and that some linguistic rules would later be acquired in T3.

## 6.2 Overview of ASR process

In order to get a better understanding of the ASR results and correlate them to learners' proficiency, we took a closer look on the substitutions, insertions and deletions performed by Whisper. The Figure 6 shows the percentage of those three processes for each times.

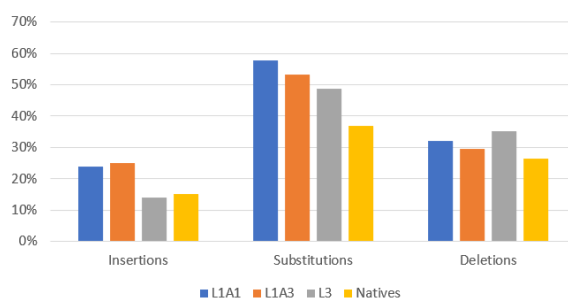


Figure 6: ASR processes

We will get a closet look to each of those three processes by comparing manual transcripts (MT) to automatic ones (AT).

### 6.2.1 Insertions

Insertions are what we define here as hyper-corrections or hyper-normalizations of the learners' speech.

- (5) a. MT: je dois je me dis toujours c'est une expérience
- b. AT: je dois je me dis toujours **que** c'est une expérience

Here Whisper adds a subordinating conjunction to the speaker's utterance.

### 6.2.2 Substitutions

Substitutions are mostly linked to morphology: number, gender, definiteness and verb tenses. French has a relatively complex orthography (van den Bosch et al., 1994) and contains a large number of silent letters which correspond to morphological markers, which make the transcription's process even more difficult.

- (6) a. MT: j'ai imaginé que institue **était** à paris
- b. AT: j'ai imaginé que j'ai institue **été** à paris

- (7) a. MT: je me suis **inscrit** pour faire le marathon
- b. AT: je me suis **inscrite** pour faire le marathon

- (8) a. MT: danse **aérobic**
- b. AT: danse **aérobique**

Most of the insertions are linked to negative forms:

- (9) a. MT: je l'aime pas
- b. AT: je **ne** l'aime pas

### 6.2.3 Deletions

Deletions are mostly about Whisper not processing normal speech disfluency, where people can repeat something twice while hesitating or thinking what to say next:

- (10) a. MT: je fais la permanence du soir qui est **jusqu'au jusqu'à** dix neuf heures  
b. AT: je fais la permanence du soir qui est **jusqu'à** dix neuf heures

We also encounter deletions where the learner produces a non-canonical form that is corrected by Whisper by deleting a character, such as contractions and speech pauses:

- (11) a. MT: j'ai imaginé **que à** la bibliothèque je rencontrerais beaucoup **des** gens  
b. AT: j'ai imaginé **qu'à** la bibliothèque je rencontrerais beaucoup **de** gens

Those three phenomena are expected in oral treatments. They show that Whisper has difficulties with elements linked to spontaneous speech, such as hesitations, repetitions, disfluencies, contractions. Those examples also show that it is difficult for the system to provide utterances that are not following a specific format, even when the pronunciation differs - like changing the words' order. There are typical corrections that one has to correct back in order to access interlanguage properly.

### 6.3 Specific SLA issues

Whisper tends to (hyper-)normalize the speech of the learners: Table 3 shows a few examples which are problematic when one is studying learners' productions for different reasons.

Those issues are extremely problematic for researchers who work in the SLA field, because it does not provide enough accuracy. The issues are linked to different specificities of a learner speech in L2 as pronunciation, prosody, fluency, pauses, morphology, syntax and different idiosyncrasies. The item **expérencier** for example is very important to acknowledge because it is a clear hint of the acquisition of verbal morphology from the learner.

Whisper rewrites the data according to the language's model deduced from the training dataset

but does not provide a systematic treatment, as show the examples below from the same learner:

- (12) a. MT: mais **le FLE** est vraiment similaire de le cours français  
b. AT: mais **le fleur** est vraiment similaire au cours français
- (13) a. MT: et puis **le FLE** c'est le français langue étrangère  
b. AT: et puis **le bleu** c'est le français langue étrangère

FLE stands for Français Langue Etrangère (French as a Foreign Language). The system here provides two different proposals for the same unknown word.

We have also see that some learners had troubles with the sound /y/ in French, and pronounced it sometimes /u/. As it is a productive difference in French, Whisper sometimes misinterpreted the second-person pronoun :

- (14) a. oui et **tout** marche très bien  
b. oui et **tu** marches très bien
- (15) a. maintenant **tout est tout** passe bien  
b. maintenant **tu es tu** passes bien

The pronunciation of French for a L2 learner differs naturally from a native pronunciation, and might not have been encountered lots by Whisper during the data training phase. Without confidence scoring or any information to understand why the system chose *fleur* in one case and *bleu* in the other, it is difficult to understand which linguistics parameters have contributed. The system is perceived as a black box for the users as one can not know which patterns and rules are applied by the system.

Using those systems as a basis for a linguistic analysis of case studies could be interesting. Unfortunately, it does not provide such information outside of the result.

Linguistic levels	Manual transcripts	Automatic transcripts
Pronunciation	la langue étranger	le long est rejeu
Prosody	co-douche	coudouche
Morphology	entendons	entendant
Syntax	je <b>toujours</b> parle le français	je <b>parle toujours</b> le français
Semantics	expérierencer	expérimenter

Table 3: Examples of errors

## 7 Conclusion

As *Tancoigne et al.* states, if we consider that transcribing is already analyzing then delegating this work to a machine can be seen as problematic in a number of cases (*Tancoigne et al., 2022*). This study aims at specifying which elements have to be taken into consideration for using such technologies on learner corpora.

Our study presents some limitations. An important one is linked to conducting this research with only one ASR system: the discrepancies we showed are inherent on Whisper. It would be needed to compare the results obtained with other ASR systems.

Secondly, we focused our analysis on advanced beginners which are one specific group of learners and would also need to add different level groups to get a bigger perspective on the performances of ASR and of the possible usage of this technology for SLA studies.

Nevertheless, Whisper appears as a good starting point for a manual correction of transcriptions. Its hyper-correction is not suited for the degree of precision needed on the actual production of the speakers. Thus it can provide a first version of the transcription, aligned on the audio with timestamps, and correcting transcriptions rather than creating them from scratch can diminish cognitive overload.

One very interesting feature that we found is that Whisper get very good results on inaudible speech for human ears, and therefore allows to double check manual transcriptions and complete them. This is something that can be useful in order to complete some data for which the sound is inaudible for the transcriber or to choose between different transcription possibilities.

Those reasons conducted us to add a new

import feature on our transcription and annotation tool *Sarramanka* (*El Ayari, 2022*) to be able to take Whisper transcripts as a starting point for manual check before any annotation process. Nevertheless the data would have been reviewed in totality and checked thoroughly to correct the elements transcribed in a native manner.

As we have discussed, transcription is a crucial part of SLA researches on speech and a crucial step that is the basis for any linguistic analysis. Therefore an automatic system could really be a very helpful tool for the study of those corpora, especially as there are many corpora open-source and available which have been documented, transcribed, annotated and analyzed. Those are precious resources that could be used to fine-tune ASR systems.

Therefore it is important to keep in mind that those systems can provide help and facilitate some treatments but that a human check is always needed in order to guarantee the quality of the data processed automatically.

## 8 Perspectives

It is needed to train the system on data from learners of French, but the question arises of the impact of source languages, which induce specificities concerning the acquisition of the pronunciation of the target language - French in this case. Next step is to train the model on learners' corpora with specific dataset matching the SLA issues we present.

We want to conduct a similar study on beginners' productions and on learners with different L1 on the VILLA corpus. The corpus is issued from the project ANR ORA *Varieties of Initial Learners in Language Acquisition: controlled classroom input and elementary forms of linguistic organisation*. The researchers observed the

acquisitional path for L2 acquisition of Polish with only 14 hours of exposure for learners from five different L1: French, Italian, German, British English and Dutch. This corpus will allow us to see the impact of pronunciation and accent on the automatic transcription provided, with a similar level of acquisition. It will also be interesting to see how efficient the system can be on beginners' productions - as they should be more distant from ASR systems' inside norm.

Next step will be to train the model on learner corpora with specific datasets matching the issues specific to SLA we have presented in Table 3. It would be really interesting to fine-tune Whisper or another ASR system like *wav2vec* (Baevski et al., 2020) on learner corpora depending on the L2 or on the L1. As we said before, those corpora can be considered like a poorly endowed language due to their specificities.

## Acknowledgments

The project **TransSLA** has been founded within the support of *Paris 8 University* and the research laboratory *Structures Formelles du Langage*.

## References

- A. Baevski, H.Zhou, A.Mohamed, and M. Auli. 2020. *wav2vec 2.0: A framework for self-supervised learning of speech representations*.
- N. Ballier, A. Meli, M. Amand, and J.-B. Yunès. 2023. *Using whisper LLM for automatic phonetic diagnosis of L2 speech, a case study with French learners of English*. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 282–292. Association for Computational Linguistics.
- S. Benazzo and M. Watorek. 2021. *Transcription de corpus oraux d'apprenants débutants en français l2 : quelques enjeux théoriques*. In *L. Spreafico, G. Bernini, A. Valentini J. Saturno (éds.) Superare l'evanescenza del parlato. Un vademecum per il trattamento digitale di dati linguistici*, pages 127–165. Bergamo: Sestante.
- C. Blanche-Benveniste. 2000. *Approches de la langue parlée en français*. Paris: Ophrys.
- R. Bley-Vroman. 1983. *The comparative fallacy in interlanguage studies: The case of systematicity*. *Language Learning*, (1):1–17.
- A. van den Bosch, A. Content, W. Daelemans, and B. de Gelder. 1994. *Measuring the complexity of writing systems*. *Journal of Quantitative Linguistics*, 1(3):178–188.
- V. Chanethom and A. Henderson. 2022. *Alignment in ASR and L1 listeners' recognition of L2 learner speech: A replication study*. In *15th International Conference on Native and Non-native Accents of English*, Łódź, Poland. Université de Łódź.
- S. P. Corder. 1980. *La sollicitation de données d'interlangue*. *Langages*, (57):29–27.
- R. Cumbal, B. Moell, J. Lopes, and O. Engwall. 2021. *"You don't understand me!": Comparing ASR Results for L1 and L2 Speakers of Swedish*. In *Proceedings Interspeech 2021*, pages 2021–2140.
- L. Dekydtspotter, B. Schwartz, and R. Sprouse. 2006. *The Comparative Fallacy in L2 Processing Research*. 8th Generative Approaches to Second Language Acquisition Conferences.
- S. El Ayari. 2022. *Sarramanka, une plateforme outillée de transcription, d'annotation et d'exploration de corpus*. In *8ème Congrès Mondial de Linguistique Française (CMLF)*, volume 138, page 10006, Orléans, France.
- S. El Ayari and M. Watorek. 2021. *Exploration outillée pour un corpus de productions orales des apprenants débutants en L2*. In *Colloque "Influence translinguistique : où en est-on aujourd'hui ?"*, Toulouse, France.
- R. Ellis and G. Barkhuize. 2005. *Analysing Learner Language*. Oxford:Oxford University Press.
- Calbert Graham and Nathan Roll. 2023. *Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits*. *Cambridge Open Engage*.
- W. Klein. 1989. *L'Acquisition de langue étrangère*. Paris: Armand Colin.
- W. Klein and C. Perdue. 1997. *The Basic Variety (or: Couldn't natural languages be much simpler?)*. *Second Language Research*, 13(4):301–347.
- P. J. Lou and M. Johnson. 2020. *End-to-end speech recognition and disfluency removal*. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2051–2061. Association for Computational Linguistics.
- B. MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.
- C. Perdue. 1993. *Adult Language Acquisition. Vol 1: Field Methods*. Cambridge University Press.

- A. Radford, J. Xu Kim, Brockman T., McLeavey G., C., and I. Sutskever. 2023. Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning*.
- L. Selinker. 1972. Interlanguage. *International Review of Applied Linguistics*, (10):209–231.
- Elise Tancoigne, Jean Philippe Corbellini, Gaëlle Deletraz, Laure Gayraud, Sandrine Ollinger, and Daniel Valero. 2022. Un mot pour un autre ? Analyse et comparaison de huit plateformes de transcription automatique. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 155(1):45–81.
- X. Wei, C. Cucchiaroni, R. van Hout, and H. Strik. 2022. Automatic speech recognition and pronunciation error detection of dutch non-native speech: cumulating speech resources in a pluricentric language. *Speech Communication*, 144:1–9.
- H. Yamasaki, J. Louradour, J. Hunter, and L. Prevot. 2023. Transcribing and aligning conversational speech: A hybrid pipeline applied to french conversations. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Taipei, Taiwan.

# Enhancing a multi-faceted Breton verb-centered resource to help a language learner

**Annie Foret**

Univ. Rennes and IRISA, France

annie.foret@irisa.fr

**Erwan Hupel**

Univ. Rennes 2 and CELTIC-BLM, France

erwan.hupel@univ-rennes2.fr

**Pêr Morvan**

An Drouizig, France

per.morvan.bzh29@gmail.com

## Abstract

This article builds on two recent resources for Breton, a verb-centered database and a set of sentences in the universal dependencies (UD) format. Our focus is on Breton, an endangered language in the Celtic family. We provide an analysis of the resource on verbs and show how it can be connected and transformed to a multi-faceted system intended to help a learner in a flexible way. We discuss several scenarios.

## 1 Introduction and objectives

Working on low-resourced languages comes with specific challenges (Vergez-Couret et al., 2024).

In this paper, we consider this issue for Breton; a discussion can be found for example in (Foret et al., 2015). We provide here a workflow that aims to facilitate the use and access by a learner to rich linguistic data, in a flexible way. The workflow is intended to be open, reproducible, with easily adaptable outputs<sup>1</sup>. The prototype is also devised for a use case within the *formal concept analysis* (FCA) paradigm<sup>2</sup>, handling several facets (kinds of information).

In contrast to a *carrier sentences* or *seed sentence approach* (Heck and Meurers, 2022), our interface starting-point is a set or subset of verb infinitives that a learner wishes to master (by viewing information in several prepared hierarchies) or that he may simply discover by serendipity or *incidental learning* (Renduchintala et al., 2019). The

<sup>1</sup>This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

<sup>2</sup>available at <https://gitlab.inria.fr/foret/termilis/-/tree/main/Breton/Verbs>

<sup>2</sup>in this project: <https://www.smartfca.org/>

outputs are thus intended as self-assisted learning systems involving small resources and lightweight technology (a browser, online or offline) but not requiring specific technical knowledge from users. We use a relational database system to store the linguistic data and generate the end systems proposed to a learner, but this is hidden to the users.

The plan of paper is as follows. In section 2 we point to some Breton specificities and difficulties; in section 3 we discuss two different kinds of resource available for Breton; these resources are used in the new workflow described in section 4; section 5 discusses scenarios enabled by the resources and workflow; section 6 concludes with perspectives.

## 2 Breton linguistic features

In Breton syntax, the verb occurs as second constituent and allows one to put the most important first. Consonant mutations are a particularity of Breton and other languages in the Celtic family. Depending of grammatical features and other features, some initial consonants change to others (Hupel, 2021; Jouitteau, 2009-2024). This is a difficulty for Breton learners and automatic processing as well.

The Breton verb varies depending on a lot of elements. There are two main categories, related to conjugations or not related. Firstly, as it is common, conjugations vary according to person, number, tense, mood, aspect and voice. Secondly, a consonant mutation may apply. For example the initial "k"[k] will become "c'h"[x], or "g"[g]: the infinitive "kanañ" (EN: to sing) occurs as "gan" in this sentence "An eostig a gan bemnoz" (EN: The nightingale sings every evening) where a *soft*

*mutation* from "k" to "g" is induced by the preceding "a" verbal particle; the other verbal particle "e" yields a *mixed mutation*, as in "Bemnoz e kan an eostig". The "pa" (EN: when) conjunction yields a *soft mutation*: the initial "k"[k] will become "g"[g], as "kregiñ" (EN: to begin) in Figure 4. But conjunction "ma" (EN: if) yields a *mixed mutation* where the verb initial "k"[k] is unchanged.

Other difficulties may arise for: a verb without ending ("kemer", "lenn", "komz"); an altered verb base in the infinitive ("skeiñ" → sko-, "mervel" → marv-); a verb base different from the infinitive ("gounit" → gounez-, "dont" → deu-).

Most Breton verbs are regular<sup>3</sup> (Desbordes, 1999). Nevertheless, several grammatical categories are distinguished in the database see section 3 and Figure 2.

### 3 Two existing Breton resources

Breton is a low-resourced language. Nevertheless, we discuss two resources of valuable interest in this verb-centered proposal.

#### 3.1 The DVB verb database

The DVB Breton verb site is handled by the association An Drouizig. Figure 1 shows the top of the page for the Breton verb "kanañ" (EN: to sing)<sup>4</sup>. Each page (in BR, FR, or EN language) provides different kinds of informations on a given verb:

- tags (such as "European level A1 verb"), other forms, translations, sources, links are on the top,
- mutation modes and examples,
- conjugations in details.

We give below a simplified explanation of the DVB verb group classification (for conjugation):

- most verbs are regular and in *d1*;
- verbs in *d2* are regular, end in -aat/-at and express an action taking place;
- verbs in *d3* are regular, end in -a and express picking up something etc.;
- verbs in *d4* are semi-regular, end in -iañ, -iiñ;
- verbs in *d5* are semi-regular, end in -liañ, -liiñ;
- verbs in *d6* are semi-regular, end in -niañ, -niñ;
- verbs in *d7* are irregular, follow the conjugation

<sup>3</sup>see also [https://arbres.iker.cnrs.fr/index.php?title=Verbes\\_irr%C3%A9guliers](https://arbres.iker.cnrs.fr/index.php?title=Verbes_irr%C3%A9guliers)

<sup>4</sup><https://displeger.bzh/en/verb/kana%C3%B1>

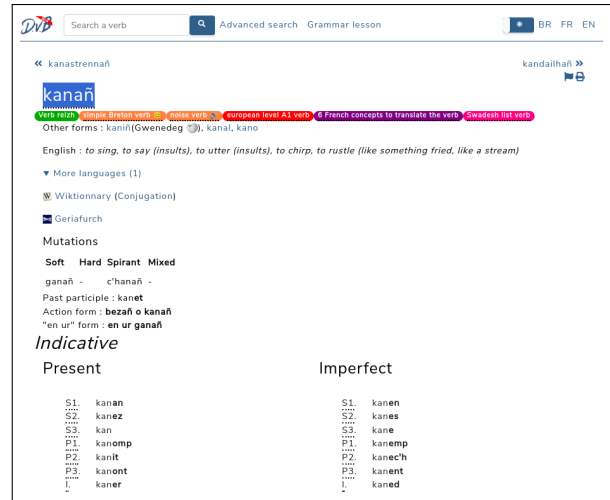


Figure 1: "kanañ" (to sing) at <https://displeger.bzh>

of "ober";

- *d8* regroups special verbs: "bezañ" / "bout" (to be), "kaout" / "endevout" (to have), "dont" (to come), "mont" (to go), "gouzout" (to know).

Figure 2 shows statistics on grammatical categories. We computed them on the DVB relational database for Breton verbs provided by An Drouizig, behind the DVB website.

category	nb_infinitive	nb_verb_id	example
bezan	2	1	bezañ
d1	11294	9308	abafañ
d2	838	823	abafaat
d3	487	471	abona
d4	562	542	adeouliañ
d5	101	99	adiziliañ
d6	100	98	adkoaniañ
d7	13	12	addizober
dont	6	3	addonet
gouzout	2	1	goût
kaout	2	1	endevout
mont	3	2	enmont

Figure 2: Verb grammatical categories in DVB.

More globally, our analysis of the relational database yields the general data model in Figure 3 with content statistics. Each table comes with a name, a list of attributes (the table columns) with its number of lines at the top. The underlined attribute is its *primary key*. Each edge stands for a *foreign key* connecting an attribute in the source table to another attribute that it refers to in the target table.

This database is a key component for the workflow proposed in section 4. We now describe the second component used in the workflow.



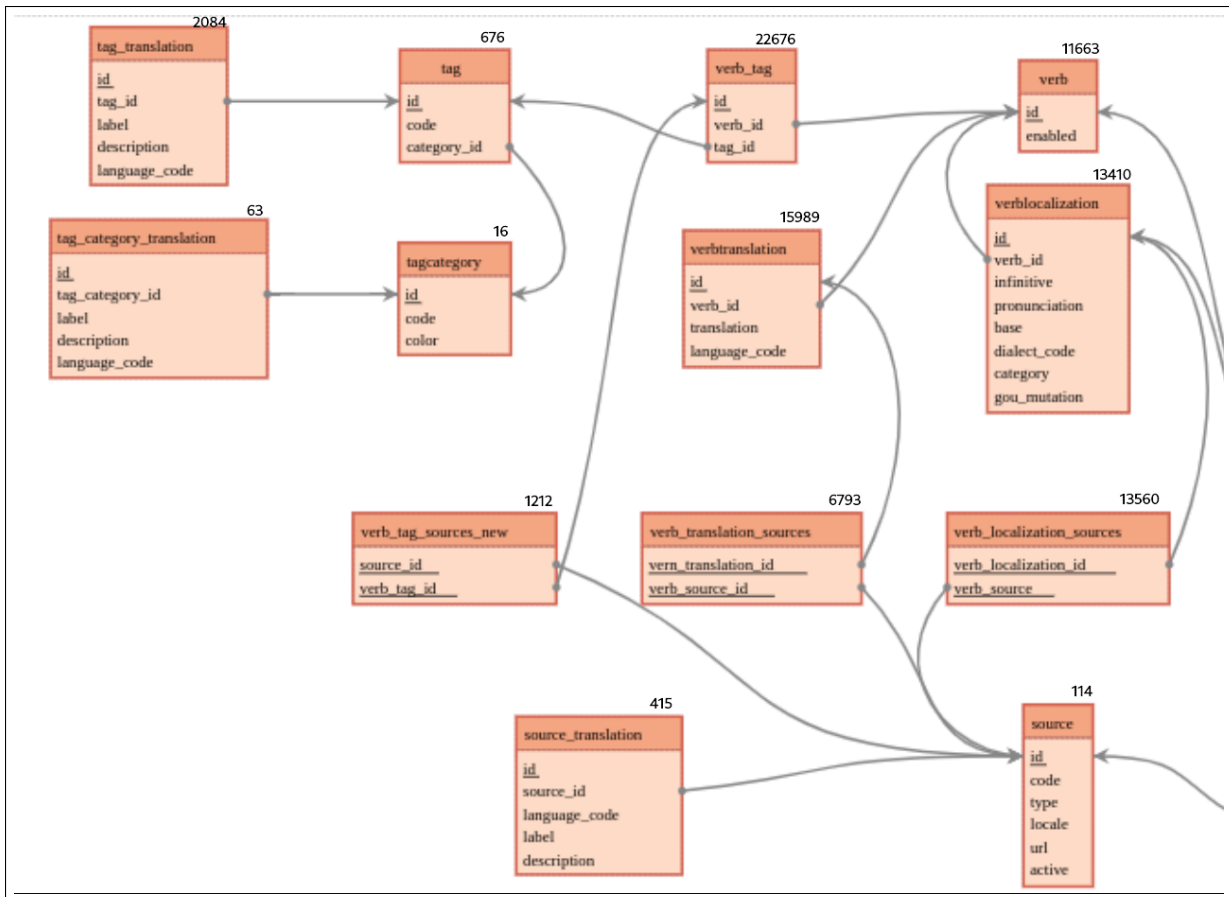


Figure 3: A relational schema for DVB (main part): each table has a name, a line count above, attributes (the key is underlined). The edges show foreign keys. We focus particularly on tag, verbtranslation and verblocalization tables and their code, translation, infinitive, category attributes.

The screenshot shows the Universal Dependencies web interface. The search query is: `pattern { X [lemma="kregiñ" ] }`. The results show 5 occurrences of the pattern. The parse tree for the sentence "Pa grog da virviñ adarre, tennit diwar an tan." is displayed. The word "grog" is highlighted in green, and its metadata is shown: `upos=VERB, lemma=kregiñ, Mood=Find, Number=Sing, Person=3, Tense=Pres, VerbForm=Fin`.

Figure 4: Grew can show (and rewrite) the treebank part that matches a linguistic pattern (on the top). The infinitive "kregiñ" (EN: to begin) is highlighted in this parse tree (its exact form is "grog", with soft mutation after "Pa"). The text[eng] metadata is: "When boiling again, draw from the fire". The second infinitive is "birviñ" (EN: to boil).

## 3.2 UD treebanks

Dependency syntax has been developed for a long time, for example in (Mel'čuk, 1988). This approach underlies the active area of universal dependencies (Nivre et al., 2016; de Marneffe et al., 2021), an annotation framework with treebanks in over 150 languages.

Several dependency treebanks are developed for Celtic languages (Lynn and Foster, 2016; Batchelor, 2019; Heinecke and Tyers, 2019). In this work we consider the UD Breton-KEB corpus V1.0<sup>5</sup> (Tyers and Ravishankar, 2018) with a 2023 revised version<sup>6</sup>. Its first annotated sentence is:

```
# sent_id = apertium.vislqg.txt:1:0
# text = N'int ket aet war-raok.
# text[eng] = They didn't progress.
# text[fra] = Ils n'ont pas progressé.
# labels = to_check
1 N' ne ADV adv Polarity=Neg 4 advmod _ SpaceAfter=No
2 int bezañ AUX vblex
Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin 4 aux _ _
3 ket ket ADV adv _ 4 advmod _ _
4 aet mont VERB vblex Tense=Past|VerbForm=Part 0
root _ _
5 war-raok war-raok ADV adv _ 4 advmod _ SpaceAfter=No
6 . PUNCT sent _ 4 punct _ _
```

We see the meta-information above at the beginning (lines starting with #), then a line by word occurrence in sentence order, with tabs separated columns: ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC. From this, the dependency parse tree can be drawn (as in Figure 4). Our workflow exploits meta-information and FORM, LEMMA, UPOS columns (crucially, UPOS='VERB' tells which word occurrences are verbs).

The universal dependencies site also collects a list of tools for working with UD. We mention two:

- Grew<sup>7</sup> (Guillaume, 2021) is a graph rewriting tool dedicated to applications in Natural Language Processing. Figure 4 shows the Breton treebank with a query.

- CoNLL-U viewer at rug.nl is a simple browser-based UD viewer. Figure 9 displays a rewritten version of the treebank and its browsing.

We will show a scenario linking verb data to related sentences that have been parsed in the universal dependency format. The workflow enabling this scenario is described in next section.

<sup>5</sup>available at [https://universaldependencies.org/treebanks/br\\_keb/index.html](https://universaldependencies.org/treebanks/br_keb/index.html)

<sup>6</sup>at <https://github.com/UniversalDependencies/UD-Breton-KEB>

<sup>7</sup><https://grew.fr>

## 4 A new workflow

The database is analyzed and processed to control, to enhance and to select appropriate fragments. The workflow outputs several versions (.csv, .html, .ttl/rdf exports) allowing different scenarios.

**DVB processing** We define different views on the DVB tables. For an HTML output, a typical generated line (in the tag part) is:

```
<a href="https://displeger.bzh/fr/verb/selaou"
data-id="5711" title="écouter" data-init="s"
data-categ="d1" data-base="selaou" data-idv="38435"
data-tag="verb_al_live_A1" data-row="179">selaou</a>
<a class="bis" href="#selaou-1" > # </a>
```

Some informations are rendered as HTML attribute-value pairs, for a basic HTML view with many CSS stylesheet possibilities, this is also shaped to show useful information on hover and useful links (to a relevant DVB page or to relevant UD sentences). The grammatical category stored in verblockalization is given as an attribute-value pair, such as data-categ="d1" for the verb "selaou", not visible in the browser, but could appear as HTML content, by a simple CSS rule.

**Verb facet selection** The DVB database contains grammatical categories in the verlocalization table. DVB also contains many tags of various kinds in the tag table, that we organize in 11 subclasses (such as level, or domain)<sup>8</sup> to view fragments in a flexible and informative way.

We produce in this way an enhanced version (see Figure 5) in turtle/RDF (semantic web) format, that enables search using the SPARQL Query Language for RDF or related tools such as Sparklis (Ferré, 2016); this approach is applied to Georgian verbs by (Ducassé and Elizbarashvili, 2022).

For ease of use, we also produce simpler HTML versions: either for verbs alone as in Figure 6, or for verbs connected with sentences from a UD treebank (parse trees) as explained below.

**Trebank preparation** Before an upload in SQL, the UD corpus file is prepared, adding a column with a line number, and handling special symbols (quote

<sup>8</sup>our current tag subclass list is: Level, Domain, Link, Substring, NbSyllabs, Ends (for verb ending), Change (for variations), Construct, Args (for transitive, etc.), Synset (number of synonyms for translation), Other.

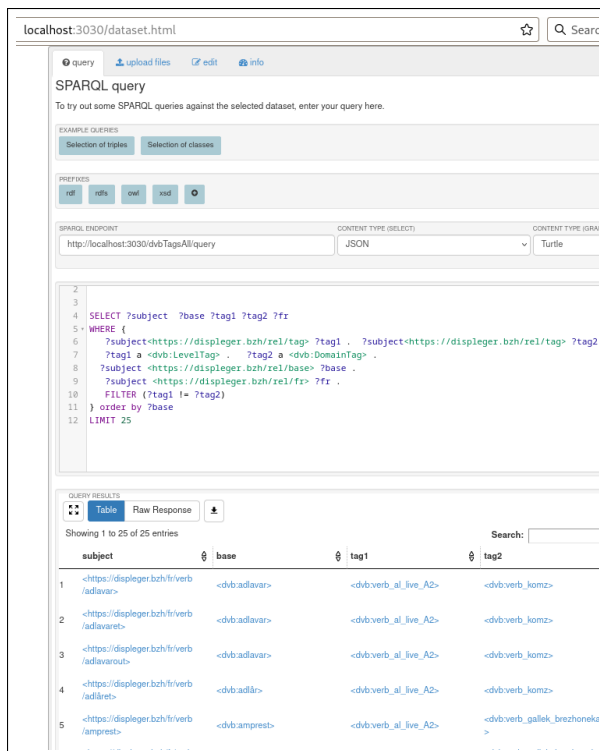


Figure 5: DVB verb facets in RDF, with tags hierarchy.

in quote, etc.). We then partition the numbered lines in two relational tables, `br_keb_ud` (line, WID, FORM, LEMMA, UPOS, XPOS, FEATS, HEAD, DEPREL, ...) for word information, `br_keb_sent`(line, sent) for whole sentence information. We then build SQL views that generate HTML lines such as:

```
<p class="sent"><a id="selaou-2"
href="https://displeger.bzh/fr/verb/selaou">selaou</a>
<span data-form="selaouit" data-root="[r]">selaouit
</span> <span class="line" data-sentnum="511">
(line 8665, sent 511) </span># text = Va
<span class="solution" title="selaouit">_</span>
<span class="w" title="selaouit">selaouit</span>!
<span class="sentFr"> # text[fra] = Écoutez-moi!
</span><p>
```

We use in particular this CSS rule:

```
span.w {visibility: hidden;}
```

to hide the conjugated form of the verb (this CSS rule may be dropped to show full sentences).

A similar output is generated for English translations, selecting lines of `br_keb_sent` containing "text[eng]" instead of "text[fra]".

## 5 Enabled use case scenarios

We first describe successive scenarios in the HTML mode, based on a navigator. We suppose

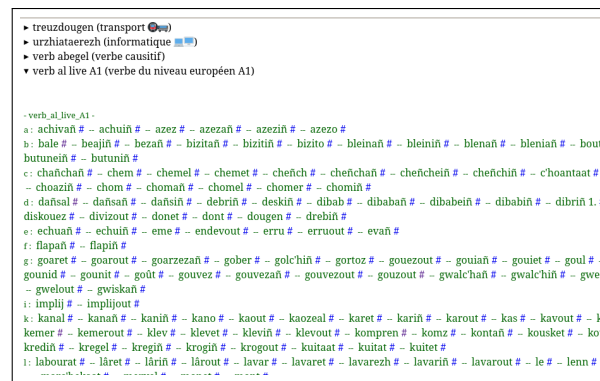


Figure 6: DVB verbs in HTML, <details> tag "european level A1" open. The top tag contents are hidden. After a click on the "verb al live A1" tag, the related verb list is visible. A translation is then directly visible by hovering over a verb.

the user has a local copy of the HTML and CSS files. In this HTML mode, the file produced by the workflow gathers several kinds of information in one place. And the user can view and interact without internet connection.

### Example scenario (1): verbs only, HTML mode

The first file version regroups information from the DVB database, showing verbs by categories (such as "level A1", then by initial letter inside a category). As explained in section 4, each verb is accompanied with attribute-value pairs that can be shown or not, depending on the chosen CSS rules (on the developer side). Figure 6 corresponds to this version. This rendering uses the HTML5 <details> element, so that the user can open and close an item (such as the "verb al live A1" tag) to display its content. For each infinitive such as `achivañ`: its translation appears on hover and a click on the infinitive links the DVB site for this verb.

An enhanced version is provided that regroups the various tags in 11 classes (supertags) as explained in section 4 and visible in the middle of Figure 7. Note that this is strict hierarchy on tags, while verbs appear in each tag they belong to (they may appear several times in the file).

In these ways, users may choose a tag or a facet (possibly several) and test their knowledge on verbs belonging to this tag or this facet, in a compact way.

### Example scenario (2): including sentences (HTML mode)

This second file has two sections, a tag section corresponding to the whole first

```

p: pakañ # -- paouez # -- pediñ # -- pellgomz # -- pleustrañ # -- plijout # -- prenañ #
r: redek # -- reiñ # -- reseo # -- resev # -- reseviñ # -- resevout # -- reskont # -- respont #
s: santout # -- sarañ # -- selaou # -- sellet # -- selliñ # -- sellout # -- serr # -- serrañ # --
skriñ # -- skriviñ # -- soñjal # --
t: tapanñ # -- tapout # -- tresañ #
v: viziñ #
  > verb al live A2 (verbe du niveau européen A2)
  > verb al live B1 (verbe du niveau européen B1)

```

► Link tag (liens)  
► Substring tags (sous-chainés difficiles)  
► Synset tags (nombre de synonymes fr)  
► Other tags (autres descriptifs)

**Phrases du corpus UD en breton br\_keb\_ud**

▼ Liste de phrases selon leur verbe  
L'infinitif à gauche est lié au site DVB, il est suivi de la phrase du corpus où ce verbe est masqué, sa fr en passant sur le souligné.

[adaozañ](#) (line 13881, sent 803) # text = Prezidant an daou guzul-rannvro ha prezidant kuzul-depar c'hall divizout reiñ lañs da \_ ar rannvro. # text[fra] = Seuls les présidents des deux conseils général de Loire-Atlantique peuvent décider de lancer la réorganisation de la région.

[adkavout](#) (line 13057, sent 762) # text = Muioch-mui e \_ ivez ar yezh vrezhon e-liamm gant a langue bretonne est aussi de plus en plus présente dans le monde économique.

Figure 7: DVB verb facets in HTML, <details> tag "european level A1" open, with sentences. In the sentence section, the infinitive on the left is linked from the A1 tag list and links to the DVB site.

```

selaou (line 7244, sent 402) # text = _ a rit an avel. # text[eng] = You listen to the wind.
selaou (line 8665, sent 511) # text = Va _ ! # text[eng] = Listen to me!
selaou (line 8674, sent 512) # text = Na _ et ac'hanon! # text[eng] = Don't listen to me!
sellout (line 2509, sent 145) # text = Ur stourm a ranko kenderc'hel (ha marteze dont da vezañ... nerz doujet evit pezh a _ ouzh skignañ o yezh er radio. # text[eng] = A fight will have to continue (and Bretons want to be respected with regard to the radio broadcasting of their language.

```

Figure 8: Breton sentences from the corpus with their EN translation and the verb form hidden (shown on hover), linked from the selected infinitive "selaou" in the DVB tag "european level A1" open.

file, and a sentence section as in Figure 7. Sentences in the sentence section are ordered by infinitive (rewritten on the left of the sentence). A click on # in the tag section points to the first sentence (in the sentence section) where the verb occurs; the verb occurrence is hidden (by the chosen CSS stylesheet) in the sentence. The exact verb form appears on hover. Sentences with a same infinitive follow each other, which enables training on the same verb with proximate sentences.

Note that a sentence appears for each infinitive that occurs in it (a sentence may appear several times in the file, depending on its number of verbs).

In these ways, users may choose a tag or a facet (possibly several) and practice or check several aspects on a verb:

- its hidden meaning as in the first file;
- its hidden conjugation in a set of sentences (the solution appears on hover over the key \_, as "selaouit" in Figure 8). For explanations, they may also consult the appropriate page of the DVB site, by a click on the infinitive (in either section).

Figure 9: Browsing the rewritten UD treebank (the infinitive in the tree node replaces the exact verb form).

**Example scenario (3): including parse trees (HTML mode)** Grew can show and rewrite a treebank part that matches a linguistic pattern.

On the preparation side, to transform the treebank we applied the Grew command<sup>9</sup> to rewrite tree features and then converted the output with the sed command to hide the "text =" metadata. The following one-rule rewriting system hides all verb-forms, replaced by their infinitive :

```

package hide-verb-form {
  rule hideVerbsForm {
    pattern { X1 [upos=VERB] }
    without { X1.mark = "x" }
    commands { X1.mark = "x" ;
               X1.form = "?(" + X1.lemma + ")"; }
  }
}
strat main { Onf(hide-verb-form) }

```

On the user side, the resulting treebank can then be loaded and searched as in Figure 9 for a small sentence (see Figure 4 for a larger tree).

**Semantic Web mode** In this mode, we assume the user has a copy of the .ttl file and has installed a SPARQL server such as Apache Jena Fuseki. The user or developer familiar with RDF web semantic standards can load the .ttl file to query it and to explore the data in these two ways: directly write a SPARQL query (as exemplified in Figure 5 or build a query with Sparklis that is a tool with guidance in a natural language as in (Ferré, 2016; Ducassé and Elizbarashvili,

<sup>9</sup>a rewriting system may also be loaded on web.grew.fr and applied on the selected corpus tree

2022). In this mode also, the file produced by the workflow gathers several kinds of information in one place. The facet filtering is very flexible, even more with `Sparklis` requiring less knowledge on the data model.

## 6 Conclusion and future work

The workflow described in this paper outputs easy-to-use language learning verb-centered contexts, aimed to help a Breton learner. The outputs gather heterogeneous information on one or few files, so that a user may train with different scenarios and facets of verbs (including flat or structured sentences). This is still work in progress, more automation and scenario variations could be provided and tested. A user study could also be added. We list some other points for future work.

- At the level of word descriptors, the hierarchy of tags could be exploited in a more elaborated way, in particular within the *formal concept analysis* (FCA) paradigm; we could test the potential of such approaches on the design and use of a self-assisted learning system. The FCA approach could also show sets of verbs sharing a same set of descriptors. Sentences from a parsed corpus (where verbs are tagged with their linguistic features as in UD) could inherit their verb descriptors as well, providing indicators per sentence in a flexible way.
- At the level of treebanks, a new<sup>10</sup> Breton UD treebank is in preparation, which may provide new insights. Other sentence structures have been proposed depending on the preferred grammatical formalism and parsing principles; the SUD (Surface-syntactic Universal Dependencies) variant (Gerdes et al., 2018) is available for UD treebanks from Grew and could have been proposed here instead of UD. Semantic structures such as AMR (Abstract Meaning Representations)<sup>11</sup> (Heinecke and Shmorina, 2022) might also bring help, but we are not aware of such data for Breton.

<sup>10</sup>see <https://arbres.iker.cnrs.fr/index.php?title=Breton.treebank.II>

<sup>11</sup>AMR page: <https://amr.isi.edu/>, AMR bibliography <https://nert-nlp.github.io/AMR-Bibliography/>

- As concerns workflow handling, the development follows a reproducibility principle and we believe the workflow should apply to the new treebank and to augmented versions of the verb database (with few adjustments).

We generated browser-based versions aimed at individualized learning solutions. Worksheets or gap filling exercises could be generated in a close way by the workflow.

We think there is a need to enhance existing resources especially on a low-resourced and endangered language such as Breton. We hope this development is a step in this direction.

## Acknowledgments.

We thank Gwenn Meynier from An Drouzig, for information on the DVB resource for Breton.

## Annex: Breton mutation system

See Figure 10 for an overview on the four mutation kinds: soft, spirant, hard and mixed.

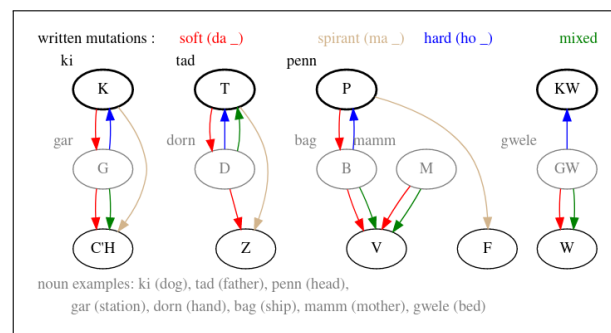


Figure 10: Breton initial mutation overview (on nouns)

## References

- Colin Batchelor. 2019. Universal dependencies for Scottish Gaelic: syntax. In *Proceedings of the Celtic Language Technology Workshop*, pages 7–15, Dublin, Ireland.
- Yann Desbordes. 1999. *Petite grammaire du breton moderne*. Mouladurioù Hor Yezh.
- Mireille Ducassé and Archil Elizbarashvili. 2022. Finding lemmas in agglutinative and inflectional language dictionaries with logical information systems: The case of Georgian verbs. In *Proceedings of XX EURALEX International Congress*. Mannheim: IDS-Verlag. Demonstration.
- Sébastien Ferré. 2016. `Sparklis`: An expressive query builder for sparql endpoints with guidance in natural language. *Semantic Web*, 8:405–418.

- Annie Foret, Valérie Bellynck, and Christian Boitet. 2015. *Akenou-breizh, un projet de plate-forme valorisant des ressources et outils informatiques et linguistiques pour le breton*. In *Actes de la Traitement Automatique des Langues Régionales de France et d'Europe*, Caen, France. Association pour le Traitement Automatique des Langues.
- Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. *SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD*. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.
- Bruno Guillaume. 2021. *Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion*. In *EACL 2021 - 16th conference of the European Chapter of the Association for Computational Linguistics*, Kiev/Online.
- Tanja Heck and Detmar Meurers. 2022. *Generating and authoring high-variability exercises from authentic texts*. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 61–71, Louvain-la-Neuve, Belgium. LiU Electronic Press.
- Johannes Heinecke and Anastasia Shimorina. 2022. *Multilingual Abstract Meaning Representation for Celtic languages*. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 1–6, Marseille, France. European Language Resources Association.
- Johannes Heinecke and Francis M. Tyers. 2019. *Development of a Universal Dependencies treebank for Welsh*. In *Proceedings of the Celtic Language Technology Workshop*, pages 21–31, Dublin, Ireland. European Association for Machine Translation.
- Erwan Hupel. 2021. *Le Breton [quelques contrastes pertinents entre le français et le breton]*. In *projet Langues et grammaires du monde dans l'espace francophone*. [lgidf.cnrs.fr](http://lgidf.cnrs.fr).
- Mélanie Joutteau. 2009-2024. *ARBRES, wikigrammaire des dialectes du breton et centre de ressources pour son étude linguistique formelle, IKER, CNRS*. <http://arbres.iker.cnrs.fr>. Licence Creative Commons BY-NC-SA.
- Teresa Lynn and Jennifer Foster. 2016. *Universal dependencies for irish*. In *Proceedings of the Celtic Language Technology Workshop*, Paris, France.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. *Universal Dependencies*. *Computational Linguistics*, 47(2):255–308.
- I. Mel'čuk. 1988. *Dependency Syntax*. SUNY Press, Albany, NY.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. *Universal Dependencies v1: A multilingual treebank collection*. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).
- Adithya Renduchintala, Philipp Koehn, and Jason Eisner. 2019. *Simple construction of mixed-language texts for vocabulary learning*. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 369–379, Florence, Italy. Association for Computational Linguistics.
- Francis M. Tyers and Vinit Ravishankar. 2018. *A prototype dependency treebank for breton*. In *Actes de la Conférence TALN. CORIA-TALN-RJC 2018 - Volume 1 - Articles longs, articles courts de TALN, Rennes, France, May 14-18, 2018*, pages 197–204. ATALA.
- Marianne Vergez-Couret, Delphine Bernhard, Michael Nauge, Myriam Bras, Pablo Ruiz Fabo, and Carole Werner. 2024. *Managing fine-grained metadata for text bases in extremely low resource languages: The cases of two regional languages of France*. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 212–221, Torino, Italia. ELRA and ICCL.

# Evaluating Automatic Pronunciation Scoring with Crowd-sourced Speech Corpus Annotations

Nils Hjortnaes, Daniel Dakota, Sandra Kübler, Francis Tyers

Indiana University

{nhjortn, ddakota, skuebler, ftyers}@iu.edu

## Abstract

Pronunciation is an important, and difficult aspect of learning a language. Providing feedback to learners automatically can help train pronunciation, but training a model to do so requires corpora annotated for mispronunciation. Such corpora are rare. We investigate the potential of using the crowdsourced annotations included in Common Voice to indicate mispronunciation. We evaluate the quality of ASR generated goodness of pronunciation scores through the Common Voice corpus against a simple baseline. These scores allow us to see how the Common Voice annotations behave in a real use scenario. We also take a qualitative approach to analyzing the corpus and show that the crowdsourced annotations are a poor substitute for mispronunciation annotations as they typically reflect issues in audio quality or misreadings instead of mispronunciation.

## 1 Introduction

Pronunciation of utterances is a difficult task for language learners, and there is limited research on how best to generate feedback automatically (Agarwal and Chakraborty, 2019; Moses et al., 2020; Neri et al., 2006; Witt, 2012). However, such feedback can be an invaluable tool for those learning a language who want to improve their speaking skills, allowing them to practice when a human teacher is not available. Ideally, the feedback should reflect the judgements of a native speaker of the targeted language variant and be targeted at the learner’s desired dialect (e.g., British vs. American English) and skill level. One current method for evaluating pronunciation is to interpret the confidence of an Automatic Speech Recognition (ASR) model as the goodness of pronunciation (Moses et al., 2020). Doing so makes a crucial

assumption that the accuracy of the transcription is representative of the learner’s pronunciation accuracy.

One of the challenges in investigating the quality of automatic feedback is that there is only one publicly available corpus with human judgements on pronunciation, L2-ARCTIC (Zhao et al., 2018). Since it does not contain examples of native speakers producing the same sentences, we cannot use it for our purposes.

The Common Voice corpus (Ardila et al., 2020) does not contain pronunciation annotation, but does contain upvote and downvote scores per utterance. We propose using these crowdsourced up- and downvote scores as a stand-in for pronunciation scores. We hypothesize that a clip receiving both up- and downvotes indicates a mispronunciation because annotators disagree on the quality, and clips with only upvotes indicate proper pronunciation as well as clear audio. To test whether these labels can be used for evaluating pronunciation scorers, we create a task to classify whether a given audio clip in the Common Voice corpus has any downvotes using the generated pronunciation scores as input. Assuming that the output of a Speech Recognition model is a measure of pronunciation accuracy (Moses et al., 2020), a neural model should be able to use that output to predict the presence of downvotes.

Typically, in ASR, the task is transcribing audio data into orthographic text. In this work we perform a zero-shot classification of downvoted clips using an ASR model (section 5.1). The final layer of this architecture is a softmax layer, providing probabilities, which form the basis of our baseline pronunciation scorer and which we compare across speakers to generate feedback (section 4).

Our results show that detecting downvotes in Common Voice is difficult. The baseline, interpreting the speech recognition softmax output as feedback, achieves only 81.4% with tuning, and in

---

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

the low 60s when comparing learner’s utterances to expert’s and predicting downvotes from the comparison. Looking closely at some of the examples and contents affirms that the voting on Common Voice utterances is a poor substitute for mispronunciation annotation. This highlights the need for a dedicated corpus annotated specifically for pronunciation for the development of tools providing pronunciation feedback to language learners.

## 2 Related Work

Pronunciation feedback systems were researched in depth in the 1990s and 2000s (Witt, 2012), as they have been shown to improve learner’s pronunciation (e.g., Agarwal and Chakraborty, 2019; Neri et al., 2006; Dalby and Kewley-Port, 1999). Early pronunciation feedback used Hidden Markov Models (HHMs; Franco et al., 2000; Dalby and Kewley-Port, 1999), following the use of HHMs for Speech Recognition at the time (Malik et al., 2021). Bratt et al. (1998) collected a corpus annotated for pronunciation during this time for evaluating these systems, but it is no longer available.

As speech recognition moved to neural network models (Malik et al., 2021; Hannun et al., 2014), pronunciation feedback followed (Agarwal and Chakraborty, 2019; Moses et al., 2020). Moses et al. (2020) use DeepSpeech (Hannun et al., 2014) to score pronunciation of Te reo Māori, an indigenous language in New Zealand, using their own speech and text corpora by calculating confidence scores for characters, as opposed to utterances, in an elicited sentence or phrase. There is no information available on how the scoring is performed. It appears to consist of the probability of the character from the target sentence appearing at its aligned timestamp, which is interpreted as the model’s confidence for that character. They “observed the model working with confident te reo speakers as expected”. (Moses et al., 2020)<sup>1</sup>

There are currently many proprietary apps for language learning which include pronunciation training in some form (Coulange, 2023). Common practice for these apps is to give the learner an elicitation phrase and an example of an expert pronouncing it, then request the learner say the phrase. Most apps, such as Memrise<sup>2</sup> and

<sup>1</sup>Only a poster is available for this work [https://pareo.nz/docs/PapaReo\\_NeurIPS2020\\_Poster.pdf](https://pareo.nz/docs/PapaReo_NeurIPS2020_Poster.pdf)

<sup>2</sup><https://www.memrise.com>

DuoLingo<sup>3</sup>, give only binary feedback (correct or incorrect), on a phrase or word level. ELSA<sup>4</sup> is able to give feedback on specific letters, based on phonemes, but only teaches English. Our long term goal is to generate feedback as narrowly as ELSA with a system that can generalize to multiple languages.

## 3 The Common Voice Dataset

We use the Common Voice English data. Common Voice is a large multilingual collection of audio data for speech recognition crowdsourced by Mozilla (Ardila et al., 2020). It consists of around 1.6 million clips ( $\leq 10$  sec.) of read sentences/phrases totalling 2 319 hours. Users can contribute recordings of sentence readings, or judgements of other’s readings by upvoting or downvoting clips<sup>5</sup>. Only clips with at least one upvote are ultimately included in the validated dataset.

Though the upvotes and downvotes do not necessarily indicate a mispronunciation, they do indicate problems as judged by human contributors. Because mispronunciation is a potential reason for an annotator to downvote a clip, these judgements give us the best indication for which clips are mispronounced.

## 4 System Overview

### 4.1 System Pipeline

The pipeline for the process of generating feedback for a given elicited phrase begins with running both the expert and the learner productions of the phrase through the speech recognizer, Coqui (see Section 5.1) and retrieving a softmax probability distribution per time slice. Coqui operates by segmenting an audio file and predicting the character, or lack of a character, present in each segment. This takes the form of a probability distribution over the candidate alphabet. It then recombines the segments into orthography, combining repeating characters<sup>6</sup> and inserting spaces as informed by a language model. Figure 1 shows this process, starting with the extraction of probability distributions in the first transition from the

<sup>3</sup><https://www.duolingo.com>

<sup>4</sup><https://elsaspeak.com>

<sup>5</sup>There is no meta data available about the individual language skills of those upvoting and downvoting.

<sup>6</sup>Double letters, such as the T’s in letter, are handled by a special character prediction.



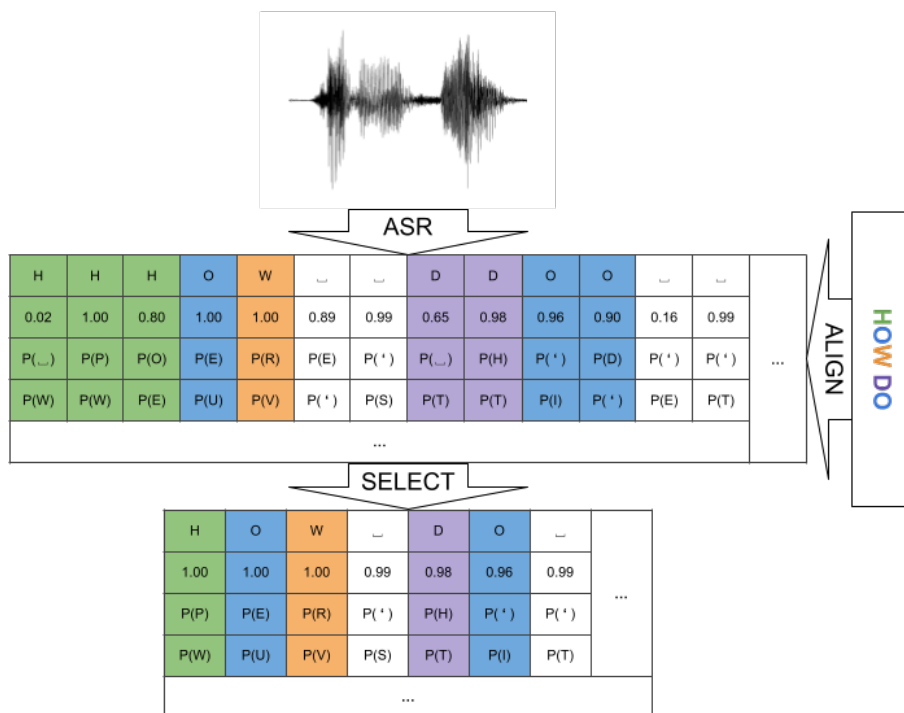


Figure 1: The extraction process for retrieving one probability distribution per character from the audio clip. 1) Extract probability distributions for time slices using via ASR. 2) Align these to the elicitation phrase. 3) Then select one representative distribution per character. The first row of each table represents the highest probability character, the second row that character’s probability, and the 3rd and 4th rows are the next highest probability characters. Each column contains a probability for each character, remaining character probabilities are represented by ellipses.

audio, represented by an arbitrary waveform, to the middle table. Each column in this table represents one time slice where the first row is the highest probability character, the second row is that character’s probability (rounded to 2 decimal points), and the remaining rows indicate probabilities for other likely characters for this time slice. The model also predicts word boundaries, represented by a space (white columns). The next step aligns the probability distributions to the elicitation phrase, using a modification of the Needleman-Wunsch algorithm (see Section 5.2). The alignment is shown via the colors, e.g., all green columns align with the first character in the elicitation phrase. Based on this alignment, the best distribution (i.e., column) per character is chosen to represent the corresponding character in the elicitation phrase. The chosen distributions for each character are shown in the lower table in Figure 1.

Once we have an alignment between the probability distributions and true character labels, we

need to choose one distribution per character in the elicitation phrase (i.e., one column per color, as shown in the lower table in Figure 1) to compare between speakers. This guarantees every character in the elicitation phrase is aligned to at least one probability distribution, even if the most probable character is not the true character. We decide which distribution, from all aligned candidates, to use for each character by choosing the single distribution where the probability of the true character is highest. These final distributions, one per true character, are what we compare between speakers to generate a score for each character.

The process to this point is executed on the learner and expert’s pronunciations of the same phrase, resulting in two probability distributions per character of the phrase which we can compare pairwise. Since similarity comparisons are dependent on the similarity metric, we use three different algorithms for this comparison: cosine similarity, Jensen-Shannon Divergence (Lin, 1991), and Cross Entropy (see Section 5.3).

Elicitation phrase	H	O	...
Best hypothesis: expert	H	O	...
Best hypothesis: learner	H	O	...
	%	0.992	0.975 ...
Comparison	Hel	0.034	0.097 ...
	JSD	0.001	0.011 ...
	XEn	0.016	0.047 ...

Table 1: Example comparing the expert and learner and probability distributions (for the first two characters shown in Figure 1), resulting in a single score per character and similarity metric.

The pairwise comparison of the two speakers’ productions per character is shown in Table 1. The probability distributions for each character per speaker is scored using the comparison algorithms, creating a single score per algorithm, which serves as feedback for each character. Since we do not know which similarity metric is the most suitable one, we experiment with three different ones (see section 5.3 for details).

## 4.2 Quantitatively Evaluating the Corpus

As discussed above, our goal is to evaluate the potential of Common Voice’s annotation as a stand in for pronunciation annotation. I.e., we use the downvotes as indication for incorrect pronunciation. We use the vote annotations as our silver standard; the task then is to predict whether a given clip has any downvotes (irrespective of the number of upvotes) using ASR generated pronunciation scores. Assuming the pronunciation scoring algorithms work well, a classifier should be able to identify clips with downvotes. Since the number of votes per clip is small, we use a binary classification problem rather than predicting the number of downvotes. Most clips have a maximum of 3 total votes, and have 1 downvote and 2 upvotes if there are any downvotes. All clips have at least one upvote.

## 4.3 Data Preprocessing

We choose to focus on sets of files which contain at least 10 different speakers producing the same sentence. We then randomly sample 1 000 of these sets, containing 34 105 total utterances. Of these, the Coqui model fails to process 9,061 clips because of problems identified in preprocessing (e.g. the transcript contains unknown characters, or the clip is longer than 10 seconds). Our final count

Dataset	WER	CER
Sampled Common Voice	0.252	0.153
LibriSpeech clean	0.052	0.019
LibriSpeech other	0.150	0.073

Table 2: Word Error Rate (WER) and Character Error Rate (CER) of sampled data used in our evaluation and Coqui AI’s reported scores for English (Coqui, 2021).

for clips is 25 044. Table 2 shows the Word Error Rate (WER) and Character Error Rate (CER) of the sampled data, along with the scores reported by Coqui for the used model when testing on the full dataset (in the version of 2021) (Coqui, 2021).

By comparing the Coqui STT output of each clip with all other clips of the same sentence (see Section 5.3), we generate 511 532 comparisons. Since we define an expert utterance as one without downvotes, we only accept comparison pairs where one clip only has upvotes (expert) and the other as the language learner. To reduce the data to a manageable size given our compute resources, we reduce these randomly to 20 000 comparisons, split into 15 000 for training and 5 000 for testing.

## 5 System Components

### 5.1 Speech Recognition

We use the freely available model, Coqui STT<sup>7</sup> (Coqui, 2021), based on Baidu’s DeepSpeech (Hannun et al., 2014). Out of the box, Coqui STT predicts an orthographic transcription of speech in an audio file by slicing it into chunks of a specified length (default: 20ms), and using an LSTM network to produce a softmaxed probability distribution over candidate characters per slice. This is illustrated in Figure 1 where the waveform is sliced into 20ms chunks, represented by the columns in the middle table. The rows represent probabilities of candidate characters.

Coqui STT was trained on approximately 47 000 hours of audio data from Common Voice (Ardila et al., 2020), LibriSpeech (Panayotov et al., 2015), and Multilingual LibriSpeech (Pratap et al., 2020). Both Librispeech corpora are comprised of segmented audiobook data.

Coqui STT’s predictions over the sliced audio results in far more characters than the transcription; it decodes this long form transcription into the final predicted words using a Connectionist Temporal Classification (CTC) decoder (Graves

<sup>7</sup><https://coqui.ai> (no longer maintained).

et al., 2006). We modify Coqui STT to preserve and return the softmax output in the form of probability distributions per 20ms time slice of the LSTM in the model’s results, where the probability space is the set of all potential orthographic characters, thus bypassing the CTC decoder.

## 5.2 Needleman-Wunsch Alignment

Since we need to align the transcripts of the time slices to the correct transcription, rather than decoding the speech signal, we modify the alignment algorithm by Needleman and Wunsch (1970).

The algorithm’s original purpose is to align two DNA sequences by calculating the distance between all possible alignments, using Levenshtein distance, and adding insertions to one or both sequences as needed. It then uses a backtrace to find the sequence resulting in the lowest divergence.

The original algorithm results in a 1 : 1 alignment, with some characters aligned to an insertion character. When there are multiple possible alignments of equal weight, Needleman-Wunsch only returns the best entirely aligned sequences. However, for our problem, we need a many to one alignment, allowing us to be intentional about selecting a distribution per elicitation phrase character, rather than relying on the 1 : 1 mappings. We modify the algorithm to allow pairing multiple items from the longer sequence (audio slices) with an item from the shorter sequence (correct transcription).

## 5.3 Comparing Distributions

We use three algorithms designed to compare probability distributions. The first is Hellinger Distance (Hellinger, 1909). It is a simple summation of comparisons between elements in the probability space normalized to be bounded by 0 and 1. The second is Jensen-Shannon divergence (JS; Lin, 1991). JS divergence is based on KL divergence (Kullback and Leibler, 1951), but it is symmetrical, making it a more consistent measure of similarity. It is also bounded by 1 when using probability distributions given the base of the log used is 2. The third is cross entropy. This is our only comparison metric which is not bounded by 0 to 1, and, like Jensen-Shannon divergence, a higher score indicates more dissimilar distributions.

Comparison Algorithm	Accuracy
Baseline	<b>81.4</b>
Jensen-Shannon	60.6
Cross Entropy	60.9
Hellinger	64.2

Table 3: Results per comparison algorithm scores as input to the downvote detection model.

## 5.4 The Downvote Detection Model

We evaluate our approach on the downvote detection task, trained on the comparison scores (see above). The downvote detection classifier consists of a Multi-Layer Perceptron with a softmax output layer, implemented using scikit-learn (Pedregosa et al., 2011). The goal of this classifier is a binary classification of whether a given clip has downvotes (indicating mispronunciation). The input features are the per character pronunciation scores from the distribution comparisons for each phrase. Phrases are of variable length, so the input is padded with ones to the length of the longest phrase. The final parameters are shown in Table 8 in the appendix. We optimized over the parameters using the Adam optimizer. The initial learning rate and beta 1 for Adam were the most impactful. More hidden layers did not improve performance, indicating that a complex network is not necessary for this task.

## 6 Quantitative Evaluation

In Table 3, we compare the accuracy of the downvote detection model when using the different comparison algorithms. The best results, 81.4%, are obtained by the baseline algorithm, using the probability of each character in the elicitation phrase from the speech recognition model’s softmax. This is a binary classification with a 50 : 50 split, i.e., random chance should yield about 50% accuracy. As an upper bound, 81.4% is therefore too low to be reliable. All of our comparison algorithm scorers perform at around 60-64%. They are similar to each other, with the Hellinger algorithm performing best after the baseline. This suggests that elaborate methods are not necessary for producing effective scores of pronunciation.

## 7 Qualitative Analysis

In this section, we probe deeper into the model, the task, and the corpus. If the vote annotation on

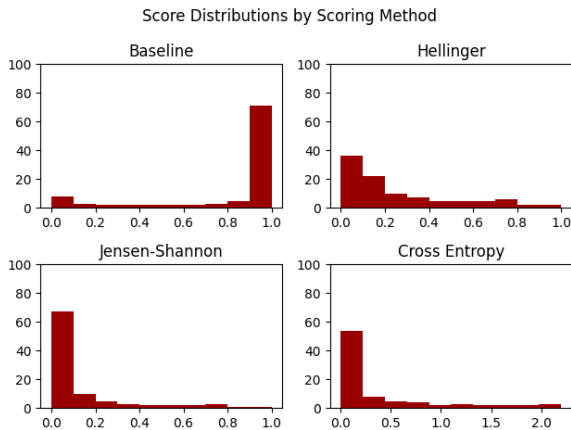


Figure 2: The distribution of pronunciation scores generated by the distribution comparison algorithms as percentages. The x-axis for Cross Entropy is different because it is not bounded by 0-1.

the clips in Common Voice are a reliable indicator of pronunciation quality, that should be reflected in the data. To test this, we choose a subset of instances we consider representative of the broader corpus with regard to both ASR performance and the mix of upvotes and downvotes.

## 7.1 Data in Aggregate

Figure 2 shows the distribution of scores by percent for each of the scoring methods. Each bin contains the output of the distribution comparison algorithm interpreted as a pronunciation score, within the bin’s width of 0.1. Since cross entropy is not bounded by 1, its scores range to 35 for our data. However, such high scores are highly infrequent, thus we do not show scores  $>2$ . The scores generated from instances both with and without downvotes are included in these histograms. Separating the instances by presence of downvote results in nearly identical graphs.

For the baseline algorithm, the majority of scores are in the 0.9-1.0 bin. Since these are the probabilities given by the baseline for the character in the elicited phrase, this indicates that the ASR model is confident and accurate most of the time. This is expected for an English model, especially given the quantity of training data this model was trained on. The baseline model rarely returns intermediate probabilities. Consequently, when it predicts the wrong character or chooses no prediction, it still tends to do so confidently. The Jensen-Shannon scorer presents a similar pattern, the majority of scores are in the bin representing the best

scores. (Since it is a distance metric, 0 represents the highest similarity and therefore a positive pronunciation score.)

The Hellinger scorer differs from the baseline, Jensen-Shannon, and Cross Entropy scorers in that it produces far fewer scores at the extremes of 0 and 1 or greater, instead making more distributed judgements. These differences indicate that some additional information is captured by the Hellinger scorer with regard to the relationship between the baseline and expert productions of the elicited phrase. The baseline scorer outperforming the Hellinger scorer (see Table 3) in our implicit evaluation task indicates that this relationship is not productive in predicting downvotes.

While the distributions in Figure 2 show an overview of the scorers, they do not directly compare the scorers to one another. We are most interested in how the comparison scorers relate to the baseline, as the baseline is representative of the model’s confidence in its transcription. Figure 3 provides a direct comparison of the baseline scorer with the 3 scorers per character in each elicitation phrase. The diagonals provides a point of reference; scores above the diagonal are scored as worse pronunciation by the respective scorer for the same character, and scores below the diagonal are scored as better.

For the comparisons with the Hellinger distance and Jensen-Shannon divergence (top and middle of Figure 3), 1 on the y axis indicates a correct pronunciation, so the diagonal indicating agreement between the comparison and baseline has a negative slope. Most of the points appear below the agreement diagonal, showing that the scorers are more forgiving overall of mispronunciation. On both extremes of the x axis, 0 and 1, there is a broad range of scores on the y axis. As discussed above, this is where the majority of baseline scores appear, especially around 1, which is why the density at those extremes is much higher. From 0.9-1.0 on the x axis, the y axis has points ranging from 0-1, but the majority tend to be low, indicating that the Hellinger scorer tends to agree with the Baseline scorer when the ASR model is confident. There is more disagreement between the scorers at the 0 x axis extreme. This may be influenced by the smaller sample size compared to the 1 extreme, but there are enough points to confirm that the Hellinger scorer is more forgiving when the ASR model has low confidence. Of the non-

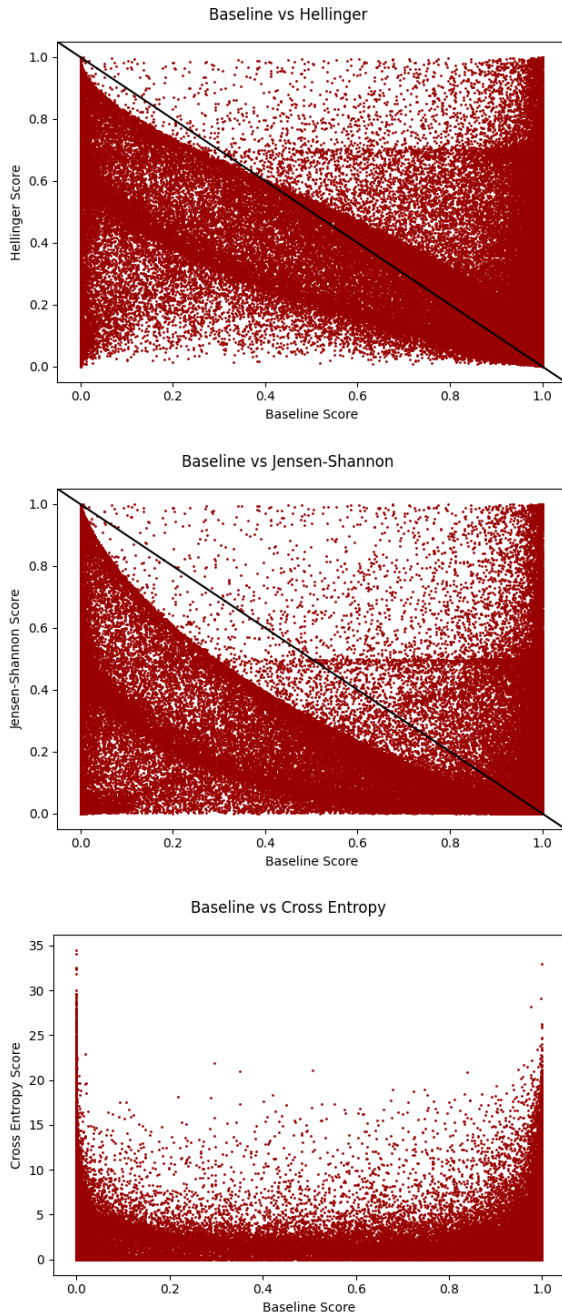


Figure 3: The relationship between scores in the baseline and the Hellinger, Jensen-Shannon, and Cross Entropy scorers. Each point represents a character’s pronunciation score with the baseline on the x axis and the graph’s respective comparison scorer on the y axis. Scores on the diagonal are equally scored by the baseline and the comparison scorer.

Baseline scorers, the Hellinger scorer performed best, which is likely due to the higher agreement it has with the Baseline.

The middle plot in Figure 3 compares the baseline with the Jensen-Shannon Divergence scorer. There is far less agreement in the Jensen-Shannon

scorer than the Hellinger/Baseline comparison in the intermediate scores, but overall the Jensen-Shannon and Baseline scorers compare very similarly, being generally more forgiving when the ASR model has low confidence in its predictions.

Cross Entropy, unlike Hellinger Distance and Jensen-Shannon, is not bounded by 0-1, so there is no agreement diagonal in the bottom plot in Figure 3. Similar to Jensen-Shannon and Hellinger, most of the points are concentrated around the 0 and 1 extremes of the x axis. Because the Cross Entropy scores are on a much larger scale, creating a threshold for a mispronunciation would be at a different value than for the other scorers, and difficult to determine.

## 7.2 Specific Examples

As discussed in Section 3, the dataset used for these experiments is intended and annotated specifically for speech recognition, not for any specific pronunciation or dialect. This is, however, the closest available annotation to our task. The annotations on the audio clips collected indicate whether the speaker in a clip “accurately [spoke] the sentence”, represented as upvotes or downvotes. Downvotes can indicate a mispronunciation, but also frequently indicate bad audio quality or missing audio. Conversely, upvotes do not distinguish between dialects, since a desired characteristic of ASR is the ability to generalize over dialect. We investigate a small number of examples further, relying on the first author’s native American English judgments. In addition to looking into different issues resulting from the data, we are also interested in the question whether the different similarity metrics we used can provide complementary information to the baseline scores.

We take a closer look at individual examples from Common Voice, illustrating a range of issues, see Tables 4, 5, 6, and 7. Scores that show a distance  $> 0.3$  from a perfect pronunciation (0 or 1, depending on the metric) are highlighted in red, indicating a mispronunciation.

Table 4 demonstrates the expected behavior in the case of a mispronunciation. The last word, *feel*, is mispronounced by learner 174840. The *f* is dropped and the *e*’s are pronounced as a lax high front vowels instead of tense. The ASR model is able to correctly transcribe the clip as *how do you feel*, though it reports being nearly equally confident that the last word is *hear*.

	h	o	w	d	o	y	o	u	f	e	e	l
Expert	0.995	0.801	0.962	0.965	0.882	0.918	0.880	0.901	1.000	1.000	0.998	0.999
Baseline	0.992	0.975	0.985	0.919	0.897	0.943	0.973	0.971	0.352	0.752	0.381	0.040
Hellinger	0.031	0.247	0.095	0.083	0.104	0.046	0.148	0.065	0.634	0.132	0.311	0.871
JSD	0.001	0.076	0.012	0.009	0.015	0.003	0.030	0.006	0.439	0.022	0.130	0.848
Cross Entropy	0.047	1.490	0.333	0.256	0.766	0.220	0.707	0.250	1.509	0.413	1.410	4.639

Table 4: Comparing Expert 167006 and Learner 174840.

	h	o	w	d	o	y	o	u	f	e	e	l
Expert	0.999	0.999	0.999	0.981	0.961	0.994	0.994	0.990	0.998	0.999	0.998	0.999
Baseline	0.992	0.975	0.985	0.919	0.897	0.943	0.973	0.971	0.352	0.752	0.381	0.040
Hellinger	0.034	0.097	0.060	0.165	0.119	0.024	0.055	0.034	0.629	0.132	0.309	0.872
JSD	0.001	0.011	0.004	0.033	0.028	0.001	0.004	0.002	0.437	0.022	0.130	0.848
Cross Entropy	0.016	0.047	0.026	0.186	0.347	0.100	0.066	0.118	1.514	0.415	1.406	4.640

Table 5: Comparing Expert 156711 and Learner 174840.

	h	o	w	d	o	y	o	u	f	e	e	l
Expert	0.999	0.999	0.999	0.981	0.961	0.994	0.994	0.990	0.998	0.999	0.998	0.999
Baseline	0.869	0.876	0.643	0.758	0.033	0.484	0.537	0.352	0.871	0.941	0.919	0.859
Hellinger	0.077	0.233	0.349	0.253	0.744	0.464	0.470	0.369	0.234	0.163	0.187	0.076
JSD	0.008	0.061	0.136	0.078	0.645	0.244	0.258	0.177	0.060	0.029	0.038	0.007
Cross Entropy	0.208	0.197	0.640	0.457	5.009	1.051	0.903	1.543	0.208	0.091	0.134	0.224

Table 6: Comparing Expert 156711 and Learner 103321.

While the Hellinger and Jensen-Shannon scorers capture these issues just as the baseline scorer does, the Cross Entropy scorer is much more critical, indicating errors where there are none in the first three words.

Table 5 shows the same learner as in Table 4, but compared with a different expert. The baseline scores are identical to Table 4 because they are independent of the expert. Though the expert scores are high in both Tables 4 and 5, the scores generated by the comparison scorers correctly indicate better pronunciation of the vowels in the first three words, especially in the Cross Entropy comparison. This demonstrates the impact that the selection of the expert has on scoring when using the comparison metrics, especially for the Cross Entropy scores. In the implicit evaluation, the comparison metrics perform worse than the baseline, but the impact of the choice of expert shows that there is at least some potential in those scorers which is not captured by that evaluation.

Table 6 contains an example where the expert speaker speaks clearly and the learner, though sounding native, does not enunciate clearly, so that the ASR model misunderstands *you* in the production, shown by the low scores. In this example, the forgiveness of the Jensen-Shannon scorer cap-

tures better that the learner pronounces the phrase correctly despite their lack of enunciation. The Hellinger scorer and cross entropy scorer closely reflect the baseline. This again shows the potential of the comparison scorers not captured by the implicit evaluation.

In Table 7, the expert speaker pronounces the phrase correctly, but the quality of the audio is very poor, and the ASR model has trouble transcribing the clip, though it is understandable to a native speaker. The learner also pronounces the clip correctly, i.e., the baseline scorer is correct in its feedback. The other scorers, however, incorrectly indicate mispronunciations in the learner’s pronunciation. This is an issue with our expert selection more than with the annotation. However, in the case of this speaker being selected as a learner instead of a speaker, several characters would still be incorrectly marked as mispronounced. Choosing an expert carefully is critical, and in this case, the Common Voice annotation is not reliable enough to do so. Overall, we reveal an issue in our methodology for selecting the expert side of the comparison, specifically that the lack of any downvotes is a poor selection criterion, as poor quality clips may get through the annotation without any downvotes.

	h	o	w	d	o	y	o	u	f	e	e	l
Expert	0.271	0.074	0.000	0.770	0.903	0.973	0.988	0.979	0.997	0.952	0.001	0.000
Baseline	0.992	0.990	0.984	0.984	0.992	0.999	0.999	0.997	0.997	0.995	0.993	0.998
Hellinger	0.434	0.804	0.750	0.283	0.125	0.050	0.047	0.028	0.020	0.109	0.680	0.704
JSD	0.235	0.770	0.572	0.097	0.020	0.003	0.003	0.001	0.001	0.015	0.496	0.499
Cross Entropy	2.675	8.502	2.385	2.218	0.578	0.114	0.095	0.070	0.041	0.384	0.086	0.004

Table 7: Comparing Expert 18456694 (bad quality audio) and Learner 18400454.

## 8 Conclusion & Future Work

Our investigation has shown that the upvote and downvote annotations make a poor substitute for a properly annotated pronunciation corpus. Clips which have native sounding speech also have downvotes because of the poor audio. There is a great deal of variation in dialect and audio quality, which is desirable for training a speech recognition model, but represents noise when grading pronunciation. A downvote is far more commonly used as an indicator of an issue with the file itself than of a mispronunciation. The issue goes both ways as well, many clips with very poor audio quality have no downvotes but are not accurately processed by the speech recognition system. Most clips also have very few votes overall (most commonly 3), which prevents us from using ratios of up- and downvotes.

Many of the issues we identified, especially in section 7, indicate that there is a need for a speech corpus annotated for pronunciation. Many of the problems, such as selection of experts and variation in dialect and audio quality, can only be addressed by a careful collection of data and having clearly defined annotations.

As demonstrated in section 7.2, the comparison scorers still demonstrate some promise. Since the data situation makes it impossible to evaluate our scorers accurately, our next step is to collect a speech corpus annotated for pronunciation. We can then evaluate and continue to develop these scorers.

## 9 Limitations

We recognize that we make several critical assumptions throughout this work necessary to interpret our results: 1) Moses et al. (2020) show that using the ASR softmax probabilities per character is a reasonable way to score goodness of pronunciation. Our results indicate that either our model (see section 4.2) does not capture the relationship between pronunciation and downvotes, or there is

none (the latter possibility being supported by our qualitative analysis). 2) There are no pronunciation corpora available with the type of annotations required for the task. In the absence of such data, we use the closest alternative. While it is possible to create such corpora for e.g. English, it may not be possible for many under-resourced languages. For the latter, using Common Voice may still be the only option. 3) We assume that the comparison metrics used are reliable. However, this can only be tested empirically once we have usable data. Finally, the ASR model is trained solely for speech recognition and not finetuned for the task of pronunciation. As we have no character level annotation to work with, finetuning is not possible in this work.

## References

- Chesta Agarwal and Pinaki Chakraborty. 2019. A review of tools and techniques for computer aided pronunciation training (CAPT) in English. *Education and Information Technologies*, 24:3731–3743.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.
- Harry Bratt, Leonardo Neumeier, Elizabeth Shriberg, and Horacio Franco. 1998. Collection and detailed transcription of a speech database for development of language learning technologies. In *ICSLP*.
- Coqui. 2021. English stt v1.0.0. Technical Report STT-EN-1.0.0, Coqui, <https://coqui.ai/models>.
- Sylvain Coulange. 2023. Computer-aided pronunciation training in 2022: When pedagogy struggles to catch up. In *Proceedings of the 7th International Conference on English Pronunciation: Issues and Practices*, pages 11–22.
- Jonathan Dalby and Diane Kewley-Port. 1999. Explicit pronunciation training using automatic

- speech recognition technology. *CALICO Journal*, 16(3):425–445.
- Horacio Franco, Victor Abrash, Kristin Precoda, Harry Bratt, Ramana Rao, John Butzberger, Romain Rossier, and Federico Cesari. 2000. The SRI Edu-Speak(TM) system: Recognition and pronunciation scoring for language learning. In *Proceedings of In-STILL*, pages 123–128.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 369–376.
- Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.
- Ernst Hellinger. 1909. Neue Begründung der Theorie quadratischer Formen von unendlich vielen Veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271.
- Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.
- Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.
- Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: A survey. *Multimedia Tools and Applications*, 80(6):9411–9457.
- Caleb Moses, Miles Thompson, Keoni Mahelona, and Peter-Lucas Jones. 2020. Scoring pronunciation accuracy via close introspection of a speech recognition recurrent neural network. In *NeurIPS2020*.
- Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.
- Ambra Neri, Catia Cucchiari, and Helmer Strik. 2006. [ASR-based corrective feedback on pronunciation: Does it really work?](#) In *Proceedings of Interspeech*.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*.
- Silke M Witt. 2012. Automatic error detection in pronunciation training: Where we are and where we need to go. In *International Symposium on Automatic Detection on Errors in Pronunciation Training*.
- Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. [L2-ARCTIC: A non-native English speech corpus](#). In *Proceedings of Interspeech*, page 2783–2787.



## A Model Parameters

Best Model Parameters	
input embedding	152
hidden layer size	128, 64, and 32
activation	ReLU
optimizer	Adam
batch size	200
learning rate	5e-4
Adam beta 1	0.80

Table 8: Optimized model parameters for the implicit evaluation.

# OPINIONS ARE BUILDINGS: Metaphors in Secondary Education EFL Essays

Anna Hülsing<sup>1</sup>, Andrea Horbach<sup>2,3,4</sup>

<sup>1</sup>University of Hildesheim, Germany

<sup>2</sup>Leibniz Institute for Science and Mathematics Education, Kiel, Germany

<sup>3</sup>University of Kiel, Germany, <sup>4</sup>FernUniversität in Hagen, Germany

huelsing@uni-hildesheim.de

horbach@leibniz-ipn.de

## Abstract

Automatic metaphor detection has been an active field of research for years. Yet, it was rarely investigated how automatic metaphor detection can aid language learning. We therefore present MEWSMET, a corpus of argumentative essays (MEWS<sup>1</sup>) written by English as Foreign Language (EFL) learners annotated for metaphors. We differentiate between two kinds of metaphors: metaphors that are comprehensible to native speakers, even though they themselves would not use them (comprehensible metaphors, CMs) and metaphors that native speakers would use (target language metaphors, TLMs). We use MEWSMET in two ways: Firstly, we analyze our annotations and find out that there is a positive linear correlation between essay score and the number of TLMs, while no correlation is found between essay score and the number of CMs. Secondly, we explore how metaphor detection models perform on MEWSMET. We find that metaphor detection is a hard task given our noisy learner data, and that metaphor detection models tend to be better at identifying all metaphors (TLMs+CMs) instead of just TLMs, even though only TLMs can be used as a feature for automatic essay-scoring.

## 1 Introduction

Conceptual Metaphor Theory claims that metaphorical linguistic expressions manifest our way of thinking. One of the most well-known examples for a metaphorical linguistic expression is *to spend time*. Here, the conceptual domain TIME

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

<sup>1</sup>Measuring Writing at Secondary Level (see Keller, 2016 and Keller et al., 2020)

is described by means of the conceptual domain MONEY. The metaphorical linguistic expression thus shows that time is considered a limited and valuable resource (Lakoff and Johnson, 1980b). Metaphorical linguistic expressions are therefore not merely ornamental, but omnipresent in our everyday life (Lakoff and Johnson, 1980a, Shutova and Teufel, 2010).

Detecting metaphorical linguistic expressions automatically is beneficial for a range of natural language processing applications, such as emotion detection (Dankers et al., 2019, Li et al., 2022), identification of mental health problems (Zhang et al., 2021, Gutiérrez et al., 2017), or propaganda detection (Baleato Rodríguez et al., 2023). Even though metaphors play an important role in education (Niebert and Gropengiesser, 2012, Mouraz et al., 2013, Oxford et al., 1998), it is only rarely investigated how metaphor detection (MD) can be employed to facilitate language learning.

Beigman Klebanov et al. (2018) have presented a corpus annotated for metaphors that is based on the ETS Corpus of Non-Native Written English<sup>2</sup> – a collection of argumentative essays provided by TOEFL test takers. They show that the use of argumentation-relevant metaphors provides information about a writer’s English language proficiency. We build on and extend this work in several ways as detailed in the following.

First, our study addresses whether the same relation between metaphoric language use and language proficiency also holds for younger writers. Although mean age of the writers in the study by Beigman Klebanov et al. (2018) is not given (Blanchard et al., 2013), we assume that – as

<sup>2</sup><https://catalog.ldc.upenn.edu/LDC2014T06>

---

[Children] are likely to take over (*adopt*) the opinion of the people [...] that are around them.

---

Young children should live their lives and should not have to build (*form*) their own opinion about something.

---

This often brings (*puts*) parents in difficult situations.

---

Table 1: Example sentences with metaphorically used verbs (underlined) taken from MEWS data (Keller et al., 2020). They are comprehensible in English, even though L1 English speakers would probably use different expressions such as the ones given in brackets.

TOEFL tests are often taken by students who want to study at a university where English is the language of instruction – most writers are in their last year of high-school or have recently graduated from high-school. In contrast, our study is based on the MEWS dataset by Keller et al. (2020), which addresses German-speaking EFL learners in earlier years of their education, while also using TOEFL writing prompts<sup>3</sup>. We assume that the general proficiency level will be lower in our dataset than in the one by Beigman Klebanov et al. (2018). In addition, our dataset comprises essays of all proficiency levels, while the one by Beigman Klebanov et al. (2018) only consists of medium- and high-proficiency essays.

Secondly, we investigate the relationship between proficiency level and metaphors that English L1 speakers comprehend, even though they themselves would not actively use them; examples are shown in Table 1. Samaniego Fernández et al. (2005) demonstrate that professional translators introduce new expressions and conceptual structures in a target culture when transferring metaphors that are non-novel in the source language to a novel metaphor in the target language. The translated expressions “seem to have been understood correctly, and this proves their [i.e. the metaphors’] transparency: they can be interpreted precisely because they appeal to our recognition of underlying symbolism.” In our dataset, students also use metaphors that seem anomalous in the target language in the sense that L1 speakers would not use them. Yet, the metaphorical expressions are perfectly comprehensible for target language speakers because they create new (and sometimes even appealing) conceptual mappings

<sup>3</sup>The prompts are different from those used in the ETS Corpus of Non-Native Written English, i.e. also different from the TOEFL dataset by Beigman Klebanov et al. (2018).

(e.g. *to build an opinion*: an opinion is – or should be – hard work just as building a house)<sup>4</sup>. We will call these metaphors **comprehensible metaphors** (CMs), as opposed to metaphors which target language speakers would actively use (**target language metaphors**, TLMs). We will examine the scores human raters gave to essays containing CMs in order to find out whether they rather occur in low- or high-proficiency essays.

Next, we investigate how well metaphor detection models perform on more noisy data from such younger, and partly less-proficient writers in detecting metaphors – both CMs and TLMs. To do so, we leverage the best-performing model from the 2020 Shared Task on Metaphor Detection (Leong et al., 2020), namely DeepMet (Su et al., 2020). Our study will focus on verbs only for several reasons. First, Cameron (2003) report that about half of all metaphors in educational discourse are found in verbs. Second, other parts of speech, especially prepositions, are often not seen as being metaphorical by laypeople (cf. Beigman Klebanov and Flor, 2013), which would pose an additional difficulty during the annotation process. Third, many metaphor detection datasets that potentially serve as training data, have been annotated just for verbs.

Our study makes the following contributions: **1)** We present the MEWSMET corpus. Here, an additional layer is added to the MEWS-dataset (Keller et al., 2020), where we annotated metaphors that are perfectly acceptable in the target language English (TLMs) as well as metaphors which are comprehensible but which native speakers would not use (CMs). **2)** We describe the relationship between TLMs and the scores human raters attributed to the student essays. We do so to confirm the trend Beigman Klebanov et al. (2018) have observed for high-school graduates also for younger and less-proficient students, namely that the use of metaphors provides insights into a learner’s proficiency level. **3)** We describe the relationship between CMs and students’ proficiency levels. **4)** We provide insights into the behaviour of metaphor detection models on noisy learner data for both TLMs and CMs.

For code and data see [https://github.com/AnHu2410/MEWSMET\\_code](https://github.com/AnHu2410/MEWSMET_code).

<sup>4</sup>The expression *to build an opinion* is based on a false friend, as the German equivalent to *to form an opinion* is (*sich*) *eine Meinung bilden*, where the word *bilden* is phonologically similar to the English verb *to build*.

## 2 Related Work

In this section we provide the scientific background to the three main fields of this study: metaphor annotation, metaphor detection and automatic essay scoring.

### 2.1 Metaphor Annotation

A widely applied example of a metaphor annotation guideline is the Metaphor Identification Procedure (MIP; [Pragglejaz Group, 2007](#)) and its extension, MIPVU ([Steen et al., 2010](#)). The underlying idea is that a token is used metaphorically if its meaning in a certain context deviates from a more “basic” meaning of this word, as defined by a contemporary dictionary. For example, the basic (i.e. first) meaning of the verb *to build* according to the online version of the Longman Dictionary of Contemporary English<sup>5</sup> is *to make something, especially a building or something large*, with examples ranging from houses and bridges to birds’ nests. In the expression *to build an opinion*, clearly this concrete basic meaning is not applicable.

We follow the annotation guideline from [Mohammad et al. \(2016\)](#). It is based on MIP ([Pragglejaz Group, 2007](#)), but condensed and enriched with examples (see Appendix A.1.1), which we deemed suitable for our annotators who had no prior experience in the identification of metaphors. While MIP and MIPVU were originally designed for annotating metaphors in English, there have been attempts to use the guidelines for other languages such as Spanish ([Sanchez-Bayona and Agerri, 2022](#)).

[Beigman Klebanov et al. \(2018\)](#) annotate argumentation-relevant metaphors, i.e. metaphors that help the writer of an argumentative essay to advance an argument. In stark contrast to [Pragglejaz Group \(2007\)](#) and [Steen et al. \(2010\)](#), they did not provide “formal definitions of what a literal sense is in order to not interfere with intuitive judgments of metaphoricality” ([Beigman Klebanov and Flor, 2013](#)). This line of thought also emerges in other annotation studies, such as [Tsvetkov et al. \(2014\)](#) and [Piccirilli and Schulte im Walde \(2022\)](#), as they rely on intuitive definitions of metaphoricality.

The distinction between – and annotation of

---

<sup>5</sup><https://www.ldoceonline.com>. We use this corpus-based dictionary for our annotation since it was also used by [Steen et al. \(2010\)](#).

– novel and conventionalized metaphors is an increasingly active research topic ([Parde and Nielsen, 2018](#), [Do Dinh et al., 2018](#), [Egg and Kordoni, 2022](#), [Reimann and Scheffler, 2024a](#)). This distinction is also relevant for our dataset, and has been annotated for future use. Another distinction which is highly relevant for our study is given by [Reijnierse et al. \(2018\)](#), who in their annotation protocol differentiate between deliberately and non-deliberately used metaphors. After all, deliberately used metaphors cannot simply be learnt from a textbook and could therefore hint at a higher language competency. We have not annotated whether or not metaphors in our dataset are used deliberately, but leave this to future work.

### 2.2 Metaphor Detection

An early approach to automatic metaphor detection was developed by [Birke and Sarkar \(2006\)](#), who used a word-sense disambiguation approach to classify literal and non-literal usages of verbs. Conceptual Metaphor Theory ([Lakoff and Johnson, 1980b](#)) claims that metaphors transfer knowledge from a concrete, familiar domain to a more abstract domain. Therefore, [Turney et al. \(2011\)](#) used abstractness scores of context words as features for their logistic regression classifier. The idea of “conceptual features” also inspired [Tsvetkov et al. \(2014\)](#) and [Köper and Schulte im Walde \(2016\)](#), who – in addition to abstractness and other scores – used semantic supersenses from WordNet ([Miller, 1994](#)) and scores representing distributional fit, respectively.

Early neural models, such as [Do Dinh and Gurevych \(2016\)](#) (a multilayer perceptron with word embeddings), showed a performance comparable to non-neural classifiers; however, they became popular because they did not require feature engineering. Later neural models clearly outperformed the non-neural classifiers: [Dankers et al. \(2019\)](#) used several multi-task learning models and reached state-of-the-art results in 2019 for both metaphor and valence/arousal/dominance (VAD) prediction. During the 2020 Metaphor Detection Shared Task ([Leong et al., 2020](#)), DeepMet overall performed best ([Su et al., 2020](#)); the authors transformed metaphor detection into a reading comprehension task and observed state-of-the-art results. We use this model in our study to compare its performance on the corpus by [Beigman Klebanov et al. \(2018\)](#) and our corpus.

Ma et al. (2021) fine-tuned BERT for MD. To perform word-based metaphor classification, they copied the input sentence and masked the target word. The original sentence and the masked copy were used as input for a sequence classification task. Uduehi and Bunesco (2024) also mask the target word, and compute the expectation of a literal meaning in the given context. Then, they compute the estimation of the realized meaning of the target word in order to predict whether the target word violates the expectation of a literal word. Li et al. (2023) exploited the fact that many datasets are based on the Metaphor Identification Process (MIP; Pragglejaz Group, 2007), where a word is annotated as metaphorical if its contextual meaning is dissimilar to its “more basic meaning” (among further criteria). While prior models (such as MelBERT by Choi et al. 2021) grounded on MIP use decontextualized representations of the target word, Li et al. (2023) successfully gathered the representation of the target word from sentences where it was used literally. While research on metaphors in English has received a lot of attention, and also metaphor detection in other low- to high-resource languages is turning into an active field (Aghazadeh et al., 2022, Lai et al., 2023, Schuster and Markert, 2023), research on metaphors in texts of English learners is rare. Stemle and Onysko (2018) used a bidirectional RNN and fastText embeddings to detect metaphors for the 2018 Shared Task on Metaphor Detection (Leong et al., 2018). As training data for their embeddings they use TOEFL tests (Blanchard et al., 2013) of different proficiency levels (among others); in contrast to our study (and that of Beigman Klebanov et al., 2018), they use these texts to detect metaphors in standard language and not in learner language.

### 2.3 Metaphors in Automatic Essay Scoring

In automatic essay scoring, the task is to predict the quality of an essay either on a holistic scale or for specific aspects of an essay such as language or structure. For holistic scoring, both linguistic form and content are usually taken into consideration and the correct usage of metaphoric expressions can be seen as one aspect of linguistic proficiency. Yet to the best of our knowledge, metaphors have so far not been integrated into automated essay scoring systems, as – until some years ago – it has been claimed that the automatic metaphor de-

tection for non-conventionalized metaphors would not work reliably enough (Graesser and McNamara, 2012). For essay scoring in Chinese, Yang et al. (2019) used a number of features, including the number of metaphors. Given their examples, though, their notion of metaphors rather corresponds to a simile with specific lexical items marking their occurrence. However, there have been recent successes integrating features on the related topic of concreteness of multi-word expressions into essay scoring (Wilkins et al., 2022) highlighting the importance to consider complex linguistic phenomena.

## 3 Annotation Study: Metaphor Annotation in Learner Essays

The goal of the following annotation study is twofold: First, we aim at investigating the relationship between essay scores and the use of metaphors. Second, our annotation results are used as train and test sets in the subsequent experimental study on automatic metaphor detection in learner texts.

### 3.1 Annotation Data

The dataset used in our study is a subset of the MEWS dataset by Keller et al. (2020), a collection of argumentative essays written by German and Swiss EFL learners. The essays are written on the basis of four different TOEFL prompts, two of them being independent prompts (the students are given a prompt only) and two of them being source-based, i.e. the prompt refers to a reading text. We focus on the independently-written essays only as source-based essays might mainly contain metaphors in standard language adopted from the text. The following two prompts were used. The students were asked whether they “agree or disagree with the following statement”:

- *Television advertising directed toward young children (aged two to five) should not be allowed.* (Prompt “TV-Ads”)
- *A teacher’s ability to relate well with students is more important than excellent knowledge of the subject being taught.* (Prompt “Teacher”)

For each essay, expert raters’ scores are available. Two raters scored the essays on a scale between one and five (with five being the best score). If the two ratings were only one point apart (e.g. rater A: 3; rater B: 4), the average was taken as the

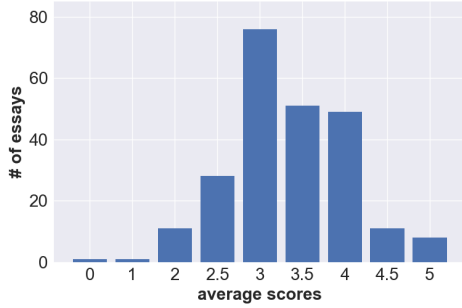


Figure 1: Number of essays per score in MEWSMET.

overall score. Otherwise, a third adjudicator rated the essay in order to obtain the overall score. The essays were written in the penultimate year before graduation; half of them at the beginning of the school year (T1) and half at the end (T2).

We randomly selected 236 essays (120 with prompt “TV-Ads”, 116 with prompt “Teacher”) from Swiss students; Figure 1 shows the number of essays per score. As can be seen, all proficiency levels are taken into account. These essays contain 8025 target verbs (excluding stop words, see Appendix B) which were automatically detected with the off-the-shelf NLTK POS-tagger which utilizes the Penn Treebank tagset (Bird and Loper, 2004).

### 3.2 Annotation Guidelines

Our goal is the annotation of verbal metaphors in learner essays. Our guidelines were adopted from Mohammad et al. (2016), who provide specific definitions for metaphorical and non-metaphorical usages compared to guidelines that rely on intuition (cf. Beigman Klebanov and Flor, 2013). We deemed this kind of guidance helpful for this structurally difficult task. In contrast to Beigman Klebanov et al. (2018), who only focus on argumentation-relevant metaphors, we chose to annotate all verbal metaphors in order to have more comprehensive material for analysis. In addition to a binary decision for metaphorical vs. literal usage, annotators had to label each target verb with one of the following four labels (examples taken from MEWSMET):

- **non-metaphorical:** for literal usages, e.g. *children learn in their small age to consume and to spend money*
- **conventional metaphor:** for frequent metaphorical usages the annotator has seen before, e.g. *Sometimes you spend even more*

*time with a particular teacher than your parents.*

- **creative metaphor:** when the annotator felt that the verb was metaphorical but rarely used in this context, e.g. *[TV-]channels [are] flooded with tons of ads.*
- **uncommon translation of a German conventionalized metaphor:** a metaphor that has a German conventionalized metaphor as basis, but the English translation used here is uncommon, e.g. *This often brings parents in difficult situations.* (literal translation of the following German sentence: *Das bringt Eltern oft in schwierige Situationen.*)

The guidelines can be found in Appendix A.1.1.

### 3.3 Annotation Procedure

The annotation procedure was conducted in three stages using the annotation platform INCEPTION (Klie et al., 2018) as detailed in the following.

**Phase 1 – Sample Annotation by Experts:** Annotating metaphors is generally considered a difficult task with rather low inter-annotator agreement. For example, Reimann and Scheffler (2024a) report a Cohen’s  $\kappa = 0.60$  for annotating metaphors in religious online forums after discussing disagreements and adjudication, and Beigman Klebanov et al. (2018) report a Cohen’s  $\kappa = 0.56$  after a first round of annotation and a Cohen’s  $\kappa = 0.62$  after showing the annotators their partner’s annotations for essays with high disagreement values, and asking them to reconsider their original annotations. Annotating not standard language, but learner language adds an extra layer of difficulty. To check the feasibility of the task and the quality of our annotation guidelines, we first asked two Swiss-German researchers in the field of English didactics to annotate a small subset (5 essays) sampled from MEWS that is not part of the subset described above (Section 3.1). The annotators were given the main annotation guideline as presented in Appendix A.1.1.

In this first annotation round, inter-annotator agreement was low (Cohen’s  $\kappa = 0.22$ ). Therefore, we discussed unclear cases and extended the main guideline (see Appendix A.1.2) to improve their clarity. Based on these improved guidelines, the experts annotated a second sample of 5 essays; as Cohen’s  $\kappa$  increased to a value of 0.37, we con-

sidered the guideline additions to be useful. Of course, the inter-annotator agreement still was not even moderate; however, given the difficulty of the task, we considered it to be sufficient for a first round of annotations.

**Phase 2 – Main Annotation Study:** Next, 236 essays taken from the MEWS corpus (see Section 3.1) were annotated by two annotators, who are pursuing their master’s degrees to become English teachers in Germany. For the purpose of training, they first annotated a MEWS-based toy corpus on the basis of our revised guidelines and discussed the results. Both annotators independently annotated the actual data. For adjudication after the first round of annotations, Annotator A was given the information whether her annotation differed from the annotations of Annotator B. The difference can be one of the following:

- A: metaphor, B: no / uncommon metaphor
- A: uncommon metaphor, B: metaphor / no metaphor
- A: no metaphor, B: uncommon metaphor/ metaphor

I.e., the nature of the difference was not disclosed to the annotator and the difference between creative and conventional metaphor was not taken into account at all.

Annotator A was asked to check these cases and correct them if she made an obvious mistake. After that, Annotator B did the same for the remaining disagreements. Finally, the first author of this paper manually checked all annotations and discussed cases which possibly contradicted the annotation guidelines with Annotators A and B.

**Phase 3 – Check by Native Speakers:** Two English L1 speakers (one American English, one British English speaker) were asked to check whether the metaphors found in Phase 2 were a) expressions that a L1 English speaker might use, b) that an L1 English speaker would not use but which are comprehensible, and c) that are incomprehensible. To avoid bias by language errors surrounding the metaphorical expression, the sentences were corrected and only the relevant part of the sentence was shown to the annotators.

### 3.4 Annotation analysis

#### 3.4.1 Inter-Annotator Agreement

After the first round of the main metaphor annotation study, agreement for the binary decision

		Ann B	
		met	non
Ann A	met	362	57
	non	32	7574

Table 2: Confusion matrix illustrating the inter-annotator agreement for the binary metaphor annotation task (*metaphorical* vs. *non-metaphorical*).

between metaphorical and non-metaphorical was moderate with Cohen’s  $\kappa = 0.42$ . As mentioned before, metaphor annotation generally is a field with rather low inter-annotator agreement, and using learner essays from all proficiency levels poses an additional difficulty. Therefore, the low level of agreement after the first round was to be expected. After the final round, Cohen’s  $\kappa$  reached a high value of 0.88. The confusion matrix for the binary decision is shown in Table 2; even though for 89 target verbs the annotators did not agree, for the vast majority they agreed in their annotations. For 362 target verbs they agreed that they are used metaphorically.

Agreement for the 4-way-task (“conventionalized”, “creative”, “uncommon”, “no metaphor”) was lower with Cohen’s  $\kappa = 0.74$ , and the annotations are represented in the confusion matrix shown in Table 3. While agreement on non-metaphorical expressions is very high and they mostly agreed on metaphors that are based on German conventionalized expressions that are uncommon in English, disagreement was high for whether a metaphor is creative or conventional. The distinction between creative and conventional is hard even for native speakers (compare Parde and Nielsen, 2018); as our annotators are not native speakers, the distinction is even harder, because they are not as familiar with certain conventionalized expressions as native speakers are.

		Ann B			
		conv	creat	unc	non
Ann A	conv	183	6	6	28
	creat	94	34	4	27
	unc	3	3	29	2
	non	23	4	5	7574

Table 3: Confusion matrix illustrating the inter-annotator agreement for the metaphor annotation (4-way-annotation: *conventional*, *creative*, *uncommon*, *non-metaphorical*).

		Ann 2		
		incompr	compr	L1
Ann 1	incompr	3	6	1
	compr	6	20	1
	L1	7	91	227

Table 4: Confusion matrix illustrating the inter-annotator agreement for the check by native speakers (*incomprehensible metaphor, comprehensible metaphor, L1 metaphor*).

For the native speaker check, i.e., the 3-way annotation whether a metaphor was L1-like, comprehensible or incomprehensible, Cohen’s  $\kappa$  reached a value of 0.24. This rather low value is mostly caused by the fact that more metaphors were considered L1-metaphors for Annotator 1 than for Annotator 2 (see Table 4). Annotator 1 was more tolerant towards metaphors such as *to fall into a down* (meaning: *to become depressed*) that could be seen as a creative invention of the writers. This may be due to Annotator 1’s Bachelor’s degree in English Language and Creative Writing (Annotator 2 had no background relevant for the task).

### 3.4.2 Quantitative Analysis

We collected annotations for a total of 8025 target words in 236 essays. We only counted those target verbs as being used metaphorically in the subsequent analyses where both annotators agreed that the verb was metaphorical, i.e. where they chose one of the following labels: conventional metaphor, creative metaphor, or uncommon translation of a German conventionalized metaphor. This was the case for 362 verbs. These 362 verb tokens consisted of 149 types. We did not perform adjudication for the individual labels (e.g., conventional metaphor), as we only take into account the binary label (metaphorical, non-metaphorical) in this study.

The 362 verbs that both German annotators annotated as being metaphorical were shown to the English native speakers. For their check, we decided to err on the side of caution and use the least optimistic label, i.e. if one annotator decided that an expression is incomprehensible while the other decided it was comprehensible, we chose the label “incomprehensible”. 23 target verbs were annotated as being incomprehensible by at least one annotator. These target verbs were counted as being non-metaphorical, even though the writer might have intended to use a metaphor here. 112 tar-

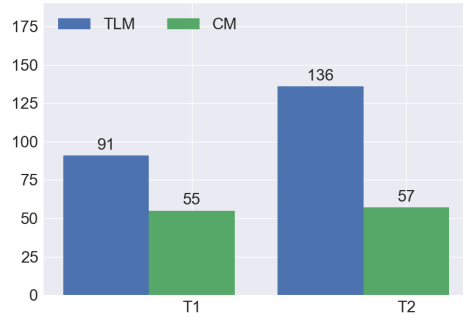


Figure 2: Number of CMs and TLMs found at beginning (T1) and end of school year (T2).

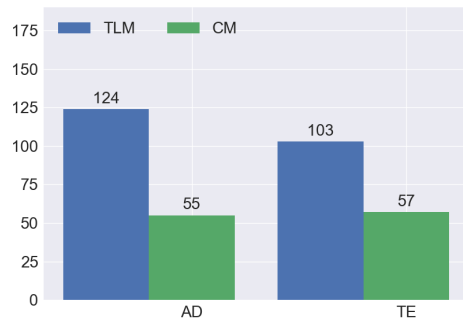


Figure 3: Number of CMs and TLMs found for the prompt “TV ads” (AD) and for the prompt “Teacher” (TE).

get verbs were annotated as being comprehensible (CMs). For 227 metaphorical expressions both annotators declared that they could have been uttered by an English L1 speaker (TLMs).

Figure 2 shows the amount of metaphors (TLMs and CMs) found at T1 and T2, respectively. As can be seen, the amount of CMs stays nearly the same for T1 and T2, but the amount of TLMs rises by 50%. This indicates that the learners’ proficiency improves within one year, and that TLMs could be a useful feature in essay scoring, whereas CMs might not be.

Figure 3 shows the amount of metaphors (TLMs and CMs) found for each prompt. While slightly more TLMs occur in the essays on TV-Ads than on Teachers, the number of CMs is roughly the same for both prompts. The balance between both prompts is important, since we split the entire MEWSMET-corpus into two parts (MEWS\_Ads and MEWS\_Teacher), and use them as training and testing data.

### 3.4.3 Relationship between Metaphors and Essay Quality Scores

In order to investigate the relationship between the number of metaphors per essay and the essay’s



holistic score, we counted the number of TLMs and CMs in each essay, and normalized the number of TLMs and CMs by the number of each essay’s characters in order to control for essay length. Then we obtained the score (1 to 5) attributed to each essay by expert raters. The correlations between the number of TLMs, CMs as well as all metaphorical expressions (TLMs plus CMs), each divided by the number of characters, and the respective essay scores in terms of Pearson’s  $\rho$  are presented in Table 5.

The results show that there is a weak, yet significant positive linear correlation (p-value < 0.05) between essay score and the number of metaphors that English L1 speakers would use (TLMs). No correlation between the essay scores and the number of comprehensible metaphors (CMs) was observed. The combined correlation between essay score and all metaphorical expressions (both TLMs and CMs) is weak, and this correlation is not significant.

	Pearson’s $\rho$	p-value
TLM/score	0.143	0.028
CM/score	-0.011	0.863
both/score	0.118	0.070

Table 5: Pearson’s correlation between target language metaphors / comprehensible metaphors / both types of metaphors combined (controlled for essay length) and essay score.

## 4 Experimental Study: Automated Metaphor Identification in Learner Essays

After having manually identified metaphors in the previous section, we now turn to the question of how well existing metaphor detection algorithms perform on MEWS learner data using our annotations as gold standard.

### 4.1 Experimental Setup

#### 4.1.1 Classifier

We use DeepMet (Su et al., 2020) to detect metaphors in learner text. DeepMet transforms metaphor detection into a reading comprehension task, i.e. the model is trained to answer questions based on a given sentence. Their model takes the global context (i.e. the whole sentence), local context (i.e. the words before and after the target word

that are enclosed by punctuation such as commas) and two types of part-of-speech as features, which are represented via BERT embeddings. These embeddings are fed into a siamese architecture based on two Transformer encoder layers. Their output is reduced to one feature vector by average pooling, which is the input to a metaphor discrimination layer. We chose this model as it showed the best performance in the 2020 metaphor detection Shared Task (Leong et al., 2020).

#### 4.1.2 Evaluation Procedure

We use the evaluation procedure presented in Su et al. (2020), where stratified 10-fold cross-validation is performed. In each fold, a model is trained based on a subset (90%) of the training data and used to make predictions on the entire set of the test data. The predictions for all training folds are summed up (leading to a number between 0 and 10 for each test instance  $i$ ). This sum is divided by the number of folds (in our case 10). A metaphor preference parameter  $\alpha$  (determined in previous experiments) indicates which prediction is the final prediction for each test instance. The default value is 0.2, so if at least two models predicted instance  $i$  to be metaphorical, the final prediction is metaphorical; else, the final prediction is non-metaphorical.

#### 4.1.3 Training Data

As mentioned before, we use two splits of MEWS-MET, namely MEWS\_Teacher and MEWS\_Ads. We train and test in both directions, i.e. we train on MEWS\_Teacher and test on MEWS\_Ads and vice versa. In addition to the data we annotated ourselves, we use two other datasets: firstly, a very large corpus annotated for metaphors, namely the VU Amsterdam Metaphor Corpus (VUA) by Steen et al. (2010). This corpus is sampled from the British National Corpus (BNC) and covers academic texts, conversation, fiction, and news texts, which means that it contains standard English. The data was annotated under the MIPVU protocol (Steen et al., 2010). Secondly, we use the TOEFL corpus which, as mentioned before, is sampled from the ETS Corpus of Non-Native Written English, and contains argumentative essays written by EFL learners shortly before or after graduating from secondary education. Even though this corpus is not as big as the VUA corpus, it contains learner language similar to MEWSMET. Only argumentation-relevant

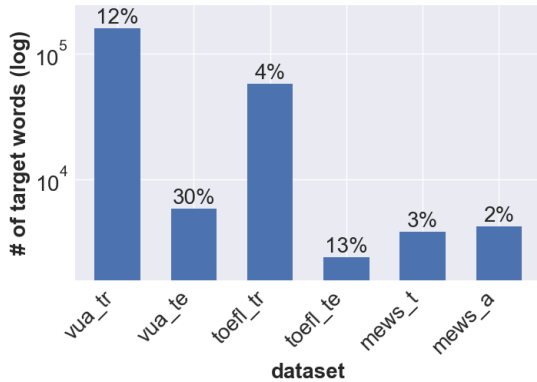


Figure 4: Amount of training (*tr*) and testing data (*te*) for VUA and TOEFL compared to size of MEWS.Teacher and MEWS.Ads on a logarithmic scale. Percentage of metaphorically used target verbs given on top of each column.

metaphors were annotated here (Beigman Klebanov et al., 2018). For both VUA and TOEFL, we used metaphor annotations for all parts of speech for training, and we evaluated on the datasets where only verbs are annotated for metaphoricity (as done by Su et al., 2020). The stark differences in the amounts of training and testing instances for the two additional corpora, compared to our dataset, are illustrated in Figure 4.

#### 4.1.4 Computing Hours and Infrastructure

It took about 30 hours to train the VUA model, 12 hours to train the TOEFL model and 2 hours to train the MEWS models. Experiments were performed on an AMD EPYC 74F3 24-Core Processor and NVIDIA RTX A6000 GPUs.

## 4.2 Performance of Metaphor Detection Method on MEWSMET

If we want to use metaphors as features for automatic essay scoring, they have to be detected automatically and reliably. Therefore we investigate how well metaphor detection models perform on noisy student data such as MEWS.

### 4.2.1 Experiment 1: Metaphor detection performance across different datasets

To assess how hard the task of metaphor detection is on our dataset compared to existing metaphor datasets, we compare performance across datasets when training and testing on data from the same dataset. Results for training and evaluating the DeepMet model on VUA and TOEFL data are reported in Su et al. (2020); to ensure comparability

with our results on MEWS data we repeated the experiments on our own GPU machines.<sup>6</sup>

To compare the performance on VUA and TOEFL to our data, we first used the MEWS\_Teacher split of our data for training and MEWS\_Ads for testing, and secondly MEWS\_Ads for training and MEWS\_Teacher for testing. In both datasets, we only considered TLMs, since we assumed that this metaphor type is closer to the metaphors annotated in VUA and TOEFL. The results are shown in Table 6. The hyperparameters were taken from the paper by Su et al. (2020) with seed = 12.

	Precision	Recall	F1
<b>VUA</b>	70.9	81.9	76.0
<b>TOEFL</b>	64.1	82.8	72.3
<b>MEWS_T</b>	35.8	19.4	25.1
<b>MEWS_AD</b>	56.3	8.7	15.1

Table 6: Results for training on VUA / TOEFL / MEWS data and testing on the corresponding test data. For MEWS, the training data is mentioned in the table (e.g. MEWS\_T refers to training on MEWS\_Teacher and evaluating on MEWS\_Ads). Here, only TLMs were taken into account. The results are determined with the preference parameter  $\alpha = 0.2$ .

In this evaluation setup, DeepMet performs best on the VUA data, closely followed by the TOEFL data. On MEWS\_Ads and MEWS\_Teacher it performs worst by a large margin.

In the course of the evaluation we observed that for the two MEWS test datasets the results also varied greatly across different training folds, while this was not the case for VUA and TOEFL data. Table 7 shows the mean and standard deviation (SD) of precision, recall and F1 across all folds for the 4 models. Here, precision, recall and F1 are calculated for each fold without using the preference parameter  $\alpha$ .

While the F1 standard deviation for VUA and TOEFL is lower than 2 F1-points, for MEWS-MET (trained on MEWS\_Teacher) it is 10.7 points and 6.5 (trained on MEWS\_Ads). During cross-validation, the test data stays the same, and as 90% of the training data are used for each fold, the difference between the individual folds does not vary largely either. The best guess is that the extreme

<sup>6</sup>Su et al. (2020) report  $F1 = 80.4$  for VUA-verb and  $F1 = 74.9$  for TOEFL-verb. We attribute differences to our results to slightly different GPU settings.

	Precision	Recall	F1
VUA	77.6 ± 2.1	69.6 ± 3.9	73.3 ± 1.6
TOEFL	72.0 ± 4.8	64.9 ± 6.5	67.9 ± 1.8
MEWS_T	34.3 ± 33.6	8.2 ± 11.1	10.3 ± 10.7
MEWS_AD	52.0 ± 32.4	4.2 ± 4.3	7.1 ± 6.5

Table 7: Mean and SD scores across precision, recall and F1 for test data on each training fold without using the preference parameter  $\alpha$ . The folds are identical with the ones used for Table 6.

differences in training data size account for this behaviour. In order to see whether the training dataset is indeed too small for the model to learn properly, we shrunk the VUA training dataset to a size comparable to MEWS\_Teacher; the Mini-VUA consists of 3600 training instances, of which 100 are tagged as being metaphorical. The result for training on Mini-VUA and testing on the VUA test dataset is a precision of  $26.6 \pm 36.8$ , a recall of  $0.2 \pm 0.5$ , and an F1-score of  $0.4 \pm 0.1$ . These numbers show that the model does not learn at all from Mini-VUA. We therefore expect our models to perform better with a larger amount of training data, too.

#### 4.2.2 Experiment 2: Model Performance for Different Training Datasets

As discussed above, larger amounts of training data are needed for DeepMet to perform well on MEWSMET. Therefore, we investigated which training data is most suitable for our task of detecting metaphors in learner language – a very large corpus based on standard English (VUA), or a medium-sized corpus based on EFL data (TOEFL). The evaluation described above was also applied here; again we used the hyperparameters from the paper (Su et al., 2020) with seed = 12. Whereas for the previous experiment we focused on TLMs only, here we present the results for TLMs only versus all metaphorical expressions (TLMs plus CMs) in Tables 8 and Table 9.

In terms of F1, the best performance for both test datasets (MEWS\_Teacher and MEWS\_Ads) and for both TLMs and TLMs+CMs was seen for the model trained on TOEFL. Across both prompts as well as across TLMs and TLMs+CMs, precision is higher than recall when training on MEWSMET. The results are generally higher for TLMs+CMs than for TLMs only.

#### 4.2.3 Experiment 3: Combining TOEFL with Target Data

As shown in Section 4.2.2, large amounts of training data alone do not lead to better results on MEWSMETS; in-domain training data seems to be necessary.<sup>7</sup> As our dataset is too small for the model to learn, we next use a combination of our data (MEWS\_Teacher) in combination with the larger TOEFL corpus as training data. We are mainly interested in detecting TLMs, so for MEWS\_Teacher we only considered TLMs as metaphors. The results are reported in Table 10. In terms of F1, DeepMet trained on both TOEFL and MEWS\_Teacher achieves the best results of all models for both TLMs and TLMs+CMs.

#### 4.3 Discussion

The results of our experiments yield five main insights. Firstly, large amounts of training data are vital for DeepMet to perform well. As is shown in Table 6, DeepMet performs much better when training and testing on VUA or TOEFL data than on MEWSMET. Here, the training datasets for the VUA and TOEFL experiments are much larger than for MEWSMET experiment (see Figure 4). When reducing the amount of VUA training data to match the size of the MEWSMET corpora, DeepMet fails at the classification task for the VUA test set (F1 = 0.4).

The second insight is that in-domain training data is needed. When we increased the training data by using VUA and TOEFL (see Tables 8 and 9), and tested on MEWSMET, the model trained on TOEFL-data outperformed the model trained on VUA-data. This behaviour was seen across prompts and for both TLMs and TLMs+CMs. This shows that large amounts of training data are needed only to an extent; after a certain threshold (that has to be determined in future work), in-domain data becomes more important than more training data. The importance of in-domain data was also highlighted by the fact that the best performance overall was seen when training on TOEFL and MEWS\_Teacher, and testing on MEWS\_Ads.

Thirdly, it became clear that the results for detecting TLMs+CMs are generally higher than for detecting TLMs only (see Tables 8, 9 and 10). This means that the models are better at detect-

<sup>7</sup>By in-domain we mean language that EFL learners used in argumentative essays for various prompts.

	TLMs only			TLMs + CMs		
	Precision	Recall	F1	Precision	Recall	F1
<b>TOEFL</b>	12.0	80.7	<b>20.8</b>	17.2	80.5	<b>28.4</b>
<b>VUA</b>	8.8	92.7	16.1	12.8	93.3	22.5

Table 8: Performance of DeepMet fine-tuned on TOEFL and VUA, and evaluated on the split of our dataset that is based on the prompt *TV-Ads*.

	TLMs only			TLMs + CMs		
	Precision	Recall	F1	Precision	Recall	F1
<b>TOEFL</b>	14.6	86.4	<b>24.9</b>	23.1	88.1	<b>36.6</b>
<b>VUA</b>	10.5	90.3	18.7	16.5	91.9	28.0

Table 9: Performance of DeepMet fine-tuned on TOEFL and VUA, and evaluated on the split of our dataset that is based on the prompt *Teacher*.

	TLMs only			TLMs + CMs		
	Precision	Recall	F1	Precision	Recall	F1
<b>TOEFL+MEWS_T</b>	18.7	63.7	28.9	27.4	64.8	38.5

Table 10: Performance of DeepMet fine-tuned on TOEFL-data plus MEWS\_Teacher, and evaluated on the split of our dataset that is based on the prompt *TV-Ads*.

ing metaphors that are comprehensible, but that a native speaker would not use. This, however, is problematic, since TLMs can be an indicator of language proficiency, while CMs apparently cannot. If metaphors were to be used as features in automatic essay scoring, an additional module would be needed that extracts TLMs.

Our fourth insight is that the model tends to overidentify metaphors, which can be seen by the high recall and low precision across all experiments that were carried out with a sufficient amount of training data. One explanation for this behaviour is that the percentage of metaphorical expressions in MEWSMET is lower than in VUA and TOEFL training data (MEWSMET: 2% and 3%, VUA: 30%, TOEFL: 4%, see Figure 4). Also, the preference parameter  $\alpha$ , originally designed to improve recall, has to be fine-tuned to MEWSMET data (we used a value of 0.2 as suggested by Su et al., 2020).

Lastly, it should be mentioned that a more reliable metaphor detection method has to be found, as our best model (trained on TOEFL and MEWS\_Teacher, see Table 10) shows a rather weak F1-score of 28.9 for detecting TLMs only.<sup>8</sup>

<sup>8</sup>In addition to the results presented above, we used the

## 5 Error Analysis

In order to get a clearer picture on why DeepMet performs rather poorly on MEWSMET data, we performed an error analysis. For this we used the best-performing model – DeepMet trained on TOEFL-data plus MEWS\_Teacher – and looked at the predictions it made for MEWS\_Ads, taking into account both TLMs and CMs. The first thing we noticed is that many differences between the annotations and the predictions concerned verbs where the concrete meaning is not the basic meaning (anymore). These verbs include *to direct*, *to confront*, *to support*, *to create*, *to target*, or *to manipulate*. For instance, the four example sentences given for the first listed meaning of *to direct* in the Longman Dictionary<sup>9</sup> are as follows:

- (1) The machine directs an X-ray beam at the patient’s body.

metaphor detection model by Ma et al. (2021), because it is in theory able to make reliable predictions with as little as 200 training instances, as has been shown by Hülising and Schulte Im Walde (2024) in a multilingual setup. However, the results we received for our MEWS-data were very poor (F1 < 14.4), which indicates that the model works well for standard language, but not for learner language.

<sup>9</sup><https://www.ldoceonline.com/dictionary/direct>, date of access: 15.08.2024

- (2) The new route directs lorries away from the town centre.
- (3) I'd like to direct your attention to paragraph four.
- (4) I want to direct my efforts more towards my own projects.

As none of these meanings entails sensual perception, the basic meaning is abstract, even though there might be instances where the word is used in a concrete way, e.g. *to direct the fire extinguisher at something*. In our guidelines based on MIP (Pragglejaz Group, 2007) we state that for metaphorical usage the meaning of a word in context “tend(s) to differ from the basic meaning”, and we ask the annotators to compare the meaning in a given context to the basic meaning, i.e. the first meaning mentioned in the Longman dictionary (see guidelines in Appendix A.1.1). Therefore, the meaning of *to direct* in a context such as *advertising directed toward young children*<sup>10</sup> does not “differ from the basic meaning” and is labelled as being literal, even though our model labels it as being metaphorical. This might be due to the fact that the majority of data used for fine-tuning stems from the TOEFL-data where the annotation is not based on MIP (Pragglejaz Group, 2007), but rather based on the annotators’ intuitions (Beigman Klebanov et al., 2018). The following sentences in the dataset by (Beigman Klebanov et al., 2018) contain the verb *to direct*, and two out of three labels are metaphorical<sup>11</sup>:

- (5) At a first sight, it can be inferred that young people [...] seem to have become more ego-directed, in order to prevent themselves from the duties that a society is asking them. → literal
- (6) it is the nature of the humen, but this in-trest need to be directed in the right way but unfortunetlly the same can be directed by some people whom not civilized.  
→ metaphorical (both)

This indicates that the different guidelines account for differences in classification.

A second source of differences between anno-

<sup>10</sup>It should be noted that the word *directed* is used in the prompt “TV-Ads” and should therefore be excluded when analyzing the correlation between proficiency level and the number of metaphors per essay.

<sup>11</sup>Only the two instances labelled as metaphorical are true verbs, the other one being a deverbal adjective.

tations and predictions are personifications. In line with conceptual metaphor theory (Lakoff and Johnson, 1980b), we explicitly consider personifications as metaphors (cf. Appendix A.1.2). Therefore, all of the following expressions in our MEWS data were annotated as being used metaphorically:

- (7) [...] parents think that advertise threatens their child [...]
- (8) If a advertise is made well it teaches the child something [...]
- (9) [...] I saw an advertisement, which was directly telling children that they should go to a certain water park [...]

However, the model predicted them not to be metaphors, which is probably again due to the different annotation guidelines used for the training and the testing data. As the guidelines by Beigman Klebanov et al. (2018) are based on intuition, personifications are not specifically mentioned, so it can be assumed that the annotators did not consider them metaphors. The fact that the verb *entertains* was labelled as being used literally in following sentence from the TOEFL-data confirms this assumption:

- (10) [...] the computer graphic which entertains many people in films or TVs can not invented without computer.

Thirdly, highly conventionalized expressions, such as *to raise a question* or *to come to the conclusion* were annotated as being used metaphorically and predicted as being used literally. Even though neither of these expression could be found in the training data by Beigman Klebanov et al. (2018), the following sentence was found where the word *to raise* is used similarly:

- (11) And even though their usage has raised certain environmental concerns [...]

Again, *raised* is not annotated as being used metaphorically in the TOEFL-data, probably because it is too conventionalized and did not “help the author advance her argument” (Beigman Klebanov et al., 2018).

These three reasons for misclassifications hint at the need for training data that was created with the help of comparable annotation guidelines.

## 6 Conclusion

In our study we set out to investigate the relationship between metaphors and essay scores. We found that EFL learners create new conceptual mappings, which are perfectly comprehensible for native speakers in spite of being uncommon (comprehensible metaphors, CMs). However, this strategy – which is absolutely serviceable in everyday life – does not give us any insights into the proficiency level of a learner, as our results suggest. Rather, language proficiency seems to correlate only with the use of metaphors that a native speaker would use (target language metaphors, TLMs).

If we want to use the number of metaphors in an essay as a feature for automatic essay scoring, we need to detect metaphors automatically. Previous studies have shown that metaphor detection methods such as DeepMet (Su et al., 2020) perform well on EFL learner data by Beigman Klebanov et al. (2018). However, such methods had not been extensively validated for younger and less proficient learners as present in our data. We showed that large amounts of training data are necessary to train a model that learns to detect metaphors in MEWSMET, however, standard English data is not useful, but new in-domain data is needed to achieve decent model performance. Here, training and testing data should ideally be annotated under the same annotation guidelines, as our error analysis revealed.

We also showed that DeepMet tends to be better at classifying CLs than TLMs. This poses a challenge, since only the number of TLMs per essay positively correlates with language proficiency. What is needed, therefore, is a method that reliably differentiates between TLMs and CMs, if we want to use the number of metaphors as features for essay scoring.

## 7 Outlook

Our MEWSMET dataset allows further analyses: First of all, differentiating between TLMs and CMs is vital. Pedinotti et al. (2021) use a dataset consisting of conventional metaphors and creative metaphors. They matched each of these metaphors with their literal counterpart and a nonsensical expression of the same syntactic structure. They used the pseudo-log-likelihood score (PLL) by Wang and Cho (2019) to measure the degree of plausibility that BERT attributes to a sen-

tence. In doing so, they show that BERT is able to discern creative metaphors from nonsense expressions. As future work, we will apply this score to see whether it can also discern TLMs from CMs.

What has not been taken into account yet is the degree of conventionalization. Although our annotators assigned the labels “creative” and “conventional” to all metaphorical instances that they believed to be acceptable English (for all others they assigned the label “uncommon translation of a German conventionalized metaphor” or “non-metaphorical”), these labels should be confirmed by native speakers, or checked against a corpus-based dictionary as is commonly done to detect creative metaphors (Reimann and Scheffler, 2024b)<sup>12</sup>. Scores indicating novelty could weigh the metaphorical labels; after all, metaphors that the learner has frequently heard or even learnt from a textbook should be treated differently than creative metaphors that learners form themselves, when looking at the correlation with essay scores.

Also, the proximity to German metaphors should be taken into account when using metaphors as features for essay scoring. In this study, we annotated metaphors that are uncommon because they are translated from a German conventional metaphor (e.g. *to build an opinion*). We did not carry out further analyses on these uncommon translations due to their small number; only 29 expressions were annotated as being uncommon by both annotators, and two were considered incomprehensible during the check by the native speakers. However, there are probably many metaphors that originate from a parallel between the source and the target language, some that are incomprehensible but certainly others which are CMs or even TLMs. One example is *das bringt uns zum nächsten Punkt*, which can be translated word for word into *this brings us to the next point*. A learner’s proficiency can be more clearly predicted when they use metaphors that do not run in parallel to German metaphors, for example, when *eine Meinung bilden* is translated into *to form an opinion* instead of *to build an opinion*.

## 8 Acknowledgements

We thank Stefan Keller and Flavio Lötscher (PH Zürich) for their help in refining the annotation guidelines. We are also grateful for the annotations provided by Rachel Castleberg, Rosalind

<sup>12</sup>We use the terms novel and creative interchangeably.

Isaacs, Jordana Plagge, and Charlotte Ventura. Anna Hülsing is supported by the German Federal Ministry of Education and Research (grant no. FKZ 01JA23S03C). Andrea Horbach’s work was partially conducted at “CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics” of the FernUniversität in Hagen, Germany.

## 9 Ethics Statement

Our annotators were paid according to German minimum wage regulations.

## References

- Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. [Metaphors in pre-trained language models: Probing and generalization across datasets and languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.
- Daniel Baleato Rodríguez, Verna Dankers, Preslav Nakov, and Ekaterina Shutova. 2023. [Paper bullets: Modeling propaganda with the help of metaphor](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 472–489, Dubrovnik, Croatia. Association for Computational Linguistics.
- Beata Beigman Klebanov and Michael Flor. 2013. [Argumentation-relevant metaphors in test-taker essays](#). In *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20, Atlanta, Georgia. Association for Computational Linguistics.
- Beata Beigman Klebanov, Chee Wee (Ben) Leong, and Michael Flor. 2018. [A corpus of non-native written English annotated for metaphor](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 86–91, New Orleans, Louisiana. Association for Computational Linguistics.
- Steven Bird and Edward Loper. 2004. [NLTK: The natural language toolkit](#). In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.
- Julia Birke and Anoop Sarkar. 2006. [A clustering approach for nearly unsupervised recognition of non-literal language](#). In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336, Trento, Italy. Association for Computational Linguistics.
- Daniel Blanchard, Joel R. Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. [TOEFL11: A corpus of non-native English](#). *ETS Research Report Series*, 2013:15.
- Lynne Cameron. 2003. [Metaphor in educational discourse](#). *Advances in Applied Linguistics*. Continuum, London, UK.
- Minjin Choi, Sunkyoung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. [MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.
- Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. [Modelling the interplay of metaphor and emotion through multitask learning](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229, Hong Kong, China. Association for Computational Linguistics.
- Erik-Lân Do Dinh and Iryna Gurevych. 2016. [Token-level metaphor detection using neural networks](#). In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33, San Diego, California. Association for Computational Linguistics.
- Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. [Weeding out conventionalized metaphors: A corpus of novel metaphor annotations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.
- Markus Egg and Valia Kordoni. 2022. [Metaphor annotation for German](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2556–2562, Marseille, France. European Language Resources Association.
- Arthur Graesser and Danielle McNamara. 2012. [Automated analysis of essays and open-ended verbal responses](#). In H. Cooper, editor, *APA handbook of research methods in psychology*, volume 1, pages 307–325. American Psychological Association.
- Pragglejaz Group. 2007. [MIP: A method for identifying metaphorically used words in discourse](#). *Metaphor and Symbol*, 22(1):1–39.
- E. Darío Gutiérrez, Guillermo Cecchi, Cheryl Corcoran, and Philip Corlett. 2017. [Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2923–2930,

- Copenhagen, Denmark. Association for Computational Linguistics.
- Anna Hülsing and Sabine Schulte Im Walde. 2024. [Cross-lingual metaphor detection for low-resource languages](#). In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 22–34, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.
- Stefan Keller. 2016. Measuring Writing at Secondary Level (MEWS). Eine binationale Studie. *Babylonia*.
- Stefan D. Keller, Johanna Fleckenstein, Maleika Krüger, Olaf Köller, and André A. Rupp. 2020. [English writing skills of students in upper secondary education: Results from an empirical study in switzerland and germany](#). *Journal of Second Language Writing*, 48:100700.
- Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. [The INCEption platform: Machine-assisted and knowledge-oriented interactive annotation](#). In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Veranstaltungstitel: The 27th International Conference on Computational Linguistics (COLING 2018).
- Maximilian Köper and Sabine Schulte im Walde. 2016. [Distinguishing literal and non-literal usage of German particle verbs](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362, San Diego, California. Association for Computational Linguistics.
- Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. [Multilingual multi-figurative language detection](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267, Toronto, Canada. Association for Computational Linguistics.
- George Lakoff. 1987. *The death of dead metaphor*, volume 2 of *Metaphor and Symbolic Activity*. De Gruyter Mouton, Berlin, Boston. 2010.
- George Lakoff and Mark Johnson. 1980a. [Conceptual metaphor in everyday language](#). *Journal of Philosophy*, 77(8):453–486.
- George Lakoff and Mark Johnson. 1980b. *Metaphors we Live by*. University of Chicago Press, Chicago.
- Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xinyang Chen. 2020. [A report on the 2020 VUA and TOEFL metaphor detection shared task](#). In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.
- Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. [A report on the 2018 VUA metaphor detection shared task](#). In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana. Association for Computational Linguistics.
- Yucheng Li, Frank Guerin, and Chenghua Lin. 2022. [The secret of metaphor on expressing stronger emotion](#). In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 39–43, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023. [Metaphor detection via explicit basic meanings modelling](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada. Association for Computational Linguistics.
- Weicheng Ma, Ruibo Liu, Lili Wang, and Soroush Vosoughi. 2021. [Improvements and extensions on metaphor detection](#). In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 33–42, Online. Association for Computational Linguistics.
- George A. Miller. 1994. [WordNet: A lexical database for English](#). In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.
- Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. [Metaphor as a medium for emotion: An empirical study](#). In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.
- Ana Mouraz, Ana Vale, and Raquel Rodrigues. 2013. [The use of metaphors in the processes of teaching and learning in higher education](#). *International Online Journal of Educational Sciences*, 5:99–110.
- Kai Niebert and Harald Gropengiesser. 2012. [Understanding and communicating climate change in metaphors](#). *Environmental Education Research - ENVIRON EDUC RES*, 19:1–21.
- Rebecca Oxford, Stephen Tomlinson, Ana Barcelos, Cassandra Harrington, Roberta Lavine, Amany Saleh, and Ana Longhini. 1998. [Clashing metaphors about classroom teachers: Toward a systematic typology for the language teaching field](#). *System*, 26:3–50.
- Natalie Parde and Rodney Nielsen. 2018. [A corpus of metaphor novelty scores for syntactically-related word pairs](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).



- Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. 2021. [A howling success or a working sea? Testing what BERT knows about metaphors](#). In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–204, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Prisca Piccirilli and Sabine Schulte im Walde. 2022. [Features of perceived metaphoricity on the discourse level: Abstractness and emotionality](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5261–5273, Marseille, France. European Language Resources Association.
- Gudrun Reijniere, Christian Burgers, Tina Krennmayr, and Gerard Steen. 2018. [DMIP: A method for identifying potentially deliberate metaphor in language use](#). *Corpus Pragmatics*, 2:129–147.
- Sebastian Reimann and Tatjana Scheffler. 2024a. [Metaphors in online religious communication: A detailed dataset and cross-genre metaphor detection](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11236–11246, Torino, Italia. ELRA and ICCL.
- Sebastian Reimann and Tatjana Scheffler. 2024b. [When is a metaphor actually novel? annotating metaphor novelty in the context of automatic metaphor detection](#). In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 87–97, St. Julians, Malta. Association for Computational Linguistics.
- Eva Samaniego Fernández, María Sol Velasco Sacristán, and Pedro Antonio Fuertes Olivera. 2005. *Translations we live by: The impact of metaphor translation on target systems*, pages 61–82. Secretariado de Publicaciones, Valladolid.
- Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. [Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection](#). In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Jakob Schuster and Katja Markert. 2023. [Nutcracking sledgehammers: Prioritizing target language data over bigger language models for cross-lingual metaphor detection](#). In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 98–106, Gothenburg, Sweden. Association for Computational Linguistics.
- Ekaterina Shutova and Simone Teufel. 2010. [Metaphor corpus annotated for source - target domain mappings](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, Tina Krennmayr, Tryntje Pasma, et al. 2010. *A method for linguistic metaphor identification*. John Benjamins Publishing Company Amsterdam.
- Egon Stemle and Alexander Onysko. 2018. [Using language learner data for metaphor detection](#). In *Proceedings of the Workshop on Figurative Language Processing*, pages 133–138, New Orleans, Louisiana. Association for Computational Linguistics.
- Chuangdong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. [DeepMet: A reading comprehension paradigm for token-level metaphor detection](#). In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online. Association for Computational Linguistics.
- Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. [Metaphor detection with cross-lingual model transfer](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.
- Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. [Literal and metaphorical sense identification through concrete and abstract context](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Oseremen Uduehi and Razvan Bunescu. 2024. [An expectation-realization model for metaphor detection](#). In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 79–84, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.
- Alex Wang and Kyunghyun Cho. 2019. [BERT has a mouth, and it must speak: BERT as a Markov random field language model](#). In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.
- Rodrigo Wilkens, Daiane Seibert, Xiaou Wang, and Thomas François. 2022. [MWE for essay scoring English as a foreign language](#). In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with READING Difficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 62–69, Marseille, France. European Language Resources Association.

Yiqin Yang, Li Xia, and Qianchuan Zhao. 2019. *An automated grader for chinese essay combining shallow and deep semantic attributes*. *IEEE Access*, 7:176306–176316.

Dongyu Zhang, Nan Shi, Ciyuan Peng, Abdul Aziz, Wenhong Zhao, and Feng Xia. 2021. *MAM: A metaphor-based approach for mental illness detection*. In *Computational Science – ICCS 2021: 21st International Conference, Krakow, Poland, June 16–18, 2021, Proceedings, Part III*, page 570–583, Berlin, Heidelberg. Springer-Verlag.

## A Appendix

### A.1 Annotation Guideline

#### A.1.1 Main Guideline

Look at each essay individually. For each essay perform the following steps:

1. Read each sentence and pay attention to the target verbs, which are already tagged.
2. The label “Metaphorical Usage” should be given to a target verb if you believe that this word is used metaphorically. Add the label “Metaphorical Usage” where missing. The following label descriptions (taken from [Mohammad et al., 2016](#)) should help you:

- Literal usages tend to be more basic, and have a more straightforward meaning; they are more physical and more closely tied to our senses (vision, hearing, touching, tasting).

**Example 1:** The enemy shot down our aircraft.

→ non-metaphorical verb usage, no labelling necessary

- Metaphorical usages tend to differ from the basic meaning and tend to be more complex and more distant from our senses. They often are more abstract, vague, and surprising. Also, they tend to bring in imagery from a different domain.

**Example 2:** He shot down the student’s proposal.

→ label: “metaphorical verb usage”

At the end of step 2, all metaphorically used verbs should have two labels (“Target Verb” and “Metaphorical Usage”).

3. Assign one of the following labels to each target verb that you labelled as being metaphorical:

- Label “Conventionalized Metaphor”: If, in your opinion, the verb represents a conventionalized metaphor, you recognize it to often be used together with one or more of the given context words.

**Example:** Susan often spends her time at the swimming pool.

→ The word *spend* is often used together with the word *time*.

- Label “Creative Metaphor”: If, in your opinion, the verb represents a creative metaphor, you do not recognize the verb being usually used together with one or more of the given context words.

**Example:** The present sews together the past and the future.

→ The word *sew* is usually not used together with words such as *present* or *past*.

- Label “Uncommon Translation Conventionalized”: If the verb represents an uncommon translation of a German conventionalized metaphor, you recognize a German conventionalized metaphor as the basis for the translation, but you think that the English translation is not common.

**Example:** *eine Meinung bilden*, student translation: *to build an opinion*.

NB: This label should only be given if you believe that the underlying German expression contains a conventionalized metaphor **and** if the resulting English phrase is uncommon or unidiomatic. It should not be given if the English phrase is unidiomatic or uncommon, but no German metaphor is the source for the error. This label should only be given in clear cases such as the afore mentioned phrase *to build an opinion*.

At the end of step 3, all metaphorically used verbs should have three labels (“Target Verb”, “Metaphorical Usage” and one of the following labels: “Conventionalized Metaphor”, “Creative Metaphor”, “Uncommon Translation Conventionalized”).

#### A.1.2 Additional Notes

As we are dealing with authentic, and therefore noisy text, there will be expressions where the metaphoricity of a verb is unclear. In order to

clarify which words should be tagged as being metaphorical and which should not, the following examples are given as anchors for the annotation.

- Words such as *to direct* or *to confront* should not be tagged as being metaphorical. These words can have a straightforward, more physical meaning (for example *to direct the extinguisher at the fire*), but this is currently not the basic meaning, as these words are in the vast majority of occurrences used in an abstract way. Therefore, here the abstract meaning is the basic meaning.
- Very frequent verbs such as *have/be/do/make...* have not been tagged as “Target Verbs” in the annotation documents, because they are mostly used as auxiliary verbs. In cases where these words occur as full verbs (e.g. *have a conversation*), the metaphorical meaning is determined mainly by the following noun, while the verb carries no or little meaning (cf. light verb phrases). As we are establishing the metaphoricity of the verbs, it is fair to say that these verbs carry no metaphorical meaning, and are therefore excluded.
- Target verbs in expressions such as *to spend time* or *to cover topics* should be tagged as being metaphorical. The expressions are highly conventionalized, but – as opposed to light verb phrases such as *have a conversation* – the meaning of the expression does not only rest on the noun, and therefore the verb carries some of the metaphorical weight.
- Idioms can be metaphors, too. For example, the verb *break* is used metaphorically in the expression *to break the ice* and should be tagged as being a metaphor. However, there are many idioms which do not have a metaphorical origin (*break a leg*, *talk to Huey on the big white telephone*) or where the origin is unclear (*it’s raining cats and dogs*). These should not obtain the label “Metaphorical Usage”.
- Phrasal verbs should be tagged as being metaphorical only if the basic meaning of the entire phrasal verb usually is more straightforward/physical/... (see example above: *to shoot down*). They should not be tagged as

being metaphorical if only the base verb usually is more straightforward/physical/.... Example: *to miss out* should not be tagged as being metaphorical, even though the basic meaning of the base verb (*to miss*) might be more straightforward/physical/....

- Personifications should be annotated as metaphors, too. Example: *Money rules the world*.
- If a verb is used as part of an extended metaphor, it should be tagged as being used metaphorically. Example: *His head was a dovecote, most thoughts flew out, only some stayed inside*. Here, the target words should be marked as being used metaphorically.
- If you are unsure what the basic meaning of a verb is, consult the online version of the Longman Dictionary: <https://www.ldoceonline.com/dictionary/>. Be aware that there are homonyms, so there might be more than one basic meaning of a verb (for example: *to lie* can refer to the position of a person or to a person not telling the truth).
- Dead metaphors, i.e. metaphors that do not exist anymore because the mapping from source to target domain can no longer be understood without historical knowledge (compare Lakoff, 1987), should not be tagged as being metaphorical. Examples: *footage*, *pedigree*.

## B Stop Words

In addition to the commonly used stop words (*be, do, should, can, have, would*) we also excluded the word *make*, because it is very often used by students as a placeholder for a verb they do not know, for example: *because school makes our future* or *make good grades*.

# Investigating strategies for lexical complexity prediction in a multilingual setting using generative language models and supervised approaches

Abdelhak Kelious<sup>1</sup>, Mathieu Constant<sup>1</sup>, Christophe Coeur<sup>2</sup>

<sup>1</sup>University of Lorraine and CNRS/ATILF, <sup>2</sup>Consultant

abdelhak.kelious@univ-lorraine.fr

mathieu.constant@univ-lorraine.fr

christophe.coeur@gmail.com

## Abstract

This paper explores methods to automatically predict lexical complexity in a multilingual setting using advanced natural language processing models. More precisely, it investigates the use of transfer learning and data augmentation techniques in the context of supervised learning, showing the great interest of multilingual approaches. We also assess the potential of generative large language models for predicting lexical complexity. Through different prompting strategies (zero-shot, one-shot, and chain-of-thought prompts), we analyze model performance in diverse languages. Our findings reveal that while generative models achieve promising performances, their predictive quality varies and optimized task-specific models still outperform them when they benefit from sufficient training data.

## 1 Introduction

Lexical complexity prediction consists in assessing the difficulty of a target word in a given context, either as a binary classification (is the word difficult or not?) or as a continuous numerical value prediction indicating the degree of complexity. Such a task is potentially useful for computer-assisted language learning: e.g. for selecting relevant textual materials for learners or for identifying complex words in texts and then providing enriched information to help the reader’s understanding.

Our study explores deep learning methodologies for multilingual lexical complexity prediction (LCP). We leverage recent advances in natural language processing models, such as transformers and generative models, to assess lexical complexity across various languages. More precisely, we first investigate various multilingual methods like transfer learning and data augmentation using

a supervised approach. We then explore the capabilities of generative pre-trained large language models (LLMs) to perform LCP applying various prompt engineering and ensemble techniques. The experiments are carried out on multilingual datasets from two shared tasks: the 2018 Complex Word Identification task (Yimam et al., 2018a) for English, French, German and Spanish, and the Multilingual Lexical Simplification Pipeline (MLSP) shared task (Shardlow et al., 2024a) for a subset of languages (English, French, Japanese and Spanish).

## 2 Related work

Lexical complexity prediction has been a growing area of research, with several works contributing to the development of graded lexical resources and methodologies aimed at understanding word complexity from both native and non-native language learners’ perspectives. For example, Gala et al. (2013) laid the groundwork for French lexical complexity by proposing a lexicon with difficulty measures. Building on this, François et al. (2014) introduced FLELex, a graded lexical resource specifically designed for French foreign learners. Tack et al. (2018) extended this research to Dutch with NT2Lex, a graded lexical resource linked to the Dutch WordNet. Meanwhile, Alfter and Volodina (2018) focused on predicting single-word lexical complexity, a task later expanded by Alfter (2021) to include multi-word expressions, highlighting the evolving nature of complexity prediction tasks. For more details on this task, North et al. (2023) provided a comprehensive overview of the computational approaches used.

### 2.1 Shared tasks

Lexical complexity prediction has also been the focus of multiple shared tasks over the last decade that strongly contributed to the advances of the field through the development of new dedicated

---

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

datasets as well as novel technical methods to perform the task.

The 2016 Complex Word Identification (CWI) task at SemEval highlighted key findings in identifying complex words, especially for non-native English speakers. The dataset, dedicated to the English language, created with input from 400 non-native speakers, showed that complex words are generally rarer, less ambiguous, and shorter. Decision trees, ensemble methods, and particularly word frequency were found to be reliable predictors of word complexity (Paetzold and Specia, 2016). Top systems, such as those by UWB and LTG, utilized features like document frequency and contextual language models, achieving high rankings (Konkol, 2016; Malmasi et al., 2016). Despite various feature explorations and innovative methods like sequence labeling (Gooding and Kochmar, 2019), the fundamental effectiveness of word frequency remained central to CWI success (Zampieri et al., 2017).

The 2018 Complex Word Identification task, thereafter CWI 2018, part of the BEA workshop at NAACL 2018, focused on identifying difficult words in texts across multiple languages, including English, German, Spanish, and French. The task was divided into binary and probabilistic classification tracks, attracting 12 teams with various approaches. Notably, ensemble-based methods and feature engineering demonstrated strong performance (Yimam et al., 2018a). Systems such as those by the NLP-CIC team compared deep learning with feature engineering, showing comparable results (Aroyehun et al., 2018). Simple models based on character n-grams also performed competitively, sometimes matching more complex systems (Alfter and Pilán, 2018). The challenge highlighted the effectiveness of both traditional feature engineering and modern deep learning approaches in CWI.

The 2021 Lexical Complexity Prediction (LCP 2021) task (Shardlow et al., 2021) at SemEval involved predicting, for the English language, the complexity of single words and multi-word expressions in context using a five-point Likert scale. The competition attracted 198 teams, with top-performing systems leveraging advanced NLP techniques such as transformers and ensemble methods. The winning system used fine-tuned pre-trained language models with stacking mechanisms, achieving high Pearson correlation scores

(Pan et al., 2021). Approaches varied widely, from logistic regression with linguistic features (De-sai et al., 2021) to ensemble-based models combining different feature types (Vettigli and Sorgente, 2021). The task highlighted the effectiveness of combining traditional linguistic features with modern deep learning models to predict lexical complexity accurately.

Recently, a dataset was developed for the MLSP 2024 shared task (Shardlow et al., 2024a). It includes 5,624 instances across 10 target languages. Each instance features a sentence from an educational text with a specific target word highlighted. For each target word, there are two types of annotations: an aggregate complexity score (rated on a scale from 1 to 5 by 10 annotators) indicating the difficulty level of the word, and a list of possible substitutions that simplify the sentence while preserving its original meaning.

## 2.2 Multilingual approaches

Although many studies concentrate on English due to a relative shortage of resources in other languages, promising approaches such as transfer learning and data augmentation have been proposed to address this gap. Cross-lingual transfer learning significantly enhances Complex Word Identification (CWI) by leveraging models trained in high-resource languages for use in low-resource languages. Zaharia et al. (2020) demonstrated the effectiveness of zero-shot, one-shot, and few-shot learning techniques with state-of-the-art NLP models, achieving high F1-scores across multiple languages. Bingel and Bjerva (2018) used cross-lingual multitask learning, showing that language-agnostic models could generalize well across different languages. Additionally, Yimam et al. (2017) employed language-independent features to train multilingual and cross-lingual models, achieving comparable performance to monolingual systems.

## 2.3 Large language models' capabilities

Large Language Models (LLMs) like ChatGPT, Mistral, and Llama3 have significantly advanced natural language processing across various domains. Given that we are currently in the era of LLMs, it is crucial to compare and assess their role in our study to understand their impact on various tasks. They excel in industrial engineering tasks, such as automation and programming, though they have limitations with complex physics

equations (Ogundare et al., 2023). In mathematical problem-solving, LLMs effectively handle arithmetic tasks using chain-of-thought reasoning (Yuan et al., 2023). Their ability to use multimodal tools is enhanced by frameworks like GPT4Tools, which improve performance in visual tasks (Yang et al., 2023). Instruction-following datasets and fine-tuning, as seen with FLACUNA, enhance their problem-solving skills (Ghosal et al., 2023). Comprehensive evaluations reveal strengths in diverse tasks like question-answering and code generation, although challenges remain (Laskar et al., 2023). Techniques like role-play prompting further improve their reasoning capabilities, making LLMs versatile tools for a wide range of applications (Kong et al., 2023). The ANU team, participating in the MLSP 2024 task to predict word complexity based on context, relied on a prompting strategy with GPT-3.5 (i.e. GPT-3.5-turbo-instruct) for the tasks using zero, one, and few-shot strategies. The zero-shot strategy included the context and target word while the non-zero strategies relied on instructing the model with one or three random samples from the trial data according to the prompting template. Overall, the authors indicate under-performance for the LCP task, while demonstrating strong performance for English in lexical simplification (Seneviratne and Suominen, 2024).

### 3 Multilingual lexical complexity prediction based on supervised learning

In this section, we investigate two main strategies for the task of lexical complexity prediction (LCP) in multiple languages using a supervised approach:

1. **Monolingual training:** the model is trained on a dataset in the target language; the training data may be composed of native data in the target language, data translated to the target language from a resource-richer language (English in our case), or a combination of both where the native data is augmented with translated data;
2. **Multilingual training:** the model is trained on a multilingual dataset including or not data in the target language; the model is based on multilingual word embeddings to deal with transfer learning.

The actual implementation of these approaches will depend on the dataset on which they will be experimented, given their different nature and composition (cf. section 3.1 and section 3.3).

#### 3.1 Datasets

Experiments to evaluate these strategies are performed on two multilingual datasets: CWI 2018 (Yimam et al., 2018b) and MLSP 2024 (Shardlow et al., 2024a), cf. section 2. The CWI 2018 dataset provided by (Yimam et al., 2018b) includes data in English, Spanish, and German for training and testing, and French solely for testing purposes, cf. table 1. Our focus is on Spanish, German, and French. We selected this dataset because it offers large possibilities of multilingual experiments using supervised learning. Two types of labels are available: binary and probabilistic. Our evaluation is conducted using the binary labels.

Language	Train	Dev	Test
English	27,299	3,328	4,252
German	6,151	795	959
Spanish	13,750	1,622	2,233
French	-	-	2,251

Table 1: The number of instances for each training, development and test set (Yimam et al., 2018b)

Additionally, we performed evaluation on the MLSP 2024 dataset (Shardlow et al., 2024a), which includes 5,624 instances across 10 target languages. The MLSP dataset provides probabilistic labels, where annotations are continuous values between 0 and 1. This dataset contains only testing and development data, the latter being limited to around 30 instances per language, i.e. 300 instances in total. We only focus on four languages (French, English, Japanese, and Spanish) in order to limit the energetic impact of our experiments and to focus on the languages studied in our working environment. Due to the lack of training data, we have decided to leverage the LCP 2021 dataset (Shardlow et al., 2021), which provides annotations highly similar to those in the MLSP task, for the English language.

#### 3.2 The model

In our research, we adopt a recent system that has proven effective in predicting lexical complexity for English (Keliou et al., 2024). We replicate this model in a multilingual version. The model

combines a pre-trained language model with frequency characteristics based on Zipf’s law. Such a system is in line with the literature showing that hybrid models using transformers (encoders) enhanced with additional linguistic features deliver more robust and effective results (Wilkins et al., 2024).

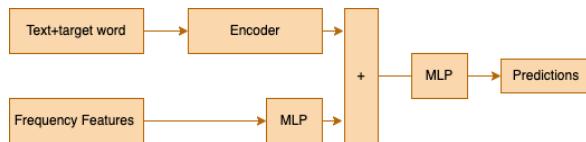


Figure 1: The overall architecture for predicting lexical complexity (Kelious et al., 2024).

Figure 1 illustrates the model described with more details in Kelious et al. (2024). This model is divided into two main parts. The first part relies on lexical embeddings: the encoder receives the target word and its context as input, formatted as follows:  $[CLS]Context[SEP]Target\ Word$ , where  $[CLS]$  and  $[SEP]$  are special tokens used in the Transformers model for processing texts. The second part incorporates five characteristics based on Zipf’s frequency, processed by a multilayer perceptron (MLP). The whole, i.e. the concatenation of the two parts, is then processed by an additional MLP layer. The model’s output is a continuous value between 0 and 1. To classify this output into binary classes, we add a sigmoid layer and apply a decision threshold set at 0.5 to convert the probabilities into binary classes for the experiments on CWI 2018.

The conversion of this model from monolingual to multilingual is relatively straightforward: for the frequency features, it suffices to extract frequency data in the target language from available corpora. As for the transformer (encoder) part, it is necessary to implement a multilingual model or a monolingual model suited to the specific language we wish to evaluate.

### 3.3 Experimental settings

The LCP model is based on various language models for encoding the input context. For multilingual training strategies, we selected the multilingual language model mdeberta-v3-base<sup>1</sup>. For monolingual training strategies, we selected Spanish BERT for Spanish (Cañete et al., 2020), German BERT for German (Chan et al., 2020), De-

<sup>1</sup><https://huggingface.co/microsoft/mdeberta-v3-base>

Berta (He et al., 2023) for English and mdeberta-v3-base for Japanese. The Zipf frequencies were computed using the python library Wordfreq<sup>2</sup>. For translating data from English to target languages such as French, German, Japanese and Spanish, we used the M2M100 model (Fan et al., 2021).

### 3.4 Experiments on CWI 2018

This section presents and evaluates the multilingual and monolingual training strategies developed on the CWI 2018 dataset using a supervised approach.

#### 3.4.1 Evaluated methods

For the multilingual training approaches, the experiments were the following:

- **Multilingual (en, de, es):** the LCP model is trained on the training data of all languages having training data, namely English (en), German (de) and Spanish (es);
- **Multilingual (zero shot):** the model is trained on the training data of all languages having training data except the target language, resulting in a zero-shot scenario.

We also experimented the following monolingual training approaches:

- **Monolingual (native data):** the LCP model is trained on the native train dataset of the target language;
- **Monolingual (native + translated data):** the model is trained on the native train dataset of the target language, augmented with a portion of the English training dataset translated to the target language;
- **Monolingual (translated data):** the model is trained on a portion of the English training dataset translated to the target language.

The experiments Monolingual (native data) and Monolingual (native + translated data) were not performed for French as it has no training data. The experiment Monolingual (translated data) was only performed for French.

<sup>2</sup><https://pypi.org/project/wordfreq/>

### 3.4.2 Results

Tables 2 and 3 shows F1 scores for Spanish, German, and French. For the sake of comparison, we also provide the results of the CWI 2018 official baseline, of the best systems of the shared task, and of a random baseline randomly selecting the output from  $\{0,1\}$ . We can derive several insights and make observations regarding the performance and trends across different types of training strategies:

**Multilingual learning.** Generally, multilingual models trained on all languages (but French) have strong performance across all languages (Spanish : 0.800, German : 0.7911, French : 0.799). The zero-shot configuration, which involves using a model in scenarios where it hasn't been explicitly trained on the target language's data, performed reasonably well but not as well as multilingual models trained on all languages (Spanish: 0.746, German: 0.744), cf. Table 2. The high score for French 0.799 in Table 3 indicates that the model benefits significantly from being part of a multilingual setup where the knowledge from other languages can be effectively transferred to French even without direct training. It suggests that the underlying representations learned by the model are robust and applicable across languages.

Model	Spanish	German
Multilingual (es, en, de)	0.800	0.791
Multilingual (zero shot)	0.746	0.744
Monolingual (native data)	0.775	0.761
Monolingual (native data + 4k translated instances))	0.789	0.781
The highest score in (Yimam et al., 2018b)	0.769	0.745
Baseline, from (Yimam et al., 2018b)	0.723	0.754
Random	0.43	0.44

Table 2: F1 Scores for Spanish and German Language Models

Model	F1 Score
Multilingual (zero shot)	0.799
Monolingual (translated data - 2k)	0.770
Monolingual (translated data - 4k)	0.713
Monolingual (translated data - 10k)	0.751
Monolingual (translated data - 27k)	0.717
The highest score in (Yimam et al., 2018b)	0.759
Baseline, from (Yimam et al., 2018b)	0.634
Random	0.38

Table 3: F1 Scores for French

**Monolingual learning.** Focused training on a

single language shows competitive results but still lags slightly behind the multilingual approach: Spanish: 0.775, German: 0.761, cf. Table 2. Augmenting the data with translations from the English data tends to be useful, as shown in Table 2, especially with an augmentation of 4k training instances translated from English to the target language. Other tested sizes tend to reach lower performance.

Regarding French, the LCP model does not use native training data but instead relies on data created by translating the English training dataset to French. This method shows varying performances as the data size increases (F1 scores: 0.770 with 2k instances, 0.713 with 4k, 0.751 with 10k, 0.717 with 27k, the full training set). The fluctuating performance with different dataset sizes indicates that the quality and consistency of translated data might vary significantly, impacting the model's learning and performance. Simply increasing the dataset size does not consistently improve performance. This approach highlights the challenges and limitations of relying on translated data for training language models, where nuances and context-specific elements of the original language might be lost or misrepresented in translation.

**Baseline and Random.** The baseline and random models provide a clear floor for performance, with baselines substantially outperforming random guessing across all languages (Baseline vs. Random: Spanish 0.7237 vs. 0.43, German 0.754 vs. 0.44, French 0.634 vs. 0.38). This reflects the effectiveness of even basic modeling techniques over uninformed strategies.

The analysis highlights that while multilingual training on all languages offers robustness and generalization across languages, targeted strategies such as monolingual training still hold importance, especially when resources are limited. The fluctuation in performance with different data sizes and types of augmentation indicates the need for careful data management and model tuning specific to each language's characteristics.

### 3.5 Experiments on MLSP 2024

In this section, we present the multilingual and monolingual experiments developed for the MLSP 2024 dataset using a supervised approach.



	English			French			Spanish			Japanese		
	Pearson	Spearman	R2	Pearson	Spearman	R2	Pearson	Spearman	R2	Pearson	Spearman	R2
Multilingual (LCP 2021)	0.80	0.76	0.64	0.52	0.49	0.27	0.67	0.64	0.46	0.63	0.65	0.40
Multilingual (LCP 2021+ Dev)	0.83	0.78	0.69	0.56	0.52	<b>0.31</b>	0.67	0.63	0.45	0.66	0.67	<b>0.43</b>
Multilingual (LCP 2021 + Dev + 2k translated data)	/	/	/	0.49	0.47	0.24	0.65	0.64	0.43	0.61	0.61	0.37
Multilingual (LCP 2021 + Dev + 4k translated data)	/	/	/	0.51	0.48	0.26	0.63	0.58	0.39	0.63	0.63	0.40
Monolingual (native data)	<b>0.87</b>	<b>0.80</b>	<b>0.72</b>	/	/	/	/	/	/	/	/	/
Monolingual (translated data)	/	/	/	0.44	0.42	0.19	0.67	0.58	0.39	0.57	0.57	0.32
Baseline MLSP 2024 (Shardlow et al., 2024b)	0.74	0.74	0.54	0.51	0.52	0.14	0.55	0.52	0.25	0.64	0.66	0.33
Highest mlsp score for English : (Goswami et al., 2024)	0.84	0.79	0.52	0.31	0.32	0.04	0.24	0.19	0.07	0.17	0.18	0.02
Highest mlsp score for French, Spanish and Japanese : (Enomoto et al., 2024)	0.81	0.75	0.51	<b>0.62</b>	<b>0.63</b>	0.27	<b>0.76</b>	<b>0.74</b>	<b>0.49</b>	<b>0.73</b>	<b>0.73</b>	0.41

Table 4: Scores for different languages and methods (Pearson, Spearman, R2)

### 3.5.1 Evaluated methods

Since we only have test and development data for the MLSP 2024 dataset, we will use for training the LCP 2021 dataset (Shardlow et al., 2021) containing 7,662 single-word instances exclusively in English. The evaluated methods using a multilingual training approach are the following:

- **Multilingual (LCP 2021)**: the LCP model is based on multilingual word embeddings and is trained exclusively on English data from LCP 2021 task;
- **Multilingual (LCP 2021 + Dev)**: the model based on multilingual word embeddings is trained on LCP 2021 (English data) augmented with the development data in the 10 languages of the MSLP 2024 task (around 30 instances per language) to improve adaptation to the target languages;
- **Multilingual (LCP 2021 + Dev + translated data)**: the model based on multilingual word embeddings is trained on the training data of Multilingual (LCP 2021 + Dev), augmented with 2k or 4k instances from LCP 2021 translated to the target language.

For the monolingual training setting, we evaluated the following approaches for which the LCP model is specific to each target language:

- **Monolingual (native data)**: the LCP model is trained on native data in the target language; this experiment is only performed for English using the LCP 2021 as training data.
- **Monolingual (translated data)**: the model is trained on the translation of LCP 2021

training data (English) to the target language; this experiment is performed on all languages but English.

### 3.5.2 Results

Table 4 presents the evaluation for predicting word complexity in English, French, Spanish, and Japanese using the learning methods presented in section 3.5.1. The evaluation metrics include the Pearson, Spearman, and  $R^2$  scores, as is usually done for this task (cf. Shardlow et al. (2021)). The results of the best MSLP 2024 systems and of the official baseline are also provided for the sake of comparison:

- **Baseline Model**: The baseline is based on linear regression and is trained using log-frequency on the trial set for each language;
- **GMU Team (Goswami et al., 2024)**: Employed a weighted ensemble of mBERT, XLM-R, and language-specific BERT models. All trial data was used for cross-lingual training and evaluation. For English, they augmented the data with the CompLex dataset (Shardlow et al., 2020).
- **TMU-HIT Team (Enomoto et al., 2024)**: Used a chain-of-thought based prompting method employing GPT-4 to generate an instruction in English, and subsequently assigned complexity scores to target words across all languages based on the English instruction.

In English, the Monolingual method, specific to the target language, achieved the best scores (Pearson 0.87, Spearman 0.80,  $R^2$  0.72), thanks to the use of specific annotated data. For French

and Japanese, Multilingual methods trained exclusively on English outperformed the monolingual method based on translation, indicating that multilingual training can be beneficial when annotated data is limited. Adding small amounts of multilingual development data (Multilingual (LCP 2021 + dev)) slightly improved performance in French and Japanese. However, increasing the data through translation (Multilingual(LCP 2021 + Dev + 2k or 4k translated data) did not yield significant improvements. The best scores for French, Spanish, and Japanese were achieved by Enomoto et al. (2024), suggesting that their approach is more effective for these languages.

#### 4 Prompting Large Language Models for multilingual lexical complexity prediction

In this section, we focus on assessing the capability of generative large language models (LLMs) to predict the complexity of a word based on its context. To do this, we use three types of prompt strategies:

- **Zero-shot prompt (base):** The model receives instructions without any specific examples on how to perform the task, relying solely on the knowledge acquired during its training. (See Appendix A)
- **One-shot prompt (instruct):** This type of prompt includes some guidelines used during data annotation, along with an example, thus providing a frame of reference for the model. (See Appendix A)
- **Chain-of-thought prompt (Advanced COT):** This prompt includes detailed annotation instructions, methodological steps to follow and analysis before delivering an evaluation, illustrated by an example (See Appendix A).

#### 4.1 Experimental settings

For this evaluation, we use five different language models: gpt-4o (June 10, 2024)<sup>3</sup>, Llama3 (Dubey et al., 2024), Mistral (Jiang et al., 2023), Phi3 (Abdin et al., 2024), and Gemma (Team et al., 2024). The last four models are used in their 4-bit quantized versions. It’s important to note that comparing these models might seem unfair if gpt-4o is

<sup>3</sup>gpt-4o : <https://openai.com>

included, however, our main goal remains to analyze the effectiveness of each type of prompt according to the model. Yet, the comparison in terms of performance remains relatively fair if gpt-4o is excluded, considering all other models share the same type of quantization. Nonetheless, the number of parameters of each model must be considered, for example, Phi3 with 3.8 billion parameters is significantly less than Gemma, which has 9 billion, while Mistral and Llama are approximately similar in size. We use Ollama<sup>4</sup>, an open-source tool, to test these different LLMs, keeping the default settings provided. All the prompts are written in English, but they explicitly indicate the target language.

Detailed evaluation of these strategies is first undertaken using the MLSP 2024 dataset (Shardlow et al., 2024a). For this task, the generative models are asked through the prompts to predict a score on a scale (0, 0.25, 0.5, 0.75, 1) for the target word in a given context in the target language, in order to mimic the human annotators of the dataset. The evaluation metrics include the Pearson, Spearman, and R<sup>2</sup> scores, as is usually done for this task (cf. Shardlow et al. (2021)). We used a subset of the available languages (English, French, Japanese, and Spanish). In addition, we also evaluate on the binary classification data from CWI 2018 in French, German, and Spanish, adapting the prompts to each task and using the F1 score for evaluation (See Appendix A).

For the sake of comparison between the supervised approach and this one, we also provide the performance of a model specifically trained on this task using a multilingual supervised approach.

#### 4.2 Results

In this part, we will evaluate the various prompt strategies for various LLMs for two different datasets: LCP 2018 and MLSP 2024.

##### 4.2.1 CWI 2018

Table 5 presents the F1 scores for predicting word complexity based on context in French, German, and Spanish. The supervised method achieves the best results across all three languages. Among the language models, gpt-4o and Llama3 display the highest performance. For gpt-4o, the **Instruct** prompt yields the best scores in German and Spanish, while the **Base** prompt performs better in French. The Mistral model shows weak

<sup>4</sup><https://ollama.com>

Model	Version	French	German	Spanish
gpt-4o	Adv COT	0.637	0.694	0.676
	Base	0.672	0.628	0.447
	Instruct	0.602	0.699	0.683
llama3	Adv COT	0.597	0.654	0.654
	Base	0.550	0.630	0.637
	Instruct	0.600	0.671	0.603
mistral	Adv COT	0.198	0.183	0.131
	Base	0.516	0.646	0.673
	Instruct	0.410	0.371	0.281
phi3	Adv COT	0.578	0.667	0.609
	Base	0.551	0.642	0.653
	Instruct	0.493	0.516	0.395
gemma	Adv COT	0.462	0.577	0.594
	Base	0.452	0.563	0.578
	Instruct	0.468	0.587	0.608
Supervised (our approach)	-	<b>0.799</b>	<b>0.791</b>	<b>0.800</b>

Table 5: F1 score comparison across different languages, models and prompting strategy for CWI 2018

Model	Version	English			French			Spanish			Japanese		
		P	S	R2	P	S	R2	P	S	R2	P	S	R2
gpt-4o	Base	0.736	0.735	0.153	0.505	0.509	0.207	0.659	0.643	0.149	0.595	0.621	0.241
	Instruct	0.759	0.665	0.142	0.545	0.555	0.205	0.667	0.645	0.194	0.421	0.404	0.381
	Adv COT	0.781	0.670	0.144	0.542	<b>0.554</b>	0.192	<b>0.680</b>	<b>0.654</b>	0.165	0.574	0.594	0.315
Phi3 3.8B	Base	0.230	0.207	0.229	-0.022	-0.036	0.299	0.233	0.214	0.221	0.110	0.210	0.259
	Instruct	0.414	0.444	0.166	0.093	0.090	0.250	0.276	0.288	0.171	0.244	0.290	0.219
	Adv COT	0.412	0.484	0.151	0.107	0.194	0.284	0.208	0.290	0.244	0.137	0.249	0.259
LLama3 8.0B	Base	0.374	0.418	0.379	0.136	0.146	0.363	0.265	0.278	0.317	0.129	0.158	0.403
	Instruct	0.555	0.519	0.147	0.180	0.170	0.229	0.382	0.376	0.152	0.252	0.253	0.184
	Adv COT	0.657	0.614	0.134	0.276	0.284	0.225	0.384	0.364	0.165	0.346	0.344	0.283
Mistral 7.2B	Base	0.461	0.489	0.394	0.166	0.149	0.309	0.400	0.397	0.355	0.125	0.122	0.388
	Instruct	0.612	0.579	0.139	0.212	0.188	0.220	0.540	0.529	0.152	0.259	0.256	0.153
	Adv COT	0.675	0.594	0.160	0.315	0.283	0.213	0.532	0.528	0.191	0.364	0.368	0.163
Gemma 9b	Base	0.123	0.169	0.482	0.038	0.063	0.433	0.175	0.180	0.384	0.137	0.135	<b>0.455</b>
	Instruct	0.322	0.360	0.320	0.185	0.189	0.311	0.395	0.407	0.227	0.260	0.270	0.279
	Adv COT	0.401	0.440	0.323	0.230	0.253	0.370	0.376	0.394	0.267	0.222	0.227	0.434
Supervised (our approach)	-	<b>0.87</b>	<b>0.80</b>	<b>0.72</b>	<b>0.56</b>	0.52	<b>0.31</b>	0.67	0.63	<b>0.45</b>	<b>0.66</b>	<b>0.67</b>	0.43

Table 6: Model performance comparison across different Languages and prompting strategies for MLSP 2024 (P:Pearson, S:Spearman, R2:  $R^2$ )

performance with the **Advanced COT** prompt but significantly improves with the **Base** prompt. These findings suggest that the effectiveness of the prompt type depends on both the model and the language, highlighting the need to adapt prompt strategies according to the language and the model in use.

We then tried to replicate the annotation process using LLMs for the CWI 2018 dataset where an instance is labeled as complex if any annotator finds the word complex, assigning a value of 1, otherwise 0. For this, given a prompt strategy, each LLM play the role of a single annotator. We will simulate the annotation process using LLMs, where 5 LLMs and 3 different prompt strategies generate a total of 15 annotations. If any of the annotations equals 1, the final annotation is set to 1, otherwise, it is set to 0. Thereafter, this method is called AT\_LEAST\_1. For comparison purposes, we also implemented a majority vote annotation

method (thereafter VOTING\_MAX), where the final label for a given instance corresponds to the most frequent label among the 15 LLM annotations.

Method	Fr	De	Es
AT_LEAST_1	0.45	0.56	0.57
VOTING_MAX	0.62	0.69	0.70

Table 7: The F1 scores for French, German, and Spanish using two voting strategies.

Table 7 shows that the score obtained using the single annotation method is significantly lower than that achieved by majority voting and is also lower than using a single LLM, gpt-4o (Base). However, the results from majority voting are relatively close to those of gpt-4o (Base) as seen in Table 5. It is also believed that VOTING\_MAX performs better than AT\_LEAST\_1, as a single vote out of 15 can lead to errors if an underperform-

ing LLM votes 1, causing the instance to be annotated as 1. Majority voting helps mitigate this issue by considering the decision of the majority of the LLMs.

#### 4.2.2 MLSP 2024

Figure 2 displays the Pearson correlation scores for each prompt type used for each LLM. It shows certain trends across different languages.

**English:** There is a progressive improvement from "base" to "advanced COT". This suggests better predictions in more complex configurations. gpt-4o notably performs better than other models with a score of 0.78. There is also a significant difference between the "base" and "instruct" prompts, while the gap between "instruct" and "advanced COT" is closer.

**French and Spanish:** gpt-4o shows continuous improvement, similarly to the trends observed in English, although the scores are more moderate. Nearly all models demonstrate improvement when going to more complex prompts.

**Japanese:** There are noticeable drops for complex prompts, which may indicate a sensitivity to the types of prompts used for Japanese.

**Supervised Model (cf. table 6):** The supervised multilingual approach described in section 3 outperforms in most cases our LLM prompting strategy, despite the lack of training data for French, Spanish, and Japanese. This has to be further investigated given the results of the best MLSP 2024 system based on a different prompting strategy with a different LLM.

The analysis of Pearson correlation scores for predicting lexical complexity (in Figure 2 and table 6) reveals a clear trend where the "advanced COT" (Chain of Thought) configurations generally achieve the best performance across various languages (French, English, and Spanish). This approach, which incorporates more detailed instructions or chain-of-thought reasoning, appears to better capture the nuances of lexical complexity compared to simpler "zero shot" and "one shot with instruction" approaches. This superiority is reflected in higher Pearson scores, indicating a stronger linear correlation between the predictions and actual values.

Observations made in English, French, and Spanish do not parallel those in Japanese, which presents a unique structure that includes mixed-script writing, the absence of clear word delimitation, and grammatical specificity. This under-

scores the necessity of using specially designed prompts for this language when predicting lexical complexity. The distinctive features of Japanese, such as kanji and grammatical particles, require a more targeted approach to effectively capture lexical complexity. By adapting prompts to the particularities of Japanese, it may be possible to enhance the accuracy of predictions by accounting for these variations.

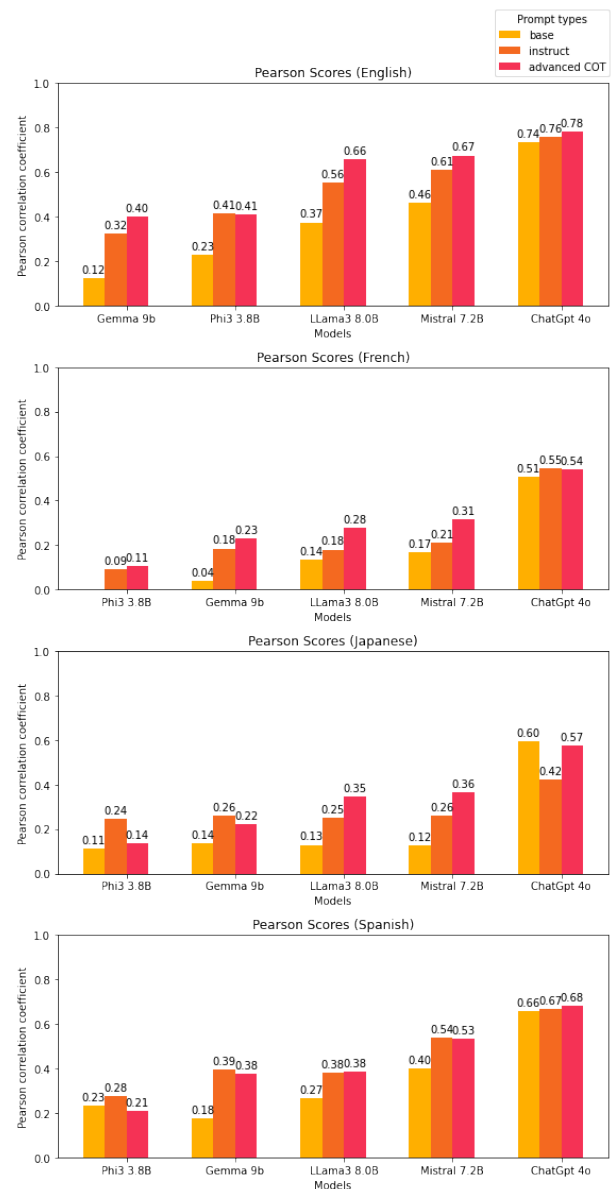


Figure 2: Correlation score for each llm based on the prompt type.

### 4.3 Are large language models (LLMs) a good alternative for multilingual lexical complexity prediction ?

While correlation scores are quite good for the MLSP 2024 dataset, R2 scores, which indicate the quality of prediction, suggest otherwise, cf. Table 6. Zero-shot generative models are not optimized for the specifics of a particular task. Although they can capture a linear relationship, they are less accurate in explaining the total variance of task-specific data, resulting in a lower R2 score. More specifically, asking an LLM to predict a score on a scale of five discrete values (0, 0.25, 0.5, 0.75, 1) penalizes it with respect to the way the dataset is annotated where each instance is annotated with a continuous value between 0 and 1 being the average of multiple human annotations. An intuitive method to address this issue with an LLM is to have it generate multiple outputs and then calculate the average, which might better disperse the data. Table 8 displays the average scores of gpt-4o with varying generation counts  $n$  (1, 10, 20, 30) for English. We have also included a model specifically trained for this task to facilitate comparison.

Models	P	S	R2
gpt-4o (n=1)	0.781	0.67	0.14
gpt-4o (n=10)	0.789	0.677	0.174
gpt-4o (n=20)	0.796	0.677	0.174
gpt-4o (n=30)	0.792	0.687	0.183
Supervised (ours)	<b>0.87</b>	<b>0.80</b>	<b>0.72</b>

Table 8: Performance metrics of gpt-4o vs Trained model for English (P:Pearson, S:Spearman, n:number of generations)

Table 8 indicates that the Pearson correlation scores do not increase significantly, with only slight improvements in the R2 score, which remains quite low compared to the 0.72 achieved by the model trained with a supervised approach.

**What are the consequences of a low R2 score in this task?** Let’s take the example of the multilingual supervised model and gpt-4o (n=30) and analyze the scatter plot of each one’s predictions. Graphs 3 and 4 illustrate the relationship between actual labels and the values predicted by two different models.

Graph 3 for gpt-4o shows a general trend that is well captured by the regression line, but with dispersion concentrated around the values (0, 0.25,

0.5, 0.75, 1), indicating larger prediction errors. On the other hand, Graph 4 displays a better fit between the predictions and the labels, with points more densely clustered around the regression line, suggesting increased accuracy and superior overall performance of the model.

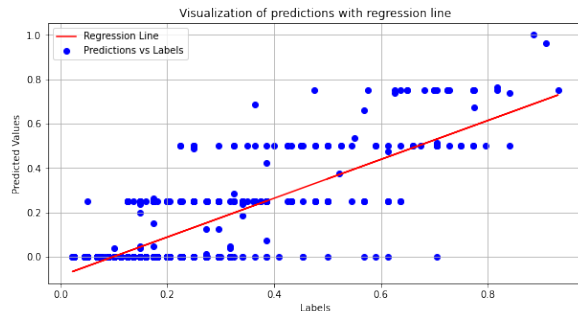


Figure 3: Scatter plot of gpt-4o’s predictions (R=0.792,R2=0.183)

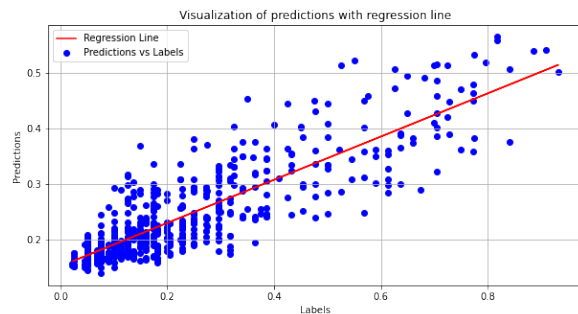


Figure 4: Scatter plot of trained model predictions (R=0.87,R2=0.72)

Graphs 5 and 6 display the dispersion of residuals  $e_i$  around the zero line.

$$e_i = y_i - \hat{y}_i$$

Each residual plot exhibits distinct characteristics reflecting the performance of two different prediction models. In Figure 6, the residuals are primarily concentrated around the mean prediction values (0.2 to 0.4), with a high density near the zero line, suggesting enhanced accuracy of the model within this range. A slight tendency to underestimate higher values is also observed, indicating a potential bias in the model. In contrast, Figure 5 shows a broader dispersion of residuals across all prediction values, with significant variations and distinct peaks at specific points (0.0, 0.2, 0.5, 0.8), suggesting a poorer fit of the model and reduced reliability, especially at the extremes.

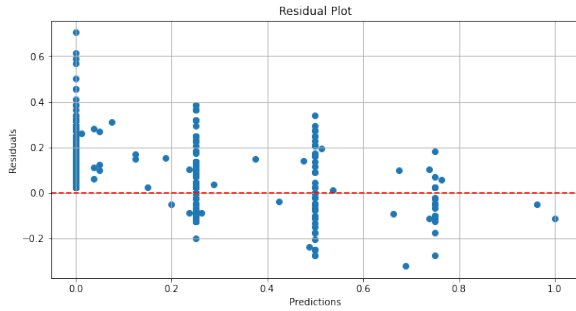


Figure 5: Residual plot for gpt-4o ( $R=0.792, R^2=0.183$ )

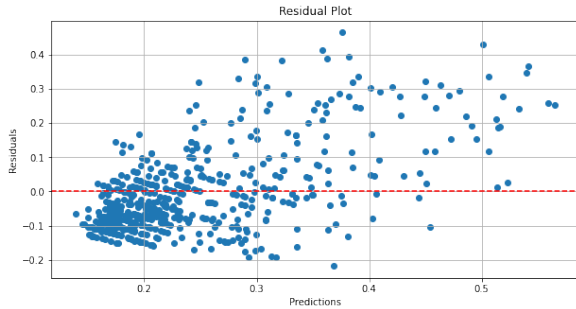


Figure 6: Residual plot for trained model ( $R=0.87, R^2=0.72$ )

#### 4.4 Can the R2 score be improved with large language models?

A good R2 score indicates better predictive quality of the model, with predicted values being closer to the actual values. In the dataset used, each instance is annotated by several evaluators who assess the complexity of a word on a five-point scale, with the final score being the average of these assessments. It is known that each evaluator may differ from each others in terms of level, and the score they assign also depends on their understanding of the instructions and their thought process before giving a score. Additionally, they can make errors. This process is very similar to that of LLMs: for example, we have seen in previous experiment that gpt-4o provides better results compared to others. Thus, we can imagine that the group of evaluators is analogous to a set of LLMs.

To test this hypothesis, we asked the five LLMs used in this experiment (gpt-4o, Llama3, Mistral, Phi3, and Gemma) to predict the score on a five-point scale (0, 0.25, 0.5, 0.75, 1) using the best prompt for English (advanced COT). We then calculated the average of these scores.

Table 9 presents the average and weighted average of LLM models compared to a single LLM and a model specifically trained for this task. The

Model	P	S	R2
One llm (gpt-4o)	0.781	0.670	0.144
Average All llm	0.710	0.673	0.450
Weighted average	0.792	0.717	0.610
Supervised (ours)	<b>0.870</b>	<b>0.800</b>	<b>0.720</b>

Table 9: Average and weighted average of large language models (LLMs) versus one LLM and a trained model.

weighted average is calculated by arbitrarily assigning weights to each LLM based on previously observed performances, as shown in Figure 2. The assigned weights are as follows: gpt-4o at 0.5, Mistral at 0.2, Llama3 at 0.1, Phi3 at 0.1, and Gemma at 0.1. These weights are used to determine if performance can be improved. Ideally and fairly, these weights should be derived from the training set and applied to the test set. As demonstrated in Table 9, the average score for all LLMs significantly improves the R2 score to 0.45, which is a substantial improvement compared to using a single LLM that scores 0.14. Performance further enhances with the use of a weighted average of 0.61, approaching the score of the model specifically trained for this task. These results strongly support our initial hypothesis. In conclusion, the use of multiple LLMs somewhat simulates the way data is annotated, providing better results in terms of R2 score.

## 5 Conclusion

In this study, we explored new methods aiming at enhancing the prediction of lexical complexity in a multilingual context using two distinct types of models: models trained specifically for the task in a supervised way and generative models not specifically trained for the task.

Regarding the supervised approach, our findings indicate that models trained on multiple languages outperform monolingual ones. Zero-shot models trained on multiple languages but the target one displayed variable performance compared with monolingual models. We also observed that data augmentation through automatic translation from English to the target language is feasible, although the required amount of augmentation instances may vary depending on the use case. Additionally, training a model directly from translated data is possible reasonable alternative, as we did

for French 3.

We further investigated the capabilities of generative models to predict lexical complexity on the MLSP 2024 dataset by varying the prompt strategy used. The results underscore the importance of prompt selection, with the "chain of thought" prompt proving particularly effective in English, French, and Spanish 2. However, this approach was not as effective for Japanese, a language that significantly differs from the others and might require a specially adapted prompt due to its unique complexity evaluation rules. Additionally, the findings for CWI 2018 reveal that the supervised approach outperforms our LLM prompting approaches. Majority voting further improved annotation quality.

Although generative models show good Pearson correlation scores, the quality of their predictions remains questionable, often due to very low R2 scores. To address this, we proposed an ensemble method using several generative models, which is akin to the human annotation process (cf. table 9). This opens new research perspectives.

## References

- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- David Alfter. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*.
- David Alfter and I. Pilán. 2018. [Sb@gu at the complex word identification 2018 shared task](#). pages 315–321.
- David Alfter and Elena Volodina. 2018. Towards single word lexical complexity prediction. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 79–88.
- S. Aroyehun, Jason Angel, D. Alvarez, and Alexander Gelbukh. 2018. [Complex word identification: Convolutional neural network vs. feature engineering](#). pages 322–327.
- Joachim Bingel and Johannes Bjerva. 2018. [Cross-lingual complex word identification with multitask learning](#). pages 166–174.
- José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PMLADC at ICLR 2020*.
- Branden Chan, Stefan Schweter, and Timo Möller. 2020. [German’s next language model](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Abhinandan Desai, Kai North, Marcos Zampieri, and C. Homan. 2021. [Lcp-rit at semeval-2021 task 1: Exploring linguistic features for lexical complexity prediction](#). *ArXiv*, abs/2105.08780.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Taisei Enomoto, Hwichan Kim, Toshio Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. Tmu-hit at mlsp 2024: How well can gpt-4 tackle multilingual lexical simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.
- Thomas François, Núria Gala, Patrick Watrin, and Cédric Fairon. 2014. [Flelex: a graded lexical resource for french foreign learners](#). In *International conference on Language Resources and Evaluation (LREC 2014)*.
- Núria Gala, Thomas François, and Cédric Fairon. 2013. Towards a french lexicon with difficulty measures: Nlp helping to bridge the gap between traditional dictionaries and specialized lexicons. In *eLex-Electronic Lexicography*.
- Deepanway Ghosal, Yew Ken Chia, Navonil Majumder, and Soujanya Poria. 2023. [Flacuna: Unleashing the problem solving power of vicuna using flan fine-tuning](#). *ArXiv*, abs/2307.02053.
- Sian Gooding and E. Kochmar. 2019. [Complex word identification as a sequence labelling task](#). pages 1148–1153.
- Dhiman Goswami, Kai North, and Marcos Zampieri. 2024. [Gmu at mlsp 2024: Multilingual lexical simplification with transformer models](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 627–634.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#).

- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Abdelhak Keliou, Mathieu Constant, and Christophe Coeur. 2024. Complex word identification: A comparative study between ChatGPT and a dedicated model for this task. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3645–3653, Torino, Italia. ELRA and ICCL.
- Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xiaoxia Zhou. 2023. Better zero-shot reasoning with role-play prompting. *ArXiv*, abs/2308.07702.
- Michal Konkol. 2016. Uwb at semeval-2016 task 11: Exploring features for complex word identification. pages 1038–1041.
- Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq R. Joty, and J. Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. pages 431–469.
- S. Malmasi, M. Dras, and Marcos Zampieri. 2016. Ltg at semeval-2016 task 11: Complex word identification with classifier ensembles. pages 996–1000.
- Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.
- O. Ogundare, S. Madasu, and N. Wiggins. 2023. Industrial engineering with large language models: A case study of chatgpt’s performance on oil gas problems. *ArXiv*, abs/2304.14354.
- Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.
- Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. 2021. Deepblueai at semeval-2021 task 1: Lexical complexity prediction with a deep ensemble approach. pages 578–584.
- Sandaru Seneviratne and Hanna Suominen. 2024. Anu at mlsp-2024: Prompt-based lexical simplification for english and sinhala. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 599–604.
- Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Pérez Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Marcos Zampieri, and Horacio Saggion. 2024a. An Extensible Massively Multilingual Lexical Simplification Pipeline Dataset using the MultiLS Framework. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding Difficulties (READI)*.
- Matthew Shardlow, Fernando Alva-Manchego, Riza Theresa Batista-Navarro, Stefan Bott, Saul Calderon-Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, and Anna Huelsing. 2024b. The bea 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589.
- Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. Complex: A new corpus for lexical complexity prediction from likert scale data. *arXiv preprint arXiv:2003.07008*.
- Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.
- Anaïs Tack, Thomas François, Piet Desmet, and Cédric Faron. 2018. Nt2lex: A cefr-graded lexical resource for dutch as a foreign language linked to open dutch wordnet. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 137–146.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Giuseppe Vettigli and A. Sorgente. 2021. Compna at semeval-2021 task 1: Prediction of lexical complexity analyzing heterogeneous features. pages 560–564.
- Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331, St. Julian’s, Malta. Association for Computational Linguistics.
- Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. Gpt4tools: Teaching large language model to use tools via self-instruction. *ArXiv*, abs/2305.18752.
- Seid Muhie Yimam, Chris Biemann, S. Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs



- Tack, and Marcos Zampieri. 2018a. [A report on the complex word identification shared task 2018](#). *ArXiv*, abs/1804.09132.
- Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018b. [A report on the complex word identification shared task 2018](#). In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.
- Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017. [Multilingual and cross-lingual complex word identification](#). pages 813–822.
- Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. [How well do large language models perform in arithmetic tasks?](#) *ArXiv*, abs/2304.02015.
- George-Eduard Zaharia, Dumitru-Clementin Cercel, and M. Dascalu. 2020. [Cross-lingual transfer learning for complex word identification](#). *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 384–390.
- Marcos Zampieri, S. Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. [Complex word identification: Challenges in data annotation and system performance](#). *ArXiv*, abs/1710.04989.

## A Appendix

### 1- Zero-shot prompt (base)

''''''

You will be given a sentence and a word included in the sentence. Evaluate the complexity of the word in the context of the sentence, and provide a rating in scale of 0.0, 0.25, 0.5, 0.75, 1.0.

Sentence: '{sentence}'

Word: '{token}'

Complexity:

return only the number (0.0, 0.25, 0.5, 0.75, 1.0) that corresponds to the complexity of the word in context.

''''''

### 2- One-shot prompt (instruct)

''''''

You are a person without specialized knowledge or expertise in any specific field. You will receive a sentence containing a word, your task is to evaluate the word based on one metric.

Evaluation Criteria:

Complexity [0.0, 0.25, 0.5, 0.75, 1.0]: This measures how difficult it is to understand the word.

1. Carefully examine the sentence and the specified word to grasp the context in which it is used.
2. Assess the complexity of the word using the criteria provided
  - 0.0: The word is simple and easily understandable to most people.
  - 0.25: The word may have some complexity or be specific to a certain field, but can still be understood with some effort.
  - 0.5: The word is moderately complex and may require some background knowledge or explanation to understand fully.
  - 0.75: The word is quite complex and may be difficult to understand without significant knowledge or explanation.
  - 1.0: The word is extremely complex and likely only understood by experts or individuals with specialized knowledge.

Your personal knowledge of a word should not influence your rating. Instead, rate the word based on the understanding an average person might have

#### Example:

Sentence: 'The professor's discourse was filled with intricate terminology that baffled the students.' Word: 'discourse'.

For this example, 'discourse' might be rated as 0.25.

Please provide a complexity rating for the '{language}' word '{token}'.

Sentence: '{sentence}'

Word: '{token}'

return only the number (0.0, 0.25, 0.5, 0.75, 1.0) that corresponds to the complexity of the word.

""""

### 3- Chain-of-thought prompt (Advanced Cot)

""""

You are a person without specialized knowledge or expertise in any specific field. You will receive a sentence containing a word, your task is to evaluate the word based on one metric.

Evaluation Criteria:

Complexity [0.0, 0.25, 0.5, 0.75, 1.0]: This measures how difficult it is to understand the word.

#### Evaluation steps:

- **1. Understand the Context:** - Read the sentence and the word carefully to understand the context in which the word is used.
- **2. Analyze the Word's Frequency and Familiarity:** - Determine how commonly the word is used in everyday language. - Consider if the word is generally known by the average person or if it is specialized.
- **3. Evaluate the Morphological Complexity:** - Examine the structure of the word, including its length, composition, and any prefixes or suffixes.
- **4. Define the Word:** - Provide a definition of the word in its common usage. - Explain the specific meaning of the word in the given context.
- **5. Assess the Overall Complexity:** - Based on the analyses above, determine the complexity of the word using the following criteria: - 0.0: The word is simple and easily understandable to most people. - 0.25: The word may have some complexity or be specific to a certain field, but can still be understood with some effort. - 0.5: The word is moderately complex and may require some background knowledge or explanation to understand fully. - 0.75: The word is quite complex and may be difficult to understand without significant knowledge or explanation. - 1.0: The word is extremely complex and likely only understood by experts or individuals with specialized knowledge.
- **6. Assign a Complexity Rating:** - Based on your evaluation, assign a complexity rating to the word.

Your personal knowledge of a word should not influence your rating. Instead, rate the word based on the understanding an average person might have

**Example:**

Sentence: 'The professor's discourse was filled with intricate terminology that baffled the students.' Word: 'discourse'

1. Understand the Context: The word 'discourse' is used in a sentence discussing a professor's speech.
2. Analyze the Word's Frequency and Familiarity: 'Discourse' is somewhat specialized but can be understood by most people with some effort.
3. Evaluate the Morphological Complexity: 'Discourse' is a relatively long word but does not have complex prefixes or suffixes.
4. Define the Word: - Common usage: 'Discourse' means written or spoken communication. - Context-specific: In the sentence, 'discourse' refers to the professor's lecture.
5. Assess the Overall Complexity: Considering its moderate frequency, moderate morphological complexity, and clear context-specific meaning, 'discourse' might be rated as 0.25.
6. Assign a Complexity Rating: For this example, 'discourse' might be rated as 0.25.

Now, Please provide a complexity rating for the '{language}' word '{token}'.

Sentence: '{sentence}'

Word: '{token}'

return only the number (0.0, 0.25, 0.5, 0.75, 1.0) that corresponds to the complexity of the word.

""""

**4- Zero-shot prompt (base-binary)**

You will receive a sentence and a specific word from that sentence. Evaluate the complexity of the word within the context of the sentence and return 1 if the word is complex, or 0 if it is easy.

Sentence: 'sentence'

Word: 'token'

Complexity:

return only the complexity score: 1 or 0.

**5- One-shot prompt (instruct-binary)**

You are an individual without specialized knowledge or expertise in a specific area.

You will be given a sentence and a word included in the sentence.

Your task is to evaluate the complexity of the word in a binary format (0 or 1).

Please read and understand these instructions carefully. Keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Complexity (0, 1): Evaluate how difficult the word is to understand for an average person.

- 0: The word is simple and easily understandable by most people. - 1: The word is complex and may be difficult for an average person to understand.

Evaluation steps: 1. Read the sentence and word carefully to understand the context.

2. Determine the complexity of the word based on the criteria above.

3. Assign a complexity rating to the word.

Note: Your own familiarity with the word should not impact your rating. Base your judgment on an average person's understanding of the word.

**Example:**

Sentence: 'The professor's discourse was filled with intricate terminology that baffled the students.' Word: 'discourse'.

For this example, 'discourse' might be rated as 1.

Please assign a complexity rating to the 'lang' word.

Sentence: 'sentence'

Word: 'token'

Complexity:

return only the number (0 or 1) that corresponds to the complexity of the word.

## **6- Chain-of-thought prompt (Advanced COT-binary)**

You are an individual without specialized knowledge or expertise in a specific area.

You will be given a sentence and a word included in the sentence.

Your task is to rate the word on one metric: complexity.

Please read and understand these instructions carefully. Keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Complexity (0 or 1): the complexity of a word in terms of how difficult the word is to understand.

**Evaluation steps:**

- 1. Understand the Context:** - Read the sentence and the word carefully to understand the context in which the word is used.

- 2. Analyze the Word's Frequency and Familiarity:** - Determine how commonly the word is used in everyday language. - Consider if the word is generally known by the average person or if it is specialized.
- 3. Evaluate the Morphological Complexity:** - Examine the structure of the word, including its length, composition, and any prefixes or suffixes.
- 4. Define the Word:** - Provide a definition of the word in its common usage. - Explain the specific meaning of the word in the given context.
- 5. Assess the Overall Complexity:** - Based on the analyses above, determine the complexity of the word using the following criteria: - 0: The word is simple and easily understandable to most people. - 1: The word is complex and may be difficult to understand for the average person.
- 6. Assign a Complexity Rating:** - Based on your evaluation, assign a complexity rating to the word.

Note: Your own familiarity with the word should not impact your rating. This should be based on an average person's understanding of the word.

**Example:**

Sentence: 'The professor's discourse was filled with intricate terminology that baffled the students.' Word: 'discourse'

1. Understand the Context: The word 'discourse' is used in a sentence discussing a professor's speech.
2. Analyze the Word's Frequency and Familiarity: 'Discourse' is somewhat specialized but can be understood by most people with some effort.
3. Evaluate the Morphological Complexity: 'Discourse' is a relatively long word but does not have complex prefixes or suffixes.
4. Define the Word: - Common usage: 'Discourse' means written or spoken communication. - Context-specific: In the sentence, 'discourse' refers to the professor's lecture.
5. Assess the Overall Complexity: Considering its moderate frequency, moderate morphological complexity, and clear context-specific meaning, 'discourse' might be rated as 0.

Now, apply this method to the given word and sentence.

Please assign a complexity rating to the 'lang' word.

Sentence: 'sentence'

Word: 'token'

Complexity:

Please return only the number (0 or 1) that corresponds to the complexity of the word. Do not include any additional information or explanations.

# Developing a Pedagogically Oriented Interactive Reading Tool with Teachers in the Loop

Mihwa Lee, Björn Rudzewitz and Xiaobin Chen

Hector Research Institute of Education Sciences and Psychology,

LEAD Graduate School and Research Network

University of Tübingen, Germany

{mihwa.lee, bjoern.rudzewitz, xiaobin.chen}@uni-tuebingen.de

## Abstract

Reading is an essential life skill and crucial for students' academic success. Particularly, the need for students to read in English as a second language (L2) has grown due to its global significance. However, L2 readers often have limited opportunities for meaningful, interactive reading practice with immediate support. This paper introduces *ARES*, a pedagogically oriented, web-based intelligent computer-assisted language learning (ICALL) system designed to enhance the L2 reading experience, developed through an action research approach involving practitioners. *ARES* offers a range of interactive features for students, including not only the autonomous identification of vocabulary and more than 650 language means, but also making them interactively explorable in the text, providing detailed explanations and practical examples in contexts. To support effective teaching, *ARES* employs a Large Language Model (LLM) for generating tailored reading comprehension questions and answer evaluations, with teachers in the loop, achieving human and Artificial Intelligence (AI) collaboration. We present the development and application of the system from both technical and pedagogical perspectives to advance L2 learning research and refine educational tools.

## 1 Introduction

In today's increasingly globalized world, the growing necessity for students to read in L2 English underscores the importance of proficient L2 reading skills (Vettori et al., 2023). Learning to read in L2 is complex, as learners must grasp literacy in an unfamiliar language (Verhoeven, 2011). Thus, it is important to support L2 learners' reading development, especially in school contexts

where L2 learning most often takes place. However, school teachers often face challenges in providing interactive and meaningful learning experiences for a large number of students due to limited time and highly heterogeneous students with different proficiency levels, native languages, and learning preferences in the same class.

Digital environments, such as ICALL systems, offer unique opportunities for new ways of learning and teaching (Amaral and Meurers, 2011). These systems have been shown to enhance learning engagement (Liu et al., 2016) and achieve better language acquisition (Oberge and Daniels, 2013) through features such as automatic feedback (Ai, 2017), intelligent tutoring (Choi, 2016), and personalized support (Heilman et al., 2010). Despite these advancements, a lot of previous systems are falling short on integrating the AI technologies (e.g., LLMs) or on addressing the practical day-to-day needs of L2 teachers.

In order to address these gaps and enhance real-life usage of ICALL systems in classrooms, we designed and developed an ICALL system that systematically and automatically provides various interactive support for L2 reading, targeting young learners of English as a foreign/second language (EFL/ESL). The development of the system is grounded in theories of text comprehension in second language acquisition (SLA), leveraging the affordances of current language technologies. The general goal of the system is to provide school teachers with a tool to easily create reading activities with interactive and individualized support for their students. Currently, the system provides a web-based platform that features (1) provision of annotations and glossing of vocabulary and language means, (2) automatic generation and evaluation of reading comprehension questions, and (3) easy management of student classes, assignments and their submissions, as well as feedback on assignments. In this article, we introduce the design

---

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

rationale and development of the system with its different elements in detail. We conclude with an outlook on the design of a study that investigates perceptions of the system in German secondary schools. By exploring these dimensions, we aim to showcase how the Natural Language Processing (NLP) and AI technologies can be used to support L2 reading learning and teaching.

## 2 Background

### 2.1 Vocabulary and Grammar Knowledge in L2 Reading Comprehension

Reading is a complex cognitive activity that requires the integration of information from the text and the reader's background knowledge. Successful reading comprehension (RC) depends on skilled processing of the visually presented text (Verhoeven, 2011). It requires a wide range of linguistic as well as non-linguistic skills including word recognition, linguistic knowledge, discourse-level meaning making, reading strategies, inferring, and comprehension monitoring (Grabe, 2014). Current theories on RC typically involve conceptual representations with several interdependent layers. There is typically a local-level representation based on text-based information (i.e., vocabulary, grammar) and a high-level representation where the content of the text becomes integrated into the reader's larger conceptual structure (i.e., integrating the textual information across sentences) (Jung, 2009; Kintsch, 1988; Kintsch and van Dijk, 1978). During the construction of semantic structures at these various levels, a reader's vocabulary and grammatical knowledge influences the entire reading process (Jung, 2009). In particular, the parsing mechanism, driven by this vocabulary and grammar knowledge, operates on text segments assembled locally. Consequently, if readers generate inaccurate or incomplete representations of these local text segments, their overall comprehension of the text can be significantly impaired (Jung, 2009; Koda, 2007). Lexical-syntactic knowledge is critical in the construction of the local-level representation, where text-based propositions are built to eventually support the high-level representation (Choi and Zhang, 2021; Kintsch, 1988). Knowledge of vocabulary and grammar thus helps with the construction of text-based information and eventually facilitates in-depth comprehension.

Following Alderson's (1984) discussion of

whether L2 reading is a reading problem or a language problem, SLA researchers have been interested in the importance of vocabulary and grammar knowledge in an effort to understand the process of L2 RC. A plethora of empirical studies have been conducted to gain a better understanding of how vocabulary and grammar knowledge affect L2 RC, whose results generally support the primacy of both L2 vocabulary and grammar knowledge in L2 RC (Choi and Zhang, 2021). For instance, in a longitudinal study examining the relation of oral language proficiency and decoding skills to L2 RC among Dutch-speaking young EFL learners, Droop and Verhoeven (2003) found that both vocabulary and morphosyntactic knowledge had an equally strong correlation to L2 RC, especially at the initial stage when the learners had relatively low L2 proficiency. Recent meta-analyses in L2 RC (Chen and Mei, 2024; Choi and Zhang, 2021) also demonstrate that L2 vocabulary and grammar knowledge are the two strongest predictors of L2 RC. Hence, it is important to accommodate both types of knowledge in the design of teaching of L2 reading. However, vocabulary and grammar knowledge varies a lot among individuals, requiring support for their development be highly personalized. From an instructional perspective, however, due to the time constraint and students' heterogeneity, it is almost impossible for teachers to pinpoint vocabulary and grammatical knowledge that each learner does not understand while they are reading.

### 2.2 Computer-based Development of L2 Reading Comprehension

Technological applications in L2 reading range from basic digital texts such as e-readers with limited interactivity to online dictionaries to collaborative annotation. Reviews of L2 RC literature (Saeidi and Yusef, 2012; Sawaki, 2001) have shown that specially designed software, ICALL systems, online lessons, animated texts, use of multimedia contexts, interactive multi-modal materials, online dictionaries, e-books and hypertext/hypermedia environments have been used to enhance L2 RC. Here, we describe two features that are highly relevant to our system.

**Online Dictionaries** Primarily used for looking up unknown words in reading, writing, and vocabulary learning activities, online dictionaries often in the form of electronic glossing have been con-



sidered highly feasible, individual learning materials (Çolak and Balaman, 2022) as they “provide controlled opportunities for linguistic input for the learner and interaction with the computer” (Chapelle, 2003, p. 25). One of the prominent examples of electronic glossing is Amazon’s Kindle, which provides users with a dictionary function that presents the definitions of words at the bottom of the screen (Lee and Lee, 2015). Another example is *Readlang* (Ridout, 2013), a commercial platform that provides instant translation of words in texts in multiple languages. In fact, it has been shown that online dictionaries such as glossing enhance L2 RC as well as L2 vocabulary acquisition, as found in a meta-analysis of studies on both electronic and textual glosses (Taylor, 2009). Studies also revealed that L2 learners prefer computerized glossing to its paper counterparts (Bowles, 2004). Traditional online dictionaries, however, constrain the selection of an appropriate meaning among all the possible meanings as well as providing a wider range of information such as collations, as they in general list only straight definitions. Previous literature suggests that examples illustrating syntax, collocation, usage and context are more helpful in clarifying meaning than straight definitions (McAlpine and Myles, 2003). Furthermore, to the best of our knowledge, there has been no attempt to integrate a dictionary on language means (i.e., explanation of forms) into language learning applications.

**Feedback** Feedback is information communicated to learners to modify their thinking or behaviors to close the gap between their actual performance and the target performance (Hattie and Timperley, 2007), thus aiming to improve learning (Shute, 2008), as well as enhance emotions and motivation during learning (Fong et al., 2019). The need for feedback on learner production has been well documented in SLA research (Mackey, 2006). Feedback can be categorized into three types: Knowledge-of-Response (KOR) feedback that only includes verification, Knowledge-of-Correct-Response (KCR) feedback that additionally includes the correct answer, and Elaborated Feedback (EF) that also includes extra-instructional information (Swart et al., 2022) such as explanations (e.g., “In the text, the author does not state that...”), follow up questions (e.g., “Why does the author of the text think...?”), location or hint of the correct information in the text (e.g.,

“Check the part in the text again where the author mentions...”), or a combination of multiple types of information (Finn et al., 2018). Among them, EF can be used to guide and direct the L2 reader, thereby providing additional support. Bown (2017), borrowing words from Mitchell et al. (2013), attests that “from a sociocultural view of L2 acquisition, this support can be considered as a form of scaffolding: a ‘process of supportive dialogue which directs the attention of the learner to key features of the environment, and which prompts them through successive steps of a problem’ (Mitchell et al., 2013, p. 25)”. In fact, in the field of educational sciences, several meta-analyses (Bangert-Drowns et al., 1991; der Kleij et al., 2015; Wisniewski et al., 2020) have demonstrated positive effects of EF over other simpler types of feedback. Despite the potential of EF in L2 RC, only a few attempts have been made to implement it in ICALL systems (Bown, 2017, 2018; Murphy, 2007, 2010). However, most of these research prototypical systems have not been tested widely in schools practically.

Overall, there have been several attempts to integrate features that support L2 RC (e.g., *Readlang*, Bown, 2018; Murphy, 2010), but most existing systems focus on a single aspect of L2 reading support (e.g., vocabulary) and fall short in offering comprehensive, pedagogically grounded support throughout the entire L2 reading process. This poses challenges for a practical implementation in classroom settings. Moreover, most of these systems were research-oriented and not designed for actual widespread classroom usage, further complicating their adoption and effectiveness. Our work seeks to address this gap between research, foreign language pedagogy, and real-life classroom usage by developing *ARES*, a pedagogically oriented, web-based ICALL system designed to enhance L2 reading experience. In the following section, we present the system architecture and each feature of the system in detail.

### 3 ARES

*ARES (Annotated Reading Enhancement System)* is designed as a multi-layer web application that strikes a balance between usability and flexibility. The system implements a responsive design that adapts the display for all devices and platforms. Therefore, it works seamlessly across multiple platforms, requiring only a computer, tablet,

or smartphone with a web browser and internet access. Using NLP tools, the system supports students by identifying and providing glossing on vocabulary and language means with examples, which they can consult as needed while reading the assigned texts. For teachers, ARES automates the process of generating questions for assignments and providing individual feedback to each student response by implementing a pre-trained LLM (ChatGPT 4o<sup>1</sup>), significantly reducing their workload and allowing them to focus more on communicative activities in classrooms.

Involving teachers or stakeholders in education research whose results will be used in schools is considered very important because schools and teachers should not only be treated as consumers of the research results (Farley-Ripple et al., 2018). Successful research that has a practical impact in schools is always the outcome of bi-directional efforts. This bi-directional effort is not a one-off process, but it will involve multiple iterations of interactions between the researchers and the teachers. Consequently, we decided to use a multi-cycle action research paradigm to guide the development and research process. The action research model (Figure 1) is a systematic, collective, collaborative, and self-reflective scientific inquiry aimed at improving educational practices and addressing the practical concerns of teachers (Kemmis and McTaggart, 1988; Rapoport, 1970), where a key characteristic of action research is the involvement of stakeholders, including teachers, students, and researchers.

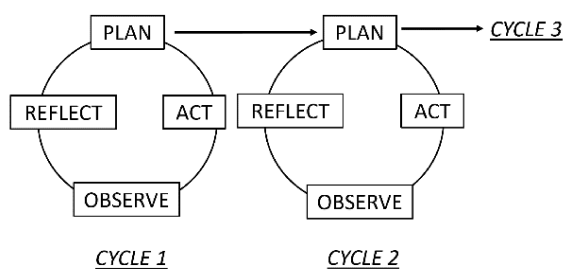


Figure 1: Action Research Model (Kemmis and McTaggart, 1988)

In the following subsections, the system that has been developed in the first phase of the action research paradigm is described in more detail, both from the teacher perspective and the student perspectives.

<sup>1</sup><https://chatgpt.com/>

### 3.1 System Architecture

Utilizing a software-as-a-service (SaaS) approach, ARES provides the software through the cloud, allowing system developers to update the application with new features and fix bugs without requiring users to download updates from app stores. The system is built on a Java backend deployed in a Jetty server. For the display layer, we use the Bootstrap framework, which provides a highly extensible component-based design for an optimized display. In order to enable Learning Analytics, all user activities such as button clicks, lookups of language means, reading comprehension question attempts, assignment submissions, viewing of specific feedback messages, and any other relevant user actions are logged through xAPI<sup>2</sup>, an interoperability specification for recording user interactions, and stored in a Learning Record Store (LRS).

### 3.2 Home Interface

Based on discussion with the involved stakeholders, the home pages that users first see when they log in offer the most commonly used functionalities as a starting point for efficient usage.

**Teacher Home** There are three main sections that teachers can select from, described in detail below:

- **Classes** Teachers can create, delete, edit classes and manage students.
- **Assignments** Teachers can manage assignments and check the results of each assignment.
- **Texts** Teachers can browse, upload, and edit texts.

To address the challenge English teachers face in finding texts appropriate for their students' English levels, the system includes a "text bank" with reading materials covering 12 topics (e.g., History, Travel and Nature, Technology). These materials are crafted by experienced ESL/EFL teachers ensuring users always have access to relevant content from a variety of themes, addressing a need by teachers to search for material to prepare their lessons. The initial target audience is classes in German secondary schools (Gymnasium) with proficiency levels roughly equivalent to A2-B1 according to the Common European Framework of

<sup>2</sup><https://xapi.com/>

Reference for Languages (CEFR) (Council of Europe, 2020). The texts are tailored to match these proficiency levels. Additionally, teachers have the option to upload their own texts, which they can later edit or delete as needed. When creating an assignment, teachers receive automatically generated suggestions for comprehension questions generated from the LLM. With the goal of keeping teachers in the loop, we designed the system so that teachers always hold the ultimate decision-making power, and are supported by the system’s suggestions and tools. They can post-edit these suggestions, confirm them, or add their own questions manually, to ensure that teachers’ expertise is involved in the process. On the technical side, we conducted an iterative approach to refine the prompt for question generation. The full final version of the prompt implemented in the system is attached in the Appendix.

Teachers can decide which annotations on language means to show students (section 3.3), allowing them to tailor assignments and annotations to specific learning goals and ensure appropriateness for their students’ proficiency levels (see Figure 2). The motivation behind this customization is that reading texts often contain a wide range of language means and grammatical structures, and it is often hard for teachers to selectively control students’ focus on a certain language mean in reading texts. By enabling teachers to customize annotations of language means based on learning goals, the system ensures that reading materials support the target structures, making the learning process more efficient and tailored to pedagogical needs.

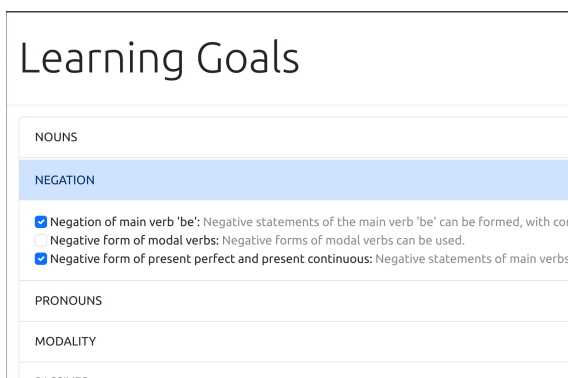


Figure 2: Selection of Annotations of Language Means

**Student Home** The system presents two main options that students need most on their start page:

- **Classes** Students can see classes they are en-

rolled and join a class using a 4-digit access code provided by the teacher.

- **Assignments** Each assignment card indicates the status of an assignment using different background colors and badges (see Figure 3).

Upon clicking or touching the assignment card, students are forwarded to the reading interface.

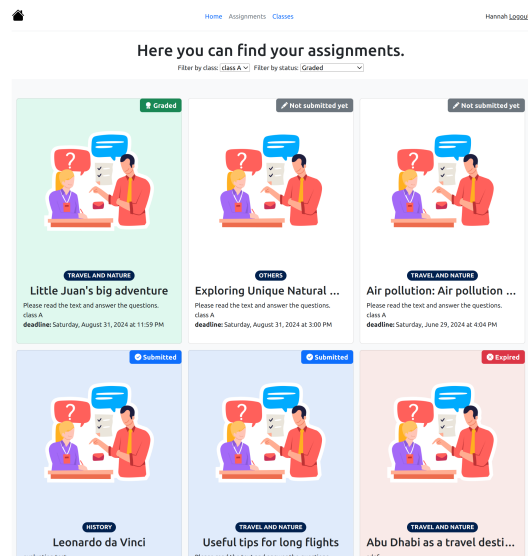


Figure 3: Student Assignments Page

### 3.3 Reading Interface

The main features of the interface are an on-demand annotation on language means that is based on the English Grammar Profile (EGP)<sup>3</sup> and an on-demand vocabulary lookup based on the LLM. Given their relevance to the overall goal of the system, the following subsections describe these functions in detail.

### 3.4 Annotations on Language Means

The annotation function of language means acts as an instant glossing on forms, allowing students to click on any word (or section of a sentence) within a reading text to access its detailed explanation with example sentences and the corresponding CEFR level of the grammatical structure. When a text is uploaded to the system, it is automatically analyzed and indexed by an NLP tool our research group has created to extract form-based language means from the EGP. The EGP is a comprehensive database listing over 650 language means spanning the entire range of CEFR levels. It is based

<sup>3</sup><https://www.englishprofile.org/english-grammar-profile/egp-online>

on an extensive analysis of the Cambridge Learner Corpus, providing insights into the typical grammar usage at each proficiency level (O’Keeffe and Mark, 2017). For each language mean, we asked experienced teachers to write an explanation and examples in both a student-directed and a more concise teacher-directed way. Along with an indication of the CEFR level, this information is shown to the users, with the respective variant of the explanation selected on the user’s role (see Figure 4).

The pipeline for this function is based on the further development of the pipeline introduced in Quimal et al. (2021). It is based on the Unstructured Information Management Architecture (UIMA, Ferrucci and Lally 2004)<sup>4</sup>, an open-source Apache framework used in large-scale text processing applications. It includes three main components: an NLP preprocessing module, an annotator built using UIMA’s Rule-based Text Annotation (Ruta)<sup>5</sup> framework, and an application to run the pipeline for analyzing texts. The NLP preprocessing module employs tools like Standard CoreNLP (Manning et al., 2014)<sup>6</sup> and DKPro Core (de Castilho et al., 2016)<sup>7</sup> for tasks such as tokenization, part-of-speech tagging, and dependency parsing. The Ruta annotator applies regular expression-based rules after the pre-processing to identify specific language means, tagging them with information like construction type and position in the text, ensuring robust and scalable text processing.

### 3.5 Vocabulary Lookup Function

The system offers an instant vocabulary glossing for students. It enables students to click on any word within a reading text and immediately access comprehensive vocabulary information about that word. When a student clicks on a word in the reading text, the system identifies and extracts the clicked word as a token and its surrounding sentence as a context. The LLM is then applied to analyze the word both as an isolated token and within the context of the sentence to understand its specific usage and meaning, including the general definition, meaning in the specific context, collocations, related vocabulary, morphosyntactic ele-

ments of the word, and additional information (see Figure 5). In order to make sure that the students understand the relevant information of the clicked vocabulary, there is an option for them to see a translation of the explanation.

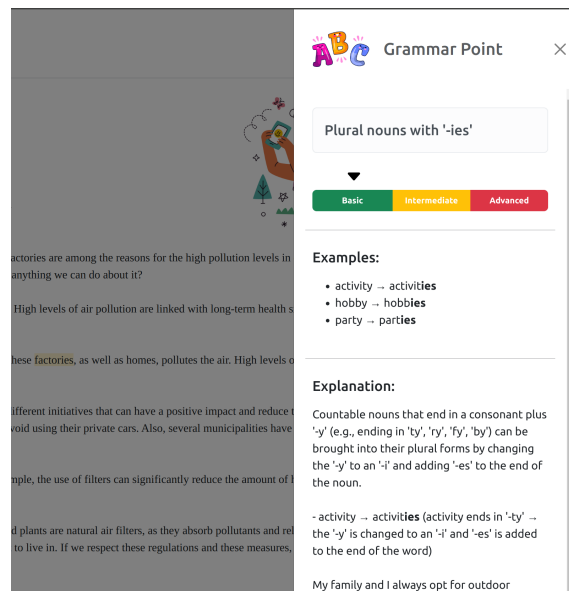


Figure 4: Grammar Lookup

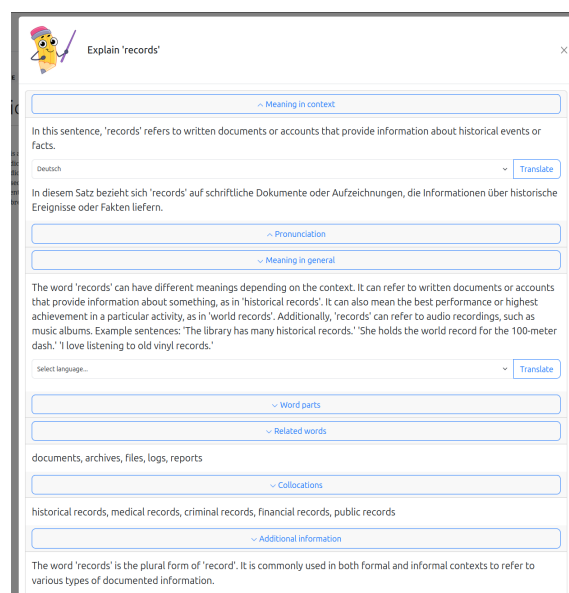


Figure 5: Vocabulary Lookup

### 3.6 Questions and Rating Functionality

For assignments that accompany RC questions, these questions are displayed below the text. Students have the flexibility to complete the assignment without answering all the questions. At the end of the assignment, when students click on the submit button, the system presents a dialogue box

<sup>4</sup><https://uima.apache.org/>

<sup>5</sup><https://uima.apache.org/ruta.html>

<sup>6</sup><https://stanfordnlp.github.io/CoreNLP/>

<sup>7</sup><https://dkpro.github.io/dkpro-core/>

asking students to rate the difficulty and interestingness of the text using a 5-star Likert scale with an option to leave free-form comments about the assignment, which will be provided to the teacher, offering insights into both the overall and individual perceptions of the assigned text to teachers, text authors, and researchers.

Once the students submit the assignment, the system forwards them to the Assignments page (section 3.2). However, students keep the right to access the reading interface even after submitting the responses in order to give them chances to review the finished assignments and view the teacher’s feedback.

## 4 Evaluation Interface

**Teacher Evaluation** On the selection of the assignment in the Teacher Home (section 3.2), the system directs them to the Evaluation Interface, which consists of two main sections as shown in Figure 6. The upper section of the page displays information about individual student submissions in a table format, including the time of submission, automatic score (calculated by the system), manual score (assigned by the teacher), percentage of the feedback read by the student, difficulty rating, interestingness rating, and comments (see Figure 6). With the purpose of reducing the teachers’ workload, we equip the system with functionalities that automate grading by integrating the LLM. Upon clicking the “Grade all automatically” button above the submission table, all student responses are sent to the LLM in a parallelized way for processing. The LLM evaluates the student responses against a target response for each question while also provided with the reading text as context. As the output of this process, the teacher sees a percentage score of correct responses displayed under the “Automatic score” column. Teachers can then transfer these automatic scores to the “Manual score” column by clicking the “Accept all corrections” button. The full final version of the refined prompt to the LLM is attached in the [Appendix](#).

In order to keep teachers in the loop, we allow teachers to review and modify the automated scores by clicking the “Grade” button within the submission table, which redirects them to the individual submission page. Here, detailed evaluation information (questions, student responses, target answers, automatic scores, and automatic

feedback) is displayed, allowing teachers to adjust scores and feedback as needed. If the teacher agrees with the automated grading, they may utilize the “Copy all” button to transfer the automated scores and feedback to the manual grading section. Alternatively, for more granular adjustments, the “Copy” button allows for the selective adoption of scores on an individual question basis. Eventually, what students see is what teachers confirm at the end. This way, although we reduce teachers’ burden of grading, we at the same time make sure that teachers are in full control of what students see.

The lower section of the page provides a summative assessment of the assignment, including the number of submissions, average automatic score, average manual score, average interestingness rating, and average difficulty rating. The average automatic and manual scores are updated automatically based on the teacher’s grading of individual submissions. The evaluation data of both the class as a whole and individual students can be downloaded as a CSV file for the teacher to bring to class for further review and discussion.

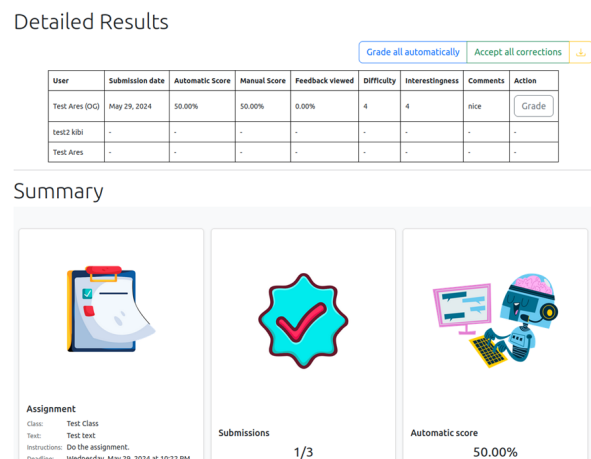


Figure 6: Assignment Grading Overview Page

**Student Evaluation** Students see only the manual evaluations confirmed by the teacher during the grading process. It is important to note that the evaluation display is only accessible to students once the teacher has entered the manual evaluation. Each answer is accompanied by different colors and icons to indicate binary feedback (correct/incorrect) (see Figure 7). Under the binary feedback icon, a chat button icon is available, which students can click to open or close the teacher’s feedback for each response. The system tracks which feedback has been viewed by the students and informs teachers about which students have

read which feedback, providing insight into student engagement and enabling more targeted support.

1. When did the last passenger pigeons die?

Septembre 1, 1914

2. Where did the last known Tasmanian tiger die?

In a cave

Meaning feedback: Nice try! But that's not where the last Tasmanian tiger died. Check the part that talks about the last known Tasmanian tiger and where it was kept.

Form feedback: Your sentence is grammatically correct. Well done!

Show answer

Figure 7: Feedback for Students

## 5 Conclusion and Outlook

Grounded in theories of text comprehension in SLA, and leveraging the affordances of language and AI technology, we present ARES, a web-based language learning system designed to support L2 RC of young EFL/ESL learners with teachers in the loop. The system provides on-demand help functions, such as glossing of vocabulary and language means, allowing students to interactively engage with texts, as well as EF on RC questions. These features not only aid students in understanding English reading texts but also alleviate teachers' workloads by automating time-consuming tasks such as question generation and evaluation. Furthermore, ARES facilitates direct interaction between students and teachers outside the classroom, enabling flexible assignment and feedback processes.

We acknowledge certain limitations in our system. First, there are challenges regarding the classification accuracy of language means (see Section 3.4). To tackle this challenge, a member of our research team is conducting a study to assess the system's classification accuracy by comparing the results of our automatic classification with labels provided by human annotators. Second, it is important to note that LLMs still lack the same level of understanding and context awareness as humans (Ray, 2023). Although they can perform a variety of tasks within seconds, LLMs struggle due to tendencies toward hallucination (Nye et al., 2023). However, this challenge is precisely why we designed the system to involve teachers

in the process, ensuring they confirm outputs before students see them, rather than relying solely on raw LLM-generated results. Although teachers might occasionally miss inaccuracies produced by the LLM, the system still significantly reduces their workload, allowing them to focus more on communicative activities in the classroom. Nevertheless, we are currently working on investigating the feasibility of leveraging the LLM to generate short answer questions and feedback. Using a human-authored evaluation method, we are investigating the linguistic and pedagogical quality of these LLM-generated outputs. For the evaluation criteria of the questions, we will employ a nine hierarchical criteria rubric (e.g., *Understandable, Grammatical, Answerable, Clear*) used in previous studies (Horbach et al., 2020; Moore et al., 2022; Steuer et al., 2021), which has been shown to be comprehensive, easy to interpret, and includes the pedagogical aspects of a question (Moore et al., 2022). For the evaluation criteria of the feedback, we will employ a four criteria rubric (*Readily applicable, Readability, Relational, Specificity*) that is formulated based on previous work on the human-authored evaluation of the quality of machine-generated feedback (Jia et al., 2021; Liang et al., 2024; Pinger et al., 2018; van der Lee et al., 2021).

Since the first version of the system is deployed, a study investigating teachers' and students' perceptions of the system is currently taking place in two intact English classes at secondary schools in southwest Germany with the purpose of evaluating the system's usability and students' interaction with the system. Over a four-week period, students will read two texts weekly as part of their homework assigned by teachers. System perceptions will be assessed through a self-report questionnaire of comprehensive evaluation of technology adapted from Lai et al. (2022). In addition to the survey data, log data will be analyzed to explore the learning behavior within the context of real-world ICALL system use.

ARES is currently available under <https://ares.kibi.group>.

## Acknowledgements

This project is funded by the German Ministry of Education and Science (BMBF) under the funding number 01IS22076.

## References

- Haiyang Ai. 2017. Providing graduated corrective feedback in an intelligent computer-assisted language learning environment. *ReCALL*, 29(3):313–334.
- Charles J. Alderson. 1984. Reading in a foreign language: A reading problem or a language problem. In Charles J. Alderson and Alexander H. Urquhart, editors, *Reading in a foreign language*, pages 1–25. Longman, London.
- Luiz A. Amaral and Detmar Meurers. 2011. On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, 23:4–24.
- Robert L. Bangert-Drowns, Chen-Lin C. Kulik, James A. Kulik, and MaryTeresa Morgan. 1991. The instructional effect of feedback in test-like events. *Review of Educational Research*, 61(2):213–238.
- Melissa A. Bowles. 2004. L2 glossing: To CALL or not to CALL. *Hispania*, 87(3):541–552.
- Andy Bown. 2017. Elaborative feedback to enhance online second language reading comprehension. *English Language Teaching*, 10(12):164–171.
- Andy Bown. 2018. Supporting online L2 academic reading comprehension with computer-mediated synchronous discussion and elaborative feedback. *The Reading Matrix*, 18(1):41–63.
- Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, , and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84.
- Carol A. Chapelle. 2003. *English Language Learning and Technology: Lectures on Applied Linguistics in the Age of Information and Communication Technology*, volume 7. John Benjamins Publishing, Amsterdam.
- Huilin Chen and Huan Mei. 2024. How vocabulary knowledge and grammar knowledge influence L2 reading comprehension: a finer-grained perspective. *European Journal of Psychology of Education*, pages 1–23.
- Inn-Chull Choi. 2016. Efficacy of an ICALL tutoring system and process-oriented corrective feedback. *Computer Assisted Language Learning*, 29(2):334–364.
- Yunjeong Choi and Dongbo Zhang. 2021. The relative role of vocabulary and grammatical knowledge in L2 reading comprehension: A systematic review of literature. *International Review of Applied Linguistics in Language Teaching*, 59(1):1–30.
- Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion Volume*. Cambridge University Press, Cambridge.
- Mienke Droop and Ludo Verhoeven. 2003. Language proficiency and reading ability in first- and second-language learners. *Reading Research Quarterly*, 38(1):78–103.
- Elizabeth Farley-Ripple, Henry May, Allison Karpyn, Katherine Tilley, and Kalyn McDonough. 2018. Rethinking connections between research and practice in education: A conceptual framework. *Educational Researcher*, 47(4):235–245.
- David Ferrucci and Adam Lally. 2004. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3–4):327–348.
- Bridgid Finn, Ruthann Thomas, and Katherine A. Rawson. 2018. Learning more from feedback elaborating feedback with examples enhances concept learning. *Learning and Instruction*, 54:104–113.
- Carlton J. Fong, Erika A. Patall, Ariana C. Vasquez, and Sandra Stautberg. 2019. A meta-analysis of negative feedback on intrinsic motivation. *Educational Psychology Review*, 31(1):121–162.
- William Grabe. 2014. Key issues in L2 reading development. In *Proceedings of the 4th CELC Symposium for English Language Teachers-Selected Papers*, pages 8–18.
- John Hattie and Helen Timperley. 2007. The power of feedback. *Review of Educational Research*, 77(1):81–112.
- Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, Maxine Eskenazi, Alan Juffs, and Lois Wilson. 2010. Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education*, 20(1):73–98.
- Andrea Horbach, Itziar Aldabe, Marie Bexte, Oier Lopez de Lacalle, and Montse Maritxalar. 2020. Linguistic appropriateness and pedagogic usefulness of reading comprehension questions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1753–1762.
- Qinjin Jia, Jialin Cui, Yunkai Xiao, Chengyuan Liu, Parvez Rashid, and Edward F Gehringer. 2021. All-in-one: Multi-task learning BERT models for evaluating peer assessments. *arXiv preprint arXiv:2110.03895*. <https://arxiv.org/pdf/2110.03895>.
- Jookyoung Jung. 2009. Second language reading and the role of grammar. *Working Papers in TESOL and Applied Linguistics*, 9(2):29–48.

- Stephen Kemmis and Robin McTaggart. 1988. *The Action Research Planner*, 3rd edition. Deakin University Press, Geelong.
- Walter Kintsch. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2):163–182.
- Walter Kintsch and Teun A. van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review*, 85(5):363–394.
- Fabienne M. Van der Kleij, Remco C. W. Feskens, and Theo J. H. M. Eggen. 2015. Effects of feedback in a computer-based learning environment on students’ learning outcomes: A meta-analysis. *Review of Educational Research*, 85(4):475–511.
- Keiko Koda. 2007. Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning*, 57:1–44.
- Jennifer W. M. Lai, John De Nobile, Matt Bower, and Yvonne Breyer. 2022. Comprehensive evaluation of the use of technology in education – validation with a cohort of global open online learners. *Education and Information Technologies*, 27:9877–9911.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Kraemer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.
- Hansol Lee and Jang Ho Lee. 2015. The effects of electronic glossing types on foreign language vocabulary learning: Different types of format and glossary information. *Asia-Pacific Education Researcher*, 24(4):591–601.
- Zhiping Liang, Lele Sha, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2024. Towards the automated generation of readily applicable personalised feedback in education. In *International Conference on Artificial Intelligence in Education*, pages 75–88.
- Chen-Chung Liu, Pin-Ching Wang, and Shu-Ju D. Tai. 2016. An analysis of student engagement patterns in language learning facilitated by Web 2.0 technologies. *ReCALL*, 28(2):104–122.
- Alison Mackey. 2006. Feedback, noticing and instructed second language learning. *Applied Linguistics*, 27(3):405–430.
- Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.
- Janice McAlpine and Johanne Myles. 2003. Capturing phraseology in an online dictionary for advanced users of English as a second language: a response to user needs. *System*, 31(1):71–84.
- Rosamond Mitchell, Florence Myles, and Emma Marsden. 2013. *Second language learning theories*, 3rd edition. Routledge, London.
- Steven Moore, Huy A. Nguyen, Norman Bier, Tanvi Domadia, and John Stamper. 2022. Assessing the quality of student-generated short answer questions using GPT-3. In *Proceedings of the 17th European Conference on Technology Enhanced Learning, EC-TEL 2022*, pages 243–257.
- Philip Murphy. 2007. Reading comprehension exercises online: The effects of feedback, proficiency and interaction. *Language Learning and Technology*, 11(3):107–129.
- Philip Murphy. 2010. Web-based collaborative reading exercises for learners in remote locations: The effects of computer-mediated feedback and interaction via computer-mediated communication. *ReCALL*, 22(2):112–134.
- Benjamin D. Nye, Dillon Mee, and Mark G. Core. 2023. Generative large language models for dialog-based tutoring: An early consideration of opportunities and concerns. In *Proceedings of the Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation 2023 co-located with 24th International Conference on Artificial Intelligence in Education (AIED 2023)*, pages 78–88.
- Andrew Oberg and Paul Daniels. 2013. Analysis of the effect a student-centred mobile learning instructional method has on language acquisition. *Computer Assisted Language Learning*, 26(2):177–196.
- Anne O’Keeffe and Geraldine Mark. 2017. The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, 22(4):457–489.
- Petra Pinger, Katrin Rakoczy, Michael Besser, and Eckhard Klieme. 2018. Implementation of formative assessment—effects of quality of programme delivery on students’ mathematics achievement and interest. *Assessment in Education: Principles, Policy & Practice*, 25(2):160–182.
- Martí Quixal, Björn Rudzewitz, Elizabeth Bear, and Detmar Meurers. 2021. Automatic annotation of curricular language targets to enrich activity models and support both pedagogy and adaptive systems. In *Proceedings of the 10th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2021)*, pages 15–27.
- Robert N. Rapoport. 1970. Three dilemmas of action research. *Human Relations*, 23(6):499–513.
- Partha Pratim Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154.



Steve Ridout. 2013. [Readlang](#). *The EUROCALL Review*, 21(2):64–68.

Mahnaz Saeidi and Mahsa Yusef. 2012. The effect of computer-assisted language learning on reading comprehension in an Iranian EFL context. In *EUROCALL Conference Proceedings: Using, Learning, Knowing*, pages 259–263.

Yasuyo Sawaki. 2001. Comparability of conventional and computerized tests of reading in a second language. *Language Learning and Technology*, 5(2):38–59.

Valerie J. Shute. 2008. [Focus on formative feedback](#). *Review of Educational Research*, 78(1):153–189.

Tim Steuer, Leonard Bongard, Jan Uhlig, and Gianluca Zimmer. 2021. [On the linguistic and pedagogical quality of automatic question generation via neural machine translation](#). In *Proceedings of the 16th European Conference on Technology Enhanced Learning, EC-TEL 2021*, pages 289–294.

Elise K. Swart, Thijs M.J. Nielen, and Maria T. Sikkema de Jong. 2022. [Does feedback targeting text comprehension trigger the use of reading strategies or changes in readers’ attitudes? a meta-analysis](#). *Journal of Research in Reading*, 45(2):171–188.

Alan M. Taylor. 2009. [CALL-based versus paper-based glosses: Is there a difference in reading comprehension?](#) *CALICO Journal*, 27(1):147–160.

Ludo Verhoeven. 2011. [Second language reading acquisition](#). In Michael L. Kamil, P. David Pearson, Elizabeth Birr Moje, and Peter Afflerbach, editors, *Handbook of reading research*, volume 4, pages 661–683. Routledge, New York, NY.

Giulia Vettori, Lidia Casado-Ledesma, S. Tesone, and Christian Tarchi. 2023. [Key language, cognitive and higher-order skills for L2 reading comprehension of expository texts in English as foreign language students: a systematic review](#). *Reading and Writing: An Interdisciplinary Journal*.

Benedikt Wisniewski, Klaus Zierer, and John Hattie. 2020. [The power of feedback revisited: A meta-analysis of educational feedback research](#). *Frontiers in Psychology*, 10:3078.

Fulya Çolak and Ufuk Balaman. 2022. [The use of online dictionaries in video-mediated L2 interactions for the social accomplishment of virtual exchange tasks](#). *System*, 106:102772.

## Appendix. Prompts for the LLM

### Prompt 1. Question Generation

The query template for asking the LLM to provide two types of reading comprehension questions (factual and inferential). The placeholder

fields with angle brackets are to be substituted for the actual data in each query.

```
You are an EFL teacher who teaches English to non-native school students between 10-18 years old. Provide simple one-sentence short-answer reading comprehension questions based on the given text to these EFL learners. Do not use too difficult words. Literal comprehension refers to an understanding of the straightforward meaning of the text, such as facts, vocabulary, dates, times, and locations. Questions of literal comprehension can be answered directly and explicitly from the text with a few words. Inferential questions ask students to infer information from the passage where the answer is not directly stated in the text. The students have to use their background knowledge to make a logical assumption about ideas in the passage and normally require a full sentence to answer, not a few words.
- text: <reading_text>
- number of factual questions: <number>
- number of inferential questions: <number>
Please provide the questions in JSON format as follows:
{
  "questions": [
    {
      "type": <factual_or_inferential>,
      "prompt": "<question>",
      "answer": "<correct_answer>"
    },
  ],
};
```

### Prompt 2. Feedback Generation

The query template for asking the LLM to provide feedback and hint to a student’s response. The placeholder fields with angle brackets are to be substituted for the actual data in each query.

```
For each question, evaluate each EFL student’s answer as follows using simple language as the students are non-native and kids:
1. Determine if the answer is correct or incorrect based on the content only.
2. Provide binary feedback for content ("Correct"/"Incorrect").
3. Offer short, kind, and friendly feedback on the content of the answers.
4. Give a concrete hint on the content explaining why the response was correct or incorrect, allowing the student to review part of the text, without revealing the target answer. When correct, do NOT provide hint.
- text: <reading_text>
Provide evaluation in JSON format using the match of answer id:
{
  "evaluation": [
    {
      "question": <question>,
      "answer_id": <answer_id>,
      "answer_text": <student’s_answer>,
      "solution": <correct_answer>,
      "binary": <binary_feedback>,
      "feedback": <content_feedback>,
      "hint": <content_hint>
    },
  ],
};
```

# Developing a Web-Based Intelligent Language Assessment Platform Powered by Natural Language Processing Technologies

Sarah Löber<sup>1,2</sup>, Björn Rudzewitz<sup>1,2</sup>, Daniela Verratti Souto<sup>1</sup>, Luisa Ribeiro-Flucht<sup>1,2</sup>,  
Xiaobin Chen<sup>1,2</sup>

<sup>1</sup>Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Germany

<sup>2</sup>LEAD Graduate School and Research Network, University of Tübingen, Germany

{sarah.loeber, bjoern.rudzewitz, luisa.ribeiro-flucht,  
xiaobin.chen}@uni-tuebingen.de,  
daniela.verratti-souto@student.uni-tuebingen.de

## Abstract

We introduce ILAP, an intelligent language assessment platform and reusable module that streamlines the creation, administration and scoring of language proficiency tests supported by Natural Language Processing (NLP) technologies. As a first implementation, we realized an automatic pipeline for the Elicited Imitation Test (EIT), a popular test format that has been widely adopted in language learning research for general proficiency and formative assessments. The platform can be extended to other test formats and assessment types. ILAP is a valuable tool for standardizing data collection in Second Language Acquisition (SLA) and Intelligent Computer Assisted Language Learning (ICALL) research as well as serving as an application for classroom assessment. In this paper, we present the design of the system and a preliminary evaluation of Large Language Models (LLMs) for generating language errors for EIT items.

## 1 Introduction

Language assessment is a way for teachers and researchers to understand the current level of a learner’s knowledge so that they can adjust their teaching or understand how language develops in the learner (Révész and Brunfaut, 2020; McNamara, 2000). Traditionally, language assessment has been done with tests of various formats, such as written tests with multiple choice, essay writing items, or spoken interviews. These tests are typically created manually, administered and graded by language teachers or researchers in school or

lab settings, except for large-scale standardized tests such as the Test of English as a Foreign Language (TOEFL) or the International English Language Testing System (IELTS), which also include automatic forms of assessment (Evanini et al., 2015). The complexity of language assessment and the labor-intensiveness of language test creation, administration and grading are a major challenge for teachers and Second Language Acquisition (SLA) researchers, especially when the need to assess the students repeatedly and frequently arises. We therefore address these issues by creating a comprehensive language assessment system incorporating NLP. These technologies accelerate test implementation and scoring, making language testing feasible for a broader audience.

In the present paper, we demonstrate ILAP (Intelligent Language Assessment Platform), which is designed to facilitate the creation, administration, scoring, and reporting of results of language tests supported by technologies such as Automatic Speech Recognition (ASR), and generative AI technologies, in particular Text-to-speech (TTS) and Large Language Models (LLMs). The system features easy test creation with NLP leveraged item construction, convenient web-based test deployment, and automatic test response scoring and reporting. As a first instance, the system’s implementation supports the Elicited Imitation Test (EIT) format, a popular test format that has been found to be effective in evaluating learners’ general proficiency and to tap into their implicit language knowledge. An EIT targeting specific linguistic constructs can potentially also be used as a formative assessment tool to facilitate adaptive teaching.

In the following section, we will first justify the

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

choice of EIT as a valuable test format to be implemented in an intelligent language assessment system by reviewing the research behind the test format. We will then specify how ILAP supports the whole procedure of EIT-based assessment and the above-mentioned technologies used in the system. Furthermore, we provide a preliminary evaluation of our automatic scoring and the use of generative AI for generating ungrammatical test items. The paper concludes with an outlook of the project and future work.

## 1.1 The Elicited Imitation Test

In SLA research, numerous types of tests have been used to characterize learners' language proficiency, implicit or explicit knowledge of a language or their cognitive abilities. EIT, a popular test format among SLA researchers, is a sentence repetition task that requires the test taker to listen to the recordings of some sentences one at a time and then repeat the sentence they have just heard. Distractor questions (e.g. simple arithmetic calculations or judgement of the truthfulness of the sentence) are often asked between the audio playback and the repetition to prevent the test taker from relying on their phonological memory but rather require them to make use of their language system based on the meaning of the sentence. EITs have been used in a variety of ways, notably as a measure of implicit knowledge or general language proficiency (Ellis, 2005; Yan et al., 2016). Several studies corroborate the high validity of the test (Yan et al., 2016; Kostromitina and Plonsky, 2022), highlighting its efficacy as well as reliability. Furthermore, EITs show potential to serve as a placement test in language education (Yan et al., 2020) and as a teacher tool to assess second language (L2) learners' oral production skills in language classes (Campfield, 2017). Better still, research has found that it is an effective assessment format for various languages (Wu et al., 2023).

So far, the EIT has been administered in different formats, with different design implementations. For example, researchers have incorporated ungrammatical sentences (Erlam, 2006). Carefully created ungrammatical sentences are often used in EITs to test learners' specific grammatical knowledge (Spada et al., 2015). That is, whether a test taker can correct specific grammar errors in the repetition stage is an indicator of their implicit knowledge of the grammatical constructs. Scoring

methods also vary: in some tests, items are scored on a binary basis, for instance, correct or incorrect for the use of the target structure only (Erlam, 2006), while others use a more fine-grained 5-point scale (Ortega et al., 2002) or even a percentage scale (Lonsdale and Christensen, 2011). Due to the different design implementations of EITs used in research, it is challenging to compare proficiency measures across studies. Therefore, there have been calls to enhance standardization of the tests (Isbell and Son, 2022; Kostromitina and Plonsky, 2022).

EIT items can also be designed to target specific grammar constructs that are the learning targets at different L2 developmental stages. For example, *third-person singular -s* or *mass/count nouns* are popular target constructs in previous L2 English studies (Kim and Godfroid, 2023). This makes the test an effective tool for formative assessment, but also poses a challenge to the test creator as they will need to not only find and write sentences with the target constructs, but also consider the sentence length, lexical frequency and other grammatical constructs in the sentence prompts. All of these factors have been found to affect the difficulty of test items as well as the validity of an EIT (Yan et al., 2016; Hendrickson et al., 2010). Users of EITs also face challenges from test administration and scoring, which traditionally requires the presence of a teacher or researcher in the classroom or lab to control the test procedure and to listen to the test responses for scoring. Hence, it is time-consuming, labor-intensive, and therefore difficult to scale.

We aim to address these issues by introducing ILAP, a web-based language assessment platform, where assessments of language proficiency can be created, administered and scored automatically. The first type of test integrated on the platform is an EIT pipeline.

## 1.2 Related work

Automating a language test requires automating several individual components involved in the testing process. While, to our knowledge, there is no fully automatic pipeline for the EIT developed yet that allows full flexibility, there have been studies on automating individual components of the test, such as item creation (Christensen et al., 2010) or scoring (Graham et al., 2008; Isbell et al., 2023). The findings of these studies show promising re-

sults for the feasibility of automated EITs.

In the case of automatic scoring, studies proposing solutions have focused on transcribing test takers' responses with automatic speech recognition (ASR) and implementing rules for scoring the transcriptions. For example, utilizing ASR and transcription scoring metrics based on string edit distance, [Isbell et al. \(2023\)](#) were able to achieve high correlations with human scoring ( $r > .90$ ) across all items on the Korean EIT. Likewise, [Graham et al. \(2008\)](#) reported high correlations for a method using ASR and binary scoring on the syllable level.

Pertaining to the further automation of EITs, [Christensen et al. \(2010\)](#) utilized a language corpus for the automatic and flexible selection of elicited imitation test items with their item selection tool. The automatically selected EIT showed higher correlations with the speaking language achievement test (SLAT) than previous EITs.

The EIT is often administered in a lab, as part of data collection for studies. An alternative would be to administer the test online, allowing for more flexibility, easier processing of the responses and potentially reaching more participants. Some studies have administered the EIT in this way, with web-based and lab-based EITs showing no significant difference in their validity ([Kim et al., 2024](#)). However, [Kim et al. \(2024\)](#) found weaker correlations, albeit non-significantly, for the web-based EIT and TOEFL scores than for the lab-based EIT and TOEFL scores when taking only ungrammatical items into account. According to the authors, this could result from a lack of immediate feedback in the web-based EIT. Informed by and building on previous efforts to automate EIT creation, administration, and scoring, we implement the process in a newly developed intelligent language assessment system that utilizes latest AI technologies. The next section provides more details.

## 2 System overview

ILAP is a web-based application that is mobile-friendly and compatible with most devices. The back end is coded in Java, while the web front end utilizes JavaScript and the Bootstrap framework. There are two interfaces offered: a test creator interface as well as a test taker view, both of which require user profiles and accounts with different roles. In the following, we describe the test cre-

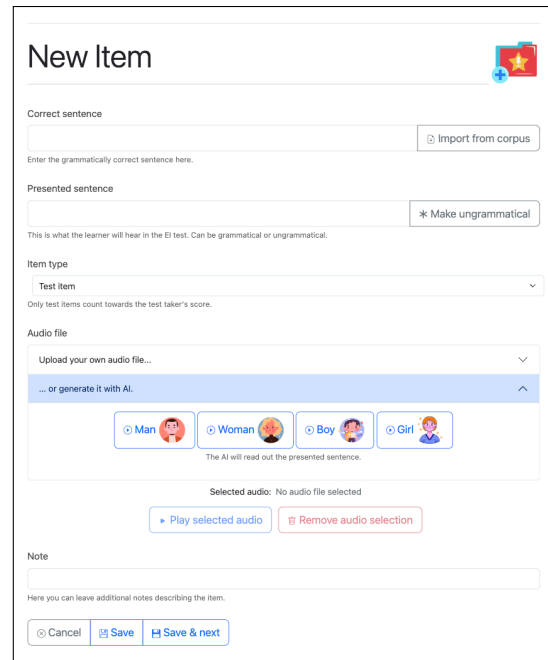


Figure 1: Interface for creating new test items

ation, administration and scoring procedure with ILAP.

### 2.1 Test creation

Test creators start by creating a test collection. This automatically generates a unique and random 4-character access code that the test creator can give to the participants to take the test. In the next step, tests can be added to the test collection. The choice for letting the user add different tests into the test collection was made with the future integration of new test types in mind. This will allow the integration of several separate test components into one test collection, e.g. an EIT followed by a reading comprehension test. When adding a new test, users can specify the name, description and visibility of the test. Tests with visibility set to public can be shared among test creators. Afterwards, users can add the test type. In the first step, we implemented an elicited imitation test type. Within the created test, users can then manage instructions, items and settings or preview their test. Figures 5, 6, and 7 in Appendix A.1 show the test creation process.

**Instructions** The instruction interface allows the user to add instructions, including their title and text. Furthermore, users specify at what point during the test an instruction is shown, e.g. before practice items or before each item. Test creators can add any number of instructions for the test,

with each instruction appearing on a separate page in the test-taker interface. Figure 8 in Appendix A.2 shows a screenshot of this interface.

**Items** Figure 1 shows the item interface from the test creator perspective, which supports adding grammatical as well as ungrammatical items. Test creators can add their own sentences or choose sentences from the provided sentence corpus. For the latter case, we annotated about 95.000 extracted sentences from the Spotlight corpus (Weiss et al., 2021) with constructs from the English Grammar Profile (EGP, O’Keeffe and Mark, 2017) using an in-house EGP annotator. Users can search, select and import sentences from the corpus, filter by grammatical construct, and also edit the sentences for their items.

The interface supports both a manual and an automatic creation of ungrammatical variants of sentences. We implemented a component incorporating GPT-4o through the OpenAI API (OpenAI, 2024a) to automatically produce ungrammatical variants of the sentence, i.e. simulating the output of mal-rules (e.g. Sleeman, 1985) on the correct sentence. The generated ungrammatical sentence is based on the user input. Users can enter the correct sentence and select the “Make ungrammatical” button, upon which the generated ungrammatical sentence is displayed in the “Sentence” field in the interface. A more elaborate evaluation on our choice for using GPT-4o for this functionality, including quantitative and qualitative human assessments on the error generation, is provided in Section 3.

Furthermore, an item can be classified as practice or test item. Audio files for items can either be uploaded or automatically generated. For this functionality, we are using the text-to-speech service from Amazon Web Services (AWS)<sup>1</sup>. Lastly, a note can be added to describe an item.

**Settings** Test creators can control all settings related to a test by overriding the default settings of tests with their own values. For example, they can control the duration of the recording of responses by test takers, whether belief statement checks are shown after items are shown, and more. This interface is displayed in Figure 9 in Appendix A.3.

## 2.2 Test administration

Each test collection is created with the status “in editing”. As long as a test has this status, it cannot

<sup>1</sup><https://aws.amazon.com/>

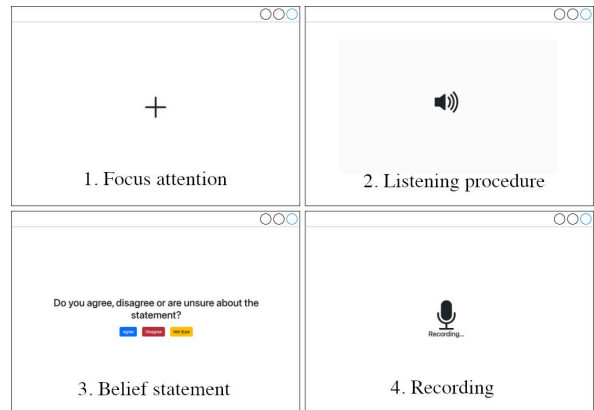


Figure 2: Test taker perspective of item procedure

be started by test takers. In case an access code for a test collection that is not released is entered, test takers get a warning message informing them that the test is not available yet. Test creators can control the release of a test by updating its status to “released”. At this point, the test becomes available and cannot be edited anymore in order to avoid problems related to test taker data referring to an outdated version of the test, ensuring a valid data collection process.

Test takers can access the released test via the test taker interface by entering the provided access code. The item procedure, using the default settings, is shown in Figure 2.

## 2.3 Scoring and results

To allow for full flexibility for the test creator, completed EITs can be scored manually as well as automatically. When a test has been taken by the user, test creators can access the result overview via the corresponding test in the “My tests” interface. In this overview, test results are grouped by the progress of test takers. Tests which have been started, but not yet finished by the user are also shown. Responses for finished tests can either be scored automatically (all at once or item by item) or manually in the performance overview, where the test takers’ audio transcriptions and the string edit distance measures to the correct sentence, converted to a percentage, are displayed.

For the automatic scoring algorithm, following the work of Isbell et al. (2023), we use the transcription of the test response and string edit distance measures to calculate the test score. The recorded audio files of the test takers are automatically transcribed and the transcription is compared to the correct sentence for each item specified by

the test creator. For transcription of test-taker responses, we are using the `Whisper-large-v2` model through the OpenAI transcriptions API (OpenAI, 2024b). Our choice of Whisper for response transcription is based on previous research (Bear et al., 2023) showing that Whisper has the lowest word error rate (WER) when compared to other commercial ASR providers on ungrammatical and grammatical sentences from L2 speakers.

For string edit distance comparison, the system first normalizes the transcription string as well as the target string by converting the characters to lowercase and removing the non-word characters as well as whitespace characters. We decided on this process of normalization after noticing that the ASR would occasionally add punctuation characters, for example adding a question mark when raising the voice at the end of a sentence. After this process, the mean of three string distance measures is computed: Levenshtein, Jaro-Winkler and Jaccard distance. We are using the mean of these measures in order to retain the different measure characteristics while also making the result more accessible by offering only one score to the test creator. For making these scores more intuitive, the mean of the three measures is converted to a percentage on the item-level as well as the test-level, ranging from 0 to 100. Figure 3 shows the scoring interface on the test-level. The system offers an additional field on the item-level for a manual score in case test creators want to apply their own scoring metric, e.g. a school grading system. A screenshot of the scoring interface on the item-level can be found in Figure 10 in Appendix A.4.

## 2.4 Preliminary testing of scoring functionality

For preliminary testing of our scoring implementation, we manually scored 22 EITs taken with our system. Scoring each item on a scale of 0-4, we followed the established scoring scheme of Ortega et al. (2002), with the sum of all item scores as the total EIT score. The EITs consisted of 24 items, resulting in a maximum total score of 96 for the manual scoring. We then correlated the total manual EIT scores with the total automatic scores of these 22 tests in ILAP. We achieved a correlation of  $r = .95$  across items, which is in line with previous studies employing this approach (Isbell et al., 2023). Figure 4 shows the correlation of manual and automatic scores. The x-axis shows

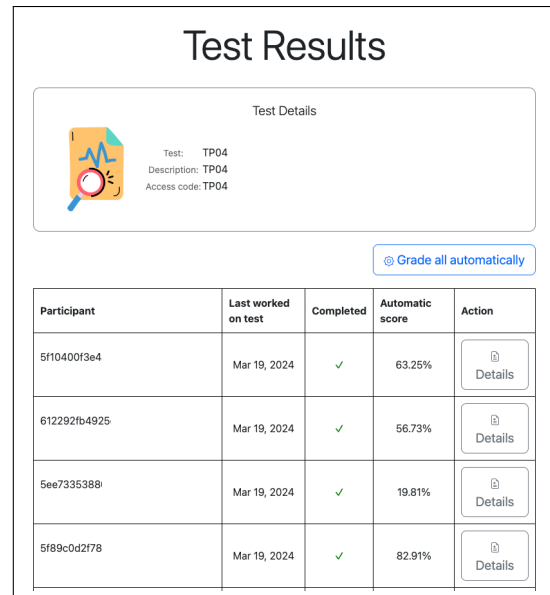


Figure 3: Interface for scoring responses on the test level

the manual scores, ranging from 0-96 points and the y-axis shows the similarity score of the target answer and the learner answer in percentages. The line shows the fitted linear model with 95% confidence interval.

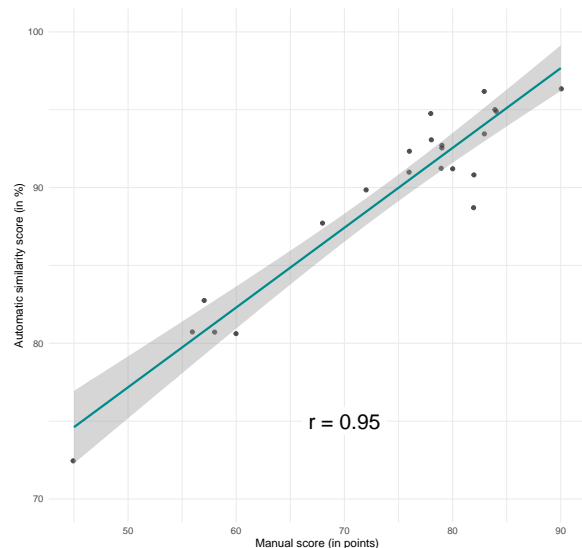


Figure 4: Correlation of manual EIT scores and automatic similarity scores. The grey area represents the 95% confidence interval of the fitted linear model.

## 3 Preliminary evaluation of LLMs for ungrammatical item generation

In order to evaluate whether the changes introduced by LLMs can be considered realistic errors, i.e. errors that are plausible to expect from learn-

ers, we conducted a preliminary evaluation for our ungrammatical sentence generation functionality in ILAP. We compared GPT-3.5-turbo, GPT-4, GPT-4o and Claude 3 Haiku on their performance of the generation of ungrammatical sentences. We used sixteen grammatical versions of existing EIT items from previous tests (Erlam, 2006; Spada et al., 2015; Godfroid and Kim, 2021) and prompted the models by providing examples from previously used ungrammatical EIT items and specifying the limit for the amount of changes to be made in the sentence. Our evaluation focused on quantitative as well as qualitative aspects.

### 3.1 Quantitative evaluation

For the quantitative evaluation of the plausibility of errors, there were no error-annotated learner corpora available containing specifically the test items we used. Therefore, we conducted our evaluation with the output of a mal-rule-based, generative approach based on actual learner error patterns. Mal-rules are patterns to parse or generate learner language that model specific misconceptions or errors (Sleeman, 1985).

An example of extensive mal-rule usage is the successful FeedBook system (Meurers et al., 2019), which is an ICALL system for English as a second language that incorporates an automatic feedback generation approach capable of generating a wide range of possible errors based on a well-formed target answer (Rudzewitz et al., 2018). The feedback generation component works by iteratively applying mal-rules derived from a corpus of actual learner errors to an input string and thereby automatically generating a wide range of ill-formed variants of input string along with error diagnoses (Ziai et al., 2018). Those variants can then be aligned with answers produced by learners, and if there is a match, the diagnosis associated with a generated variant is used to display a scaffolding feedback message to the learner.

Since the mal-rules included in FeedBook represent generalizations of actually observed learner errors, we employed the overlap between the output of the FeedBook feedback generation and the output of the LLMs as a criterion to assess the plausibility of the errors generated by the LLMs. To this end, we let the FeedBook feedback generation component generate all possible variants based on ten experimental test items, and com-

puted the degree of overlap between the sentences from this approach with the sentences generated by the LLMs. Table 1 shows the results.

Model	FeedBook Overlap
GPT-3.5-turbo	27.3
GPT-4	27.3
GPT-4o	81.8
Claude 3 Haiku	63.6

Table 1: Overlap (in percentages) between the output of different LLMs with the output of the FeedBook mal-rule-based generative approach

Since not all constructs in all sentences were covered by the ICALL system’s generative approach due to the fact that the FeedBook was designed for a specific grade, we restricted the comparison to those sentences where the FeedBook generated alternative variants, which were ten out of sixteen experimental test items. An example of generated errors by the four LLMs and FeedBook can be found in Table 2.

The results show that GPT-4o produced the highest overlap with the output from the mal-rule generation approach.

### 3.2 Human evaluation

We also conducted human evaluation with the sentences generated by the four models. We asked human raters to evaluate the ungrammatical sentences on 5-point Likert scales on three different dimensions, namely

- *Naturalness of Error (NoE)*: this sentence contains an error that is characteristic of an error produced by language learners
- *Retention of meaning*: this sentence retains the meaning of the correct sentence
- *Adherence to prompt*: the output adheres to the prompt given to the LLM

Seven human evaluators, all experts in linguistics with teaching experience, rated the same 16 sentences generated by each model without knowing which model the sentences were from, resulting in a total of 64 sentences per evaluator. Evaluators indicated their agreement to the dimensions above on a 5-point Likert scale ranging from 1 - Strongly disagree to 5 - Strongly agree. Table 3 shows the results of the human evaluation <sup>2</sup>.

<sup>2</sup>The data and analysis scripts are available under <https://osf.io/tjn4v/>

Model	Generated error sentence(s)
GPT-3.5-turbo	Family names is often changed after marriage.
GPT-4	Family names are often changed after marriage it.
GPT-4o	Family name are often changed after marriage.
Claude 3 Haiku	Family names are often change after marriage.
FeedBook	Family names are often change after marriage. Family names often changed after marriage. Family names often changes after marriage. Family names is often changed after marriage. Family names be often changed after marriage. ...

Table 2: Examples of LLM-generated errors and FeedBook generated error variants on the input sentence “Family names are often changed after marriage.”. Input sentence taken from Spada et al. (2015).

We performed additional statistical analyses on the evaluation data in R. Given the small sample, we conducted Shapiro-Wilk tests to assess the normality of the distribution for each dimension. The Shapiro-Wilk tests showed significant deviations from normality, confirming our data were not normally distributed. Therefore, we opted for non-parametric tests for further analysis. A Kruskal-Wallis test showed a significant difference of means on Naturalness of errors ( $H(3) = 14, p = 0.002$ ) and Retention ( $H(3) = 10, p = 0.02$ ). Pairwise Mann-Whitney U comparisons with Bonferroni corrections were conducted to determine which specific models differed. The results revealed that GPT-4o significantly outperformed GPT-3.5-turbo on Naturalness of errors ( $p = 0.01$ ). Furthermore, GPT-4o significantly outperformed Claude 3 Haiku on Retention ( $p = 0.03$ ).

Model	NoE	Retention	Adherence
GPT-3.5-turbo	3.41	4.42	4.55
GPT-4	3.79	<b>4.71</b>	4.79
GPT-4o	<b>4.09</b>	4.66	<b>4.83</b>
Claude 3 Haiku	3.88	4.38	4.71

Table 3: Mean ratings of LLM generated ungrammatical sentences on three dimensions.

### 3.3 Results and discussion

Based on the results of the preliminary evaluation, we decided to use GPT-4o for generating ungrammatical variants of sentences in our system. Our quantitative evaluation shows the high overlap between GPT-4o output and the mal-rule-based approach, suggesting that GPT-4o generates plausible learner errors. The human evaluation strength-

ened this finding, with GPT-4o achieving the highest ratings in naturalness of errors as well as adherence (although not significantly). GPT-4 also seemed to perform well in the human evaluation, achieving the highest ratings in retention of meaning. However, the quantitative evaluation showed a low overlap between the mal-rule-based generative approach and the sentences generated by GPT-4, which might be due to the ICALL system’s limited scope in producing more advanced learner errors, since it only covers specific grammatical constructs. For example, the sentence “Birthday cards have been emailed since hundreds of years.”, with the same error being generated by both GPT-4 and GPT-3.5-turbo, had no matching variant in the FeedBook output, but was rated plausibly by humans. This is possibly due to the since/for construct not being included in the ICALL system.

It is also noteworthy that out of all dimensions, all models score lowest on the NoE dimension, meaning that the errors generated by the models were not rated as highly natural or being highly characteristic of language learners by the human evaluators. This observation could indicate that commercial LLMs might not excel at generating mal-formed language, but rather have been demonstrated to be highly effective for grammatical error correction (e.g. Katinskaia and Yangarber (2024)). Better results for error generation could be achieved with a model fine-tuned for this specific task (Bryant et al., 2023).

## 4 Limitations

There are some limitations to our system. First, the small amount of validation data demands cautiousness when making any claims about the effec-



tiveness of ILAP or our scoring functionality. For this reason, we are striving for wider deployment of the platform to collect data from different and larger groups, for example in schools or learning environments. Secondly, questions about the effect of the use of technology in the creation of the EIT items remain open. In the future, we plan to investigate to what extent technology can be used in creating language proficiency test items and the effect on test validity. Thirdly, there currently is a lack of test types in ILAP. We are working on implementing more test types, which would make the platform more versatile and adaptable to more use cases.

Additionally, the system could provide even more support in the item creation process, for instance an assessment of an item’s difficulty. This functionality is currently not integrated into the platform. Arguably, this would make it easier to create EIT items of varying difficulty. Future research could focus on the automatic item difficulty prediction of EIT items, where important progress has been made in the context of computer-adaptive testing (Settles et al., 2020).

As discussed in Section 3.3, we conducted a novel but preliminary evaluation for assessing the output of LLMs for error generation in both a quantitative and qualitative way. Our quantitative approach for evaluation is arguably not without flaws and might benefit from including a larger set of data as well as more diverse resources and approaches, such as an error-annotated corpus, in order to evaluate the generation of ungrammatical sentences. This would also enable further research to go beyond the scope of EIT items.

## 5 Conclusion and future work

We presented a system for automatic language assessment as well as data collection and computer-assisted scoring. We included a pipeline for elicited imitation tests, which can be used for both research and education. Furthermore, we described our preliminary evaluation of the integrated scoring functionality and presented an approach for the evaluation of LLMs for generating ungrammatical sentences. To the best of our knowledge, this type of evaluation has not been performed before. With the integration of the EIT we have made an important first step in enabling automatic language assessment and standardizing proficiency tests in SLA research use-

ful for teachers, researchers and test creators. The benefit of such a system can be of importance to other domains, such as ICALL. As Ruiz et al. (2023) stated, not all ICALL systems currently offer a built-in functionality for collecting test results for SLA research, leading the authors to emphasize the need for reusable modules. Since ILAP can potentially be integrated into other systems, it can be used to simplify the process of testing for ICALL systems.

In the future, we will expand the platform with a teacher dashboard view and implement more test types to make the system more relevant for usage in schools. Currently, we have started the deployment of the system for studies, including a study to test the effects of automatic speech synthesis on test validity and other studies on factors (e.g., speech rate) that might affect test performance and scoring, and, consequently, the reliability of EITs. As for our ungrammatical item generation analysis, we plan to build on and extend this analysis by increasing the sample size for the analysis to cover more types of learner errors. Furthermore, we intend to error-annotate the output of the LLMs according to annotation criteria for learner corpora in order to be able to compare the frequency of the generated error types with the frequency of the error type in learner corpora and use this information as an additional criterion for the plausibility of the errors and quality of the LLM output. Additionally, we plan to explore the effects of fine-tuning a large language model for this task specifically based on error-annotated learner corpora.

The ILAP system is currently available at <https://ilap.kibi.group>.

## Acknowledgements

We would like to thank the two anonymous NLP4CALL reviewers for their helpful feedback. This project is funded by the German Ministry of Education and Science (BMBF) under the funding number 01IS22076.

## References

Elizabeth Bear, Stephen Bodnar, and Xiaobin Chen. 2023. *Learner and linguistic factors in commercial ASR use for spoken language practice: A focus on form*. In *Proc. 9th Workshop on Speech and*

- Language Technology in Education (SLaTE)*, pages 161–165.
- Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. [Grammatical error correction: A survey of the state of the art](#). *Computational Linguistics*, 49(3):643–701.
- Dorota E. Campfield. 2017. [Lexical difficulty – using elicited imitation to study child L2](#). *Language Testing*, 34(2):197–221.
- Carl Christensen, Ross Hendrickson, and Deryle Lonsdale. 2010. [Principled construction of elicited imitation tests](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Rod Ellis. 2005. [Measuring implicit and explicit knowledge of a second language: A psychometric study](#). *Studies in Second Language Acquisition*, 27(2):141–172.
- Rosemary Erlam. 2006. [Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study](#). *Applied Linguistics*, 27(3):464–491.
- Keelan Evanini, Michael Heilman, Xinhao Wang, and Daniel Blanchard. 2015. Automated scoring for the TOEFL Junior® comprehensive writing and speaking test. *ETS Research Report Series*, 2015(1):1–11.
- Aline Godfroid and Kathy Minhye Kim. 2021. [The contribution of implicit-statistical learning aptitude to implicit second-language knowledge](#). *Studies in Second Language Acquisition*, 43(3):606–634.
- C. Ray Graham, Deryle Lonsdale, Casey Kennington, Aaron Johnson, and Jeremiah McGhee. 2008. [Elicited imitation as an oral proficiency measure with ASR scoring](#). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- Ross Hendrickson, Meghan Aitken, Jeremiah McGhee, and Aaron Johnson. 2010. [What makes an item difficult? A syntactic, lexical, and morphological study of elicited imitation test items](#). In *Selected Proceedings of the 2008 Second Language Research Forum*, pages 48–56. Cascadilla Proceedings Project Somerville, MA.
- Daniel R. Isbell, Kathy Minhye Kim, and Xiaobin Chen. 2023. [Exploring the potential of automated speech recognition for scoring the Korean Elicited Imitation Test](#). *Research Methods in Applied Linguistics*, 2(3):100076.
- Daniel R. Isbell and Young-A Son. 2022. [Measurement properties of a standardized elicited imitation test: An integrative data analysis](#). *Studies in Second Language Acquisition*, 44(3):859–885.
- Anisia Katinskaia and Roman Yangarber. 2024. [GPT-3.5 for grammatical error correction](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7831–7843, Torino, Italia. ELRA and ICCL.
- Kathy Minhye Kim and Aline Godfroid. 2023. [The interface of explicit and implicit second-language knowledge: A longitudinal study](#). *Bilingualism: Language and Cognition*, 26(4):709–723.
- Kathy Minhye Kim, Xiaoyi Liu, Daniel R Isbell, and Xiaobin Chen. 2024. [A comparison of lab-and web-based elicited imitation: Insights from explicit-implicit L2 grammar knowledge and L2 proficiency](#). *Studies in Second Language Acquisition*, pages 1–22.
- Maria Kostromitina and Luke Plonsky. 2022. [Elicited imitation tasks as a measure of L2 proficiency: a meta-analysis](#). *Studies in Second Language Acquisition*, 44(3):886–911.
- Deryle Lonsdale and Carl Christensen. 2011. [Automating the scoring of elicited imitation tests](#). In *Proc. Machine Learning in Speech and Language Processing (MLSPLP 2011)*, pages 16–20.
- Timothy Francis McNamara. 2000. *Language Testing*. Oxford University Press.
- Detmar Meurers, Kordula De Kuthy, Florian Nuxoll, Björn Rudzewitz, and Ramon Ziai. 2019. [Scaling up intervention studies to investigate real-life foreign language learning in school](#). *Annual Review of Applied Linguistics*, 39:161–188.
- OpenAI. 2024a. [OpenAI API](#). Software.
- OpenAI. 2024b. [OpenAI API Whisper](#). Software.
- Lourdes Ortega, Noriko Iwashita, John M Norris, and Sara Rabie. 2002. An investigation of elicited imitation tasks in crosslinguistic SLA research. In *Second Language Research Forum, Toronto*, pages 3–6. Paper presentation.
- Anne O’Keeffe and Geraldine Mark. 2017. [The English Grammar Profile of learner competence: Methodology and key findings](#). *International Journal of Corpus Linguistics*, 22(4):457–489.
- Andrea Révész and Tineke Brunfaut. 2020. [Validating assessments for research purposes](#). In *The Routledge Handbook of Second Language Acquisition and Language Testing*, pages 21–32. Routledge.
- Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, Verena Möller, Florian Nuxoll, and Detmar Meurers. 2018. [Generating feedback for English foreign language exercises](#). In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 127–136.

Simón Ruiz, Patrick Rebuschat, and Detmar Meurers. 2023. [Supporting individualized practice through intelligent CALL](#). In Yuichi Suzuki, editor, *Practice and Automatization in Second Language Research*, 1 edition, pages 119–143. Routledge, New York, NY.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. [Machine Learning–Driven Language Assessment](#). *Transactions of the Association for Computational Linguistics*, 8:247–263.

Derek Sleeman. 1985. Inferring (mal) rules from pupils’ protocols. In *Selected and updated papers from the proceedings of the 1982 European conference on Progress in artificial intelligence*, pages 30–39.

Nina Spada, Julie Li-Ju Shiu, and Yasuyo Tomita. 2015. [Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies](#). *Language Learning*, 65(3):723–751.

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021. [Using broad linguistic complexity modeling for cross-lingual readability assessment](#). In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54.

Shu-Ling Wu, Yee Pin Tio, and Yuening Zhao. 2023. [Examining the comparability of parallel English and Chinese elicited imitation tasks](#). *Research Methods in Applied Linguistics*, 2(3):100058.

Xun Yan, Yuyun Lei, and Chilin Shih. 2020. [A corpus-driven, curriculum-based chinese elicited imitation test in US universities](#). *Foreign Language Annals*, 53(4):704–732.

Xun Yan, Yukiko Maeda, and April Ginther. 2016. [Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis](#). *Language Testing*, 33(4):497–528.

Ramon Ziai, Björn Rudzewitz, Kordula De Kuthy, Florian Nuxoll, and Detmar Meurers. 2018. [Feedback strategies for form and meaning in a real-life language tutoring system](#). In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 91–98.

## Appendix A

### A.1 Test creation process

**New Test Collection**

Note: your new test collection will only become available to test takers when you release it in a later step. You can take your time to set it up.

Name  
EI Test  
Choose a speaking name for your test.

Description  
This is an example test.  
Internal description for you. Not shown to test takers.

Access code  
URM4 Re-generate  
Test takers will be able to access the test by entering this code.

Test visibility  
Private  
Public tests can be seen and duplicated by other test creators.

Cancel Create

Figure 5: User interface to create a new test

**Add test**

Name  
EI Test

Description  
This is an EI Test

Type  
Elicited Imitation

Add

Figure 6: Add test page with the test type Elicited Imitation

**EI Test**

This is an example test.  
access code: URM4

Test preview  
Preview

Add new test  
Add

**EI Test**  
This is an EI Test  
Items Instructions Settings

Cancel

Figure 7: Test component management, where instructions, items and settings can be edited and added

## A.2 Instructions

The interface for adding an instruction page consists of several sections. At the top, there is a title 'Add instruction page' and a small icon of a globe with a document. Below this, a paragraph explains that users can compose an instruction shown as a separate page and control its position. The 'Title' section has a text input field containing 'Instruction 1' and a note that the title will be displayed as a heading. The 'Instruction' section has a larger text input field containing 'After the beep, repeat the sentence in correct English.' Below this, there is a note about linebreaks and a 'Position' dropdown menu set to 'Before the practice items block'. At the bottom, there are three buttons: 'Cancel', 'Save', and 'Save & next'.

Figure 8: Interface for adding instruction pages for tests

## A.3 Settings

The 'Add setting' interface has a title 'Add setting' and a gear icon. A paragraph explains that settings can be added for a test, with defaults if none are added. The 'Setting' section features a dropdown menu with the text 'Et: Should a belief statement be displayed? [true/false]'. Below it is a note to select the parameter. The 'Value' section has a text input field containing 'true'. A note below explains that for binary answers, 'true' or 'false' should be entered, and for time units, the time in seconds should be entered. At the bottom, there are three buttons: 'Cancel', 'Save', and 'Save & next'.

Figure 9: Interface for adding test settings

## A.4 Item-level scoring

The 'Score result' interface has a title 'Score result' and a small icon of a person with a speech bubble. A paragraph states that users can override the automatic score of single responses with manual scores. Below this, there is a 'Test details' section with a small icon of a document and a magnifying glass, and a list of details: 'Test taker: Test Kibi', 'Test: Test Et', 'Description: Et for testing', and 'Access code: 3GP5'. A 'Score automatically' button is located to the right. The main content area is titled 'Test 1' and shows 'Name: Et Test', 'Description: Et Test', and 'Type: Et'. Below this is 'Item 1' with an 'Audio' section containing a 'Play' button. The 'Sentence' is 'The weather in Australia is nice in December.' and the 'Correct sentence' is the same. The 'Type' is 'Text' and the 'Note' is 'Item 1'. There are two 'Response' sections. The first has a 'Type' of 'belief\_statement' and a 'Response content' of 'Agree'. The second has a 'Type' of 'audio', a 'Recording' section with a 'Play' button, and a 'Response given' of 'Sep 30, 2024'. The 'Transcription' is 'The weather in Australia is nice in December.', the 'Automatic score' is '100.00%', and there is a 'Manual score' input field with a 'Save' button.

Figure 10: Interface to score responses to individual items

# Jingle BERT, Frozen All the Way: Freezing Layers to Identify CEFR Levels of Second Language Learners Using BERT

Ricardo Muñoz Sánchez, David Alfter, Simon Dobnik, Maria Irena Szawerna, Elena Volodina

University of Gothenburg, Sweden

ricardo.munoz.sanchez@gu.se

## Abstract

In this paper, we investigate the question of how much domain adaptation is needed for the task of automatic essay assessment by freezing layers in BERT models. We test our methodology on three different graded language corpora (English, French and Swedish) and find that partially fine-tuning base models improves performance over fully fine-tuning base models, although the number of layers to freeze differs by language. We also look at the effect of freezing layers on different grades in the corpora and find that different layers are important for different grade levels. Finally, our results represent a new state-of-the-art in automatic essay classification for the three languages under investigation.

## 1 Introduction

Automated essay scoring (AES) is the “process of evaluating and scoring written prose via computer programs” (Shermis and Burstein, 2003). Even though the implied use of computers nowadays might suggest so, AES is not a recent phenomenon. Ellis Batten Page, also known as “the father of AES” (Wresch, 1993), started to develop his ideas in the 60’s (Page, 1966; Page and Paulus, 1968) and implemented a rather sophisticated program to analyze and grade student essays. Even though work on AES started around 55 years ago, it is still an active area of research to this day (e.g. Beigman Klebanov and Madnani, 2020; Wilkens et al., 2023; Lagutina et al., 2023).

When dealing with pretrained language models, two of the most common approaches are to fine-tune the whole model or to just train any extra classification layers that have been added. Despite that, there have been studies that show that partly fine-tuning the models allows for better domain adaptation by maintaining part of the original knowledge of the model while learning domain-specific features at the same time (Zhu et al., 2021).

The reason for this is that different layers of neural models encode different kinds of features, with the first few encoding lower-level features and the later ones encoding higher-level features.

In this paper we aim to determine how much domain adaptation is required for AES. We limit our experiments to BERT models for a couple of reasons. There has been a lot of studies focusing on which layers of these models encode which aspects of linguistic knowledge (e.g. Clark et al., 2019; Jawahar et al., 2019). On the other hand, the more recent generative decoder-only models tend to vary a lot from each other, which can complicate both comparison among themselves and between different languages. Finally, the performance of these decoder-only models in terms of second-language assessment has had mixed results so far (Naismith et al., 2023; Yancey et al., 2023), which in turn means that BERT-based models are still an important part of AES for second language assessment.

Thus we analyze which layers of a pretrained BERT model are important for the task at hand and which ones should be fine-tuned. We assume that the knowledge embedded in the frozen layers (semantics, syntax, grammaticality, etc.) is important for the model to properly determine the proficiency level an essay has been annotated as. We further analyze whether this varies depending on the CEFR level of the essays. That is, we want to determine whether the same encoded knowledge of the language model is equally important for all levels.

We work with the CEFR<sup>1</sup> framework (COE, 2001). It is used to evaluate foreign/second language learning by assigning one of the six levels (A1, A2, B1, B2, C1, C2) that determine the proficiency of second language (L2) speakers. Furthermore, we work with three different languages: English, French and Swedish. While CEFR-labeled

<sup>1</sup>Common European Framework of Reference for Languages

data can be scarce, there is a growing societal need for automated grading in CEFR terms. An example of this is how different governments are either planning to require a language test for applicants for residence and citizenship or already do so (Code civil français, 2011; Swedish Government, 2021, 2023; U.S. Citizenship and Immigration Services, 2023; Government of Canada, 2024). Because of this, we expect that the need for support in AES will drastically increase in the near future, both as a way to support self-studying learners and for high-stakes essay grading.

The rest of our paper is organized as follows. Section 2 introduces the context for our experiment in terms of previous research. In Section 3.2 we describe our approach, as well as the considerations we have taken into account while designing it. Section 3.1 describes the datasets used for our experiments, while Section 3.3 describes the state-of-the-art we compare our models to. We present our results as well as a discussion of these in Section 4. Finally, we present our conclusions in Section 5, as well as possible directions in which to expand our work.

## 2 Related Work

The state-of-the-art in AES has long been dominated by systems using feature engineering and linguistic variables that measure textual quality, such as number of words (Shermis and Burstein, 2003; Parslow, 2015), number of grammatical errors (Yannakoudakis et al., 2018; Ballier et al., 2019), type-token ratio (Vajjala and Lõo, 2014; Lee and Hasebe, 2020), or lexical density (Hancke, 2013; Hancke and Meurers, 2013; Pilán and Volodina, 2018). It is only recently that deep learning approaches have begun to set new standards (Hussein et al., 2019; Bestgen, 2020).

Alikaniotis et al. (2016) and Taghipour and Ng (2016) were the first ones to use deep learning for AES. Even though they used an LSTM<sup>2</sup> architecture (Hochreiter and Schmidhuber, 1997), other network architectures such as Convolutional Neural Networks (CNN) and Recurrent Convolutional Neural Networks (RCNN) have also been successfully applied in the past (Dong and Zhang, 2016; Dong et al., 2017; Dasgupta et al., 2018; Shin and Gierl, 2021).

Recent experiments using GPT for CEFR classification have found that GPT-4 (OpenAI, 2024) can

reach performances approaching those of sophisticated automated scoring systems (Banno et al., 2024), although agreement with human annotators remained inconclusive (Yancey et al., 2023). Large Language Models have also been used for other tasks related to computational approaches to language learning, such as learner-adapted definition generation (Yuan et al., 2022), learner-centered text simplification (Baez and Saggion, 2023), or proficiency-adapted text generation (Bezirhan and von Davier, 2023).

As with most fields in NLP, most of the work in this field has been done in English (Søgaard, 2022). A consequence of that is that other languages are often not paid enough attention to.

For instance, very little work has been done on essay classification in Swedish, some examples being Östling et al. (2013) on grading upper-secondary essays written by native speakers, Pilán (2018) on CEFR classification of L2 learner essays, Lilja (2018) on assigning grades to high-school essays, and Ruan (2020) on assigning grades to essays written as a part of national exams. Some of these works use the Uppsala Corpus of Student Writings (Megyesi et al., 2016). This corpus mainly consists of native speaker upper secondary level writings but also contains some texts, around 8%, written by learners of Swedish as a second language. However, it is not aligned with the CEFR scale.

Both Lilja (2018) and Ruan (2020) use deep learning to classify these essays by assigned grades. Lilja (2018) uses an LSTM and explores whether pre-trained embeddings are better or not than a fine-tuned version or randomly initialized ones. They conclude that pre-trained fine-tuned embeddings produce the best results, but due to high standard deviations, they are not significantly different from randomly initialized embeddings.

Ruan (2020), explores the use of hand-crafted features in combination with deep neural networks. The feature categories are virtually identical to those in Pilán (2018), namely count-based, morphological, syntactic and lexical. Semantic features were not included. The chosen architecture is a recurrent neural network. Using each feature group separately, they find that all feature groups perform similarly, although each feature group separately performs better than using all features simultaneously. Overall, they find that a feature-based system outperforms the word embeddings based system by Lilja (2018).

---

<sup>2</sup>Long Short-Term Memory

A similar situation presents itself for French, with a limited number of studies on essay classification. For non-L2 French, [Lemaire and Dessus \(2001\)](#) use Latent Semantic Analysis to grade a limited number of student essays (31), and [Zaghouni \(2002\)](#) presents a conceptual design for grading essays using a multi-agent system. [Parslow \(2015\)](#) presents a preliminary study on automatic grading of L2 French essays written by Swedish native speakers using feature-based methods and Naive Bayes classifiers. Finally, [Ranković et al. \(2020\)](#) use CamemBERT to extract word-level features and a deep recurrent network to grade essays written by French learners in German-speaking parts of Switzerland.

[Mayfield and Black \(2020\)](#) argue that the move to deep neural models for AES comes with considerable computational costs while producing performance comparable to the classical models. Their conclusions indicate, however, that there is a further need to explore deep learning approaches.

### 3 Materials and Methods

Second language assessment is a high-stakes situation, given that its outcome can affect the educational and professional opportunities that a student has available to them. While deep learning models tend to out-perform feature-based models, they tend to be obscure, with little to no explanation both of where specific predictions come from and which kind of features they focus on ([Guidotti et al., 2018](#)).

In this section, we first introduce the datasets we used in Section 3.1, followed by our approach to obtain a more explainable BERT ([Devlin et al., 2019](#)) model in Section 3.2. Finally, we talk about the state-of-the-art we compare our approach to in Section 3.3.

#### 3.1 Datasets

##### 3.1.1 English Dataset

We are using the EFCamDat corpus ([Geertzen et al., 2013](#)) for experiments on English. The corpus consists of essays collected from the EF Education First online platform. The essays were assigned a grade on a 16-level scale with equivalents to some of the major standards in L2 language learning, including CEFR levels. However, it should be noted that the grades were assigned according to the level the students reached in the platform as opposed to direct evaluation of the essays themselves.

Level	# essays	# train	# valid	# test
A1	192K	2,299	767	767
A2	130K	1,555	518	518
B1	62K	738	246	246
B2	18K	218	73	73
C1	5K	62	20	20
C2	0	0	0	0
Total	406K	4,872	1,624	1,624

Table 1: Number of essays in the English L2 learner corpus (EFCamDat) for each of the CEFR levels. The letter *K* denotes that the numbers we are dealing are in the thousands. Note that there are no C2 level essays in the corpus. We randomly sample a small percentage of the corpus for faster training while keeping the label distributions the same.

The corpus contains over 400,000 essays from CEFR levels ranging from A1 to C1, as seen in Table 1. The students are placed into one of the platform’s 16 levels either through a placement test or by progressing through the course. Each level has eight possible writing tasks, which gives a wide array of possible topics for each CEFR level. Given that we are training the models several times, we sampled 2% of the data to keep the use of computational resources within a reasonable margin. The essays were randomly sampled and stratified by CEFR level, to maintain the proportion of each label. Moreover, this leaves us with a dataset of a comparable size to TCFLE-8, the French corpus we are using.

##### 3.1.2 French Dataset

For French, we use the recently released TCFLE-8 corpus ([Wilkens et al., 2023](#)). This is a corpus based on the French language certification exam TCF (test de connaissance du français ‘French knowledge test’) administered by the France Éducation International. It is the biggest French corpus for AES to date with over 6.5k essays and covers a wide variety of prompts.

All essays are graded by at least 2 professional raters and cover all six levels of the CEFR scale, as seen in Table 2. Different data cleaning and quality assurance steps were taken by the corpus creators to ensure that the corpus contains representative samples at each level.

Level	# essays	# train	# valid	# test
A1	689	413	138	138
A2	1,375	825	275	275
B1	1,466	880	293	293
B2	1,427	856	285	285
C1	1,127	676	226	226
C2	485	0	0	0
Total	6,569	3,650	1,217	1,217

Table 2: Number of essays in the French L2 learner corpus (TCFLE-8) for each of the CEFR levels. Note that this is the only corpus of the three that we are working with that contains C2 level essays. We have removed the essays of this level to allow for better comparison across languages.

Level	# essays	# train	# valid	# test
A1	59	35	12	12
A2	143	85	29	29
B1	86	52	17	17
B2	105	63	21	21
C1	96	58	19	19
C2	7	0	0	0
Missing	6	0	0	0
Total	502	293	98	98

Table 3: Number of essays in the Swedish L2 learner corpus (Swell-Pilot) for each of the CEFR levels. Note that there are very few essays of level C2 in the corpus and that some are missing a level.

### 3.1.3 Swedish Dataset

For Swedish, we use the Swell-pilot corpus (Volodina et al., 2016a; Volodina, 2024). It consists of three subcorpora of L2 Swedish learners (see below) and is annotated with CEFR levels. All CEFR levels are well represented in the corpus, with the exception of C2 level (advanced) essays, as seen in Table 3. Thus we remove the C2 essays as their low number would not be representative of the model’s classification capabilities. Moreover, there are six essays that lack a level which have been ignored for the purposes of this experiment.

**SpIn** consists of 256 essays from a course for refugees that had recently arrived to Sweden. The course was introductory in nature and the essays were part of a mid-term exam.

**SW1203** consists of 141 essays from a preparatory course for foreign students that intended to study an undergraduate program in Sweden.

**TISUS** consists of 105 essays from the written part of the Test In Swedish for University Studies (TISUS)<sup>3</sup>. The essays are argumentative, the topic being “stress”.

## 3.2 Methodology

In order to classify the essays, we use language-specific versions of BERT. For the experiments themselves, we explore how freezing different layers of BERT during training affects its performance. We freeze the layers in a bottom-up manner, given that lower layers learn more basic linguistic features such as surface-level features, while higher layers learn more task-specific features, such as semantic and contextual features (Clark et al., 2019; Jawahar et al., 2019). Thus, we compare different configurations ranging from a completely fine-tuned model to one where only the classification layer was trained.

For the classification task itself, we truncate the essays to fit the maximum token length of BERT and feed them to the model.<sup>4</sup> We then take the top layer representation of the [CLS] token and feed it to a linear layer for classification. Taking the output of the same layer all the time allows us to compare the differences between how the models are learning depending on how many layers we have frozen.

In terms of hyperparameters, we explore using different learning rates<sup>5</sup> and find that the best performing on average is 5e-5. We also run the experiments for 10 epochs, loading the best performing checkpoint at the end.

Given that none of the corpora used has standard train/test splits, we run our experiments five times, generating new train/validation/test splits with a 60/20/20 distribution each run to account for variance. We maintain the proportions of the different CEFR levels across the splits. The number of each label per level can be seen in Tables 1, 2, and 3.

As for our models, we use specific versions of BERT according to the language.

For English, we use the original version of BERT<sup>6</sup> (Devlin et al., 2019). It was trained using BooksCorpus (Zhu et al., 2015) and an English Wikipedia dump. Note that we are using the cased

<sup>3</sup><https://www.su.se/tisus/english/>

<sup>4</sup>Note that we are not using Longformer as it is not available in all of the languages we are working with.

<sup>5</sup>We experimented with learning rates of 1e-4, 5e-4, 1e-5, 5e-5, 1e-6, 5e-6, and 1e-7.

<sup>6</sup><https://huggingface.co/google-bert/bert-base-english-cased>



version of BERT as the Swedish model has no uncased version available.

We use CamemBERT<sup>7</sup> (Martin et al., 2020) for French, which is based on RoBERTa (Liu et al., 2019) rather than on vanilla BERT. It was trained using the French section of the OSCAR corpus (Suárez et al., 2019), a language annotated version of CommonCrawl.<sup>8</sup>

The final model we use is Swedish BERT<sup>9</sup> (Malmsten et al., 2020), a Swedish version of BERT implemented by KBLab at the National Library of Sweden<sup>10</sup>. It was trained on a combination of corpora containing newspapers, social media, official reports from the Swedish government, legal documents, and Wikipedia in Swedish.

### 3.3 State-of-the-art

In this section we talk about the current state-of-the-art in AES within the context of the datasets we are using. These results are summarized in the top row of Table 4.

#### 3.3.1 English

The most similar work to our own for English is by Schmalz and Brutti (2021) who use BERT for the classification of the EFCamDat data. They also work on subsets of the whole data (10k, 50k, 100k) due to space and computational constraints with using the whole corpus. We report their best results as state-of-the-art.<sup>11</sup>

#### 3.3.2 French

Wilkens et al. (2023) perform a series of essay classification experiments on the TCFLE-8 corpus in order to establish some first baselines: (1) a transformer-based approach using CamemBERT, (2) a feature-based approach using XGBoost, and (3) a simple logistic regression. For the feature-based algorithm in (2) and (3) they use a set of 119 features – distilled from over 5k features – from nine subcategories: errors, graded lexicons, lexical diversity, lexical frequency, lexical sophistication, orthographic neighbors, morphology, tenses, likelihood, and word length. They find the transformer-

<sup>7</sup><https://huggingface.co/almanach/camembert-base>

<sup>8</sup><https://commoncrawl.org/about/>

<sup>9</sup><https://huggingface.co/KB/bert-base-swedish-cased>

<sup>10</sup><https://www.kb.se/in-english/research-collaboration/kblab.html>

<sup>11</sup>It would arguably be fairer to compare against the results they obtained with the smallest subsample, approaching our own sample size.

based model to perform best, followed by XGBoost. For brevity, we will only report results from their best-performing model (i.e., the transformer-based model) as state-of-the-art.

#### 3.3.3 Swedish

For Swedish, we compare our model with a feature-based approach to be able to draw a comparison between performance and explainability. Pilán et al. (2016) and Volodina et al. (2016b) use a feature set of about 60 features divided into five subcategories: length-based, lexical, morphological, syntactic, and semantic features. They use an SVM to classify the data. Both studies found that lexical features perform the best.

Pilán and Volodina (2018) specifically investigate the importance of features for the classification of (1) sentences, (2) reading texts from textbooks, and (3) learner essays from SweLL-pilot. Using analysis of variance (ANOVA), they determine the most predictive features for each of the three subgenres of text. In general, this study corroborates findings from Crossley and McNamara (2011) for L2 English in that lexical diversity and lexical frequency are strong predictors in both studies, and Vajjala and Lõo (2014) who also found verb variation and lexical variation to be strong predictors for L2 Estonian.

### 3.4 Evaluation

We evaluate our system both in terms of accuracy and of “adjacent accuracy”. The idea behind adjacent accuracy is that an A1 essay misclassified as A2 is a smaller mistake as opposed to it being misclassified as a B2 essay.

In more formal terms, we say that a prediction is correct in terms of adjacent accuracy if: (1) our classes are ordinal and (2) the prediction is either the correct class or the immediate predecessor or successor of it.

Moreover, we use F1 score calculated using both usual and adjacent accuracy. We report both macro and weighted F1 scores as they aggregate the F1 scores for the individual classes assuming either that the classes are equally important (for macro averaging) or that the number of examples for each class matter (for weighted averaging).

## 4 Results and Discussion

### 4.1 Performance Across Languages

In this section we present the results of our experiments, noting the performance across languages

Layers Frozen	English	French	Swedish
State-of-the-art	0.974	0.56	0.23
None	0.975 ± 0.000	0.555 ± 0.003	0.722 ± 0.018
All layers	0.319 ± 0.000	0.443 ± 0.005	0.188 ± 0.001
Embedding Layer	0.971 ± 0.000	0.526 ± 0.005	0.727 ± 0.008
1 Encoder Layer	0.974 ± 0.000	0.517 ± 0.011	0.731 ± 0.019
1 and 2	0.974 ± 0.000	0.524 ± 0.010	<b>0.744 ± 0.011</b>
1 to 3	0.974 ± 0.000	0.538 ± 0.002	0.718 ± 0.006
1 to 4	<b>0.977 ± 0.000</b>	0.529 ± 0.011	0.720 ± 0.003
1 to 5	0.972 ± 0.000	0.537 ± 0.008	0.725 ± 0.010
1 to 6	0.966 ± 0.000	0.532 ± 0.017	0.705 ± 0.006
1 to 7	0.967 ± 0.000	0.542 ± 0.018	0.671 ± 0.009
1 to 8	0.962 ± 0.000	0.548 ± 0.006	0.664 ± 0.020
1 to 9	0.957 ± 0.000	0.552 ± 0.004	0.612 ± 0.011
1 to 10	0.946 ± 0.000	0.564 ± 0.004	0.596 ± 0.013
1 to 11	0.919 ± 0.000	<b>0.572 ± 0.001</b>	0.541 ± 0.004

Table 4: Weighted F1 scores for the different languages. Even though the number of layers to freeze to obtain the best-performing model varies across languages, the best model is always partially fine-tuned.

and CEFR levels. More detailed tables and results for each language can be found in Appendix B for the metrics based on accuracy and in Appendix C for those based on adjacent accuracy. Table 4 compares the weighted F1 scores among languages.

First of all we can notice that all BERT models that were even partially fine-tuned performed better than the fully frozen model. That is, fine-tuning even one layer led to large improvements in the performance.

Even though the best performing model was always partially fine-tuned, which layers should be frozen varied depending on the language. For instance, for English, the only model that performed better than the fully fine-tuned one was the one where we froze all layers up to the fourth encoder layer, indicating a reliance on surface-level features for classification. Meanwhile, the French model showed a preference towards fine-tuning just the last few encoder layers, indicating that a broad range of linguistic features may be necessary to accurately classify the essays. Finally, the Swedish model worked the best when few of the encoder layers were frozen, which again points to the importance of surface-level features for AES in Swedish.

Based on this, we can assume that maintaining basic knowledge of the language within the model is an important part of automated essay grading. This sounds reasonable, given that second language

learners tend to demonstrate an imperfect usage of the language. Moreover, we would prefer not to have this usage of the language overwrite the knowledge of the model.

Something notable is that when the model misclassified an essay, it usually assigned that essay to one of the adjacent levels. Even though the CEFR levels are ordinal to us humans, this information was not provided to the model at any point during training. This points to the model learning how to identify the level of the essay according to the linguistic characteristics, as students from adjacent levels are more likely to create similar texts than those for levels that are farther apart.

## 4.2 Performance Across CEFR Levels

Figures 1, 2, and 3 show how the different levels react to fine-tuning different layers of the models. We have cut-off the values that are below a certain threshold for each of the plots as they do not help us identify which layers are important for that specific class. Nevertheless, the full figures can be found in Appendix A.

For French and for Swedish we notice that the levels where the model performs the best are those that are closer to the edges of the CEFR scale, regardless of the language. This points to these levels being easier to classify as they are the most likely to be different from the other essays. On the other hand, levels B1 and B2 are the ones that have lower

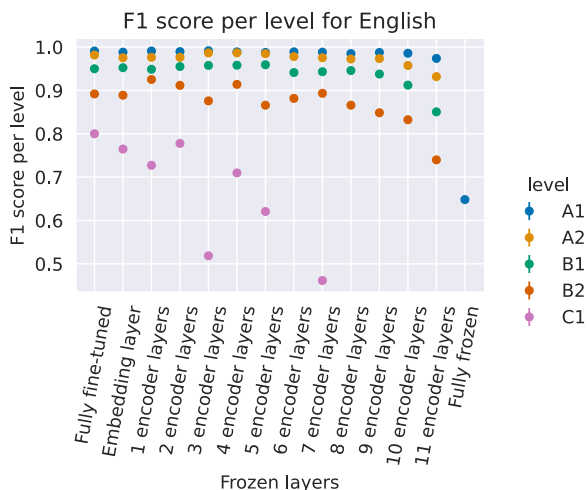


Figure 1: Performance per CEFR level when freezing different layers of BERT. Note that the performance tends to drop as the levels increase.

F1 scores. This might be due to them being more similar to their adjacent levels and thus harder to properly identify.

In more language specific notes, we see that the different levels tend to follow the same trend as the overall performance of each model.

We begin by looking at how the English BERT behaves across levels in Figure 1. We can note that the performance is inversely correlated to the level. That is, lower levels get higher F1 scores, while higher levels get lower F1 scores. This might be due to the prompts given to the students. For example, A1 essays have an almost perfect classification. However, most of them begin with a salutation (hi, hello, etc.) and address someone called Anna. This could in turn lead to leakage, which would explain the high performance seen in Table 4 compared to French and Swedish. Moreover, the levels are inferred from the course level, which Muñoz Sánchez et al. (2024b) argue is not necessarily a good proxy for CEFR levels. As for the individual levels, we notice that the general trend is for their accuracy to drop the more layers we freeze. Even though there are some layers that have either higher or lower perplexity, they do not seem to follow a pattern.

When looking at the French model in Figure 2 we notice that most of the levels have a slight increase in their performance as we approach the latter layers. However, different levels behave differently. For instance, the performance for level C1 is mostly stable with a very slight decrease when freezing just the first few layers and a very slight increase when fine-tuning just the last few layers.

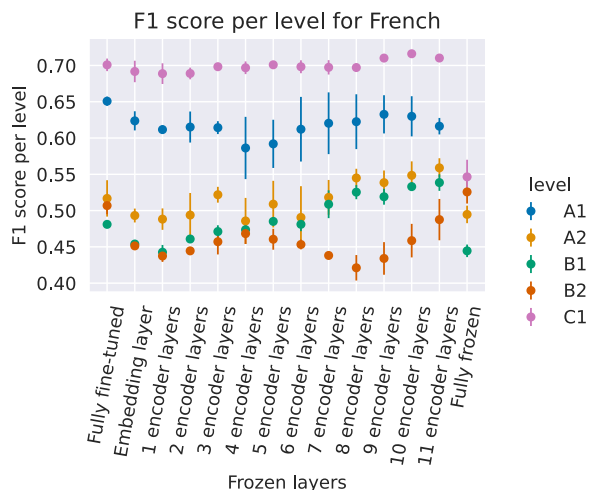


Figure 2: Performance per CEFR level when freezing different layers of CamemBERT. Note that even though all levels perform differently, most of them have a slight uptick in performance when we finetune only the last few encoder layers.

Meanwhile, level A1 has its highest performance when fine-tuning all of the model and another increase when freezing layers up to the ninth or tenth encoder layers, which points to the importance of a broad range of features. With levels A2, B1, and B2 we see a similar pattern: fine-tuning the whole model leads to higher performance but fine-tuning just the final encoder layer leads to the highest performance for these levels. Thus, we can assume that low-, mid- and high-level features play an important role in French AES. Even though the performance of our best model is similar to the one reported by Wilkens et al. (2023), we still see an increase in performance when freezing layers compared to fully fine-tuning the base model.

Finally, we take a look at Swedish BERT in Figure 3. Here we notice that there are two humps in the performance for levels A1 and A2. The first is when freezing just the first few layers and the second one is when freezing up to the first four or five encoder layers. This points to the importance of lexical and syntactic features. A similar pattern can be observed for level B1, albeit in a more erratic manner. For levels B2 and C1 we notice that freezing the first two decoder layers leads to the highest performance, pointing to the importance of lexical features.

## 5 Conclusions and Future Work

In this study we analyzed different fine-tuning strategies for AES using BERT-based models.

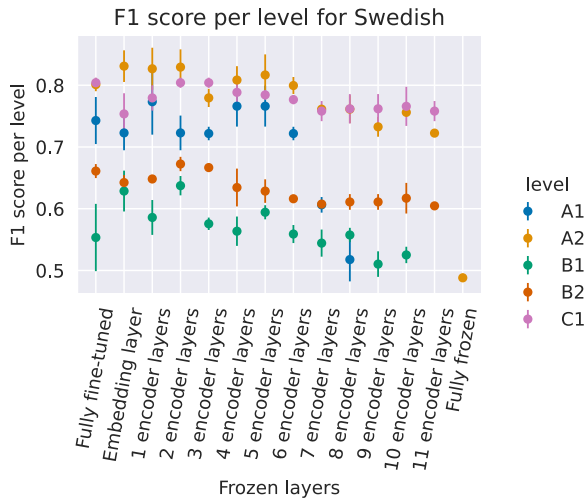


Figure 3: Performance per CEFR level when freezing different layers of Swedish BERT. Note that even though all levels perform differently, most of them have a sharp drop in performance when we finetune only the last few encoder layers.

Even though there was no unified pattern across languages on which layers are crucial, we show that the best-performing modes are ones that have gone through domain adaptation by partial fine-tuning. We also show that even though the importance of layers when taking into account the performance on each individual class differs, it tends to closely follow that of the whole model.

There are several directions in which our work can be expanded upon. The most immediate one would be to expand the languages used, as this would allow to identify if there are patterns depending on language families. On a similar note, we would be interested in seeing the effects of the L1 of a student on which layers and/or features are more important for the assessment.

Another important follow-up of our work would be to determine whether freezing specific layers leads to more fair systems. The idea behind this would be that a fair model should focus on the knowledge and skills of the students as opposed to spurious correlations such as (indirectly) using demographic data for classification. Human graders do tend to show slight biases based on these characteristics (Aldrin, 2017) and study on how deep learning models deal with these has been limited to perceived ethnicity of names (Muñoz Sánchez et al., 2024a).

Finally, we consider that it is important to do a deeper analysis both of the terms appearing in the essays and of the kinds of prompts given to

the students. As we mentioned, almost all of the essays in the A1 level in the English dataset include salutations as their first word. This is because the prompts for this level ask the students to greet or to introduce themselves to someone in specific. This can lead to a dataset in which it is not easy to identify whether our model is behaving as we expect or if it is looking as spurious correlations.

We consider that this work is an important step towards understanding which features are important when using transformer-based models for AES. This will in turn help create better and more interpretable models for this task, as well as will contribute to their fairness.

## Limitations

The present work only reports on works for the automatic assessment of written language. It should be mentioned that there is a substantial body of work done on automatic assessment of speech as well. Speech has its own specificities, for example fluency. Fluency is the rate at which one speaks, as operationalized in the Complexity, Accuracy, Fluency (CAF) framework (Skehan et al., 1998).

On top of that, the datasets and the approach we use in this paper aggregate several characteristics such as the grammar, vocabulary, relevance, among others into a single label for the whole essay. Naismith et al. (2023) note that this can lead to issues when automatically assigning a level to the essay, as some of these characteristics are harder to capture computationally, such as discourse coherence.

Another thing to note is that the models we used were originally trained using vastly different amounts of data. This could lead to differences in how they model language. For example, the models performed extremely well for the English dataset, while the performance was lower for both the French and the Swedish datasets. We recommend further analysis and cross-examination to ensure that none of these datasets were included in the training data for any of these models. On top of that, the French model is based on RoBERTa not on BERT, which might affect the results. To the best of our knowledge, CamemBERT is the most commonly used model derived from BERT in French.

## Ethics Statement

It is important to note that our model should not be used as a substitute for expert human graders.

As noted during the results, not even our model achieves perfect accuracy, which could impact the lives of students. Thus, we suggest always keeping a human-in-the-loop approach with this kind of technology.

## Acknowledgements

The research presented here has been enabled by the Swedish national research infrastructure Nationella Språkbanken, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions.

This work is part of a continuous endeavour on exploring the use of machine learning techniques to study second language learner texts in Swedish. Moreover, we aim to modernise the L2 tools that are currently available through the research infrastructure of Språkbanken Text, in particular via Språkbanken’s learning platform Lärka,<sup>12</sup> which allows researchers, as well as teachers and learners, to interact and analyse these kinds of texts in an automated manner.

The third author is also supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

## References

- Emilia Aldrin. 2017. [Assessing Names? Effects of Name-Based Stereotypes on Teachers’ Evaluations of Pupils’ Texts.](#) *Names*, 65(1):3–14.
- Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. [Automatic text scoring using neural networks.](#) In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany. Association for Computational Linguistics.
- Anthony Baez and Horacio Saggion. 2023. [LSLlama: Fine-tuned LLaMA for lexical simplification.](#) In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 102–108, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Nicolas Ballier, Thomas Gaillat, Andrew Simpkin, Bernardo Stearns, Manon Bouyé, and Manel Zarrouk. 2019. [A Supervised Learning Model for the Automatic Assessment of Language Levels Based on Learner Errors.](#) In Maren Scheffel, Julien Broisin, Viktoria Pammer-Schindler, Andri Ioannou, and Jan Schneider, editors, *Transforming Learning with Meaningful Technologies*, volume 11722, pages 308–320. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.
- Stefano Banno, Hari Krishna Vydan, Kate Knill, and Mark Gales. 2024. [Can GPT-4 do L2 analytic assessment?](#) In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 149–164, Mexico City, Mexico. Association for Computational Linguistics.
- Beata Beigman Klebanov and Nitin Madnani. 2020. [Automated evaluation of writing – 50 years and counting.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.
- Yves Bestgen. 2020. [Reproducing monolingual, multilingual and cross-lingual CEFR predictions.](#) In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5595–5602, Marseille, France. European Language Resources Association.
- Ummugul Bezirhan and Matthias von Davier. 2023. [Automated Reading Passage Generation with OpenAI’s Large Language Model.](#) *Computers and Education: Artificial Intelligence*, 5:100161.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. [What does BERT look at? an analysis of BERT’s attention.](#) In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.
- Code civil français. 2011. [Article 21-24 \(version en vigueur depuis le 18 juin 2011 \[entered into force on june 18, 2011\]\).](#)
- Council of Europe. COE. 2001. *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge University Press.
- Scott A. Crossley and Danielle S. McNamara. 2011. [Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing.](#) *International Journal of Continuing Engineering Education and Life Long Learning*, 21(2-3):170–191.
- Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha. 2018. [Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring.](#) In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 93–102, Melbourne, Australia. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding.](#) In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

<sup>12</sup><https://spraakbanken.gu.se/larka/>

- Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Fei Dong and Yue Zhang. 2016. [Automatic features for essay scoring – an empirical study](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077, Austin, Texas. Association for Computational Linguistics.
- Fei Dong, Yue Zhang, and Jie Yang. 2017. [Attention-based recurrent convolutional neural network for automatic essay scoring](#). In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.
- Jeroen Geertzen, Theodora Alexopoulou, Anna Korhonen, et al. 2013. [Automatic linguistic annotation of large scale I2 databases: The ef-cambridge open language database \(efcamdat\)](#). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*, pages 240–254.
- Government of Canada. 2024. [Documents for Express Entry: Language requirements](#). Accessed: 14-06-2024.
- Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. [A survey of methods for explaining black box models](#). *ACM computing surveys (CSUR)*, 51(5):1–42.
- J Hancke. 2013. [Automatic prediction of CEFR proficiency levels based on linguistic features of learner language](#). *Master Thesis. University of Tübingen, Tübingen, Germany*.
- Julia Hancke and Detmar Meurers. 2013. [Exploring CEFR classification for German based on rich linguistic modeling](#). In *Proceedings of the Learner Corpus Research (LCR) conference*, pages 54–56.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long Short-term Memory](#). *Neural computation*, 9:1735–80.
- Mohamed Abdellatif Hussein, Hesham Hassan, and Mohammad Nassef. 2019. [Automated language essay scoring systems: a literature review](#). *PeerJ Computer Science*, 5(208):1–16.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. [What does BERT learn about the structure of language?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.
- Nadezhda Stanislavovna Lagutina, Kseniya Vladimirovna Lagutina, Anastasya Mikhailovna Brederman, and Natalia Nikolaevna Kasatkina. 2023. [Text classification by cefr levels using machine learning methods and bert language model](#). *Modelirovanie i Analiz Informatsionnykh Sistem*, 30(3):202–213.
- Jae-Ho Lee and Yoichiro Hasebe. 2020. [Quantitative Analysis of JFL Learners’ Writing Abilities and the Development of a Computational System to Estimate Writing Proficiency](#). *Learner Corpus Studies in Asia and the World*, 5:105–120.
- Benoit Lemaire and Philippe Dessus. 2001. [A System to Assess the Semantic Content of Student Essays](#). *Journal of Educational Computing Research*, 24(3):305–320.
- Mathias Lilja. 2018. [Automatic Essay Scoring of Swedish Essays using Neural Networks](#). PhD Thesis. Uppsala University.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#).
- Martin Malmsten, Love Börjesson, and Chris Haffenden. 2020. [Playing with words at the national library of sweden – making a swedish bert](#).
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. [CamemBERT: a tasty French language model](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.
- Elijah Mayfield and Alan W Black. 2020. [Should you fine-tune BERT for automated essay scoring?](#) In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, Seattle, WA, USA → Online. Association for Computational Linguistics.
- Beáta Megyesi, Jesper Näsman, and Anne Palmér. 2016. [The Uppsala corpus of student writings: Corpus creation, annotation, and analysis](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3192–3199, Portorož, Slovenia. European Language Resources Association (ELRA).
- Ricardo Muñoz Sánchez, Simon Dobnik, Maria Irena Szawerna, Therese Lindström Tiedemann, and Elena Volodina. 2024a. [Did the names I used within my essay affect my score? diagnosing name biases in automated essay scoring](#). In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 81–91, St. Julian’s, Malta. Association for Computational Linguistics.
- Ricardo Muñoz Sánchez, Simon Dobnik, and Elena Volodina. 2024b. [Harnessing GPT to study second language learner essays: Can we use perplexity to determine linguistic competence?](#) In *Proceedings*

- of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), pages 414–427, Mexico City, Mexico. Association for Computational Linguistics.
- Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.
- OpenAI. 2024. *Gpt-4 technical report*.
- Robert Östling, Andre Smolentzov, Björn Tyrefors Hinnerich, and Erik Höglin. 2013. Automated essay scoring for Swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47, Atlanta, Georgia. Association for Computational Linguistics.
- Ellis B. Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243. Publisher: JSTOR.
- Ellis B. Page and Dieter H. Paulus. 1968. *The Analysis of Essays by Computer. Final Report*. Technical report, The University of Connecticut.
- Nicholas Parslow. 2015. *Automated Analysis of L2 French Writing: a preliminary study*. Master’s thesis. Publisher: Unpublished.
- Ildikó Pilán and Elena Volodina. 2018. Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58, Santa Fe, New-Mexico. Association for Computational Linguistics.
- Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111, Osaka, Japan. The COLING 2016 Organizing Committee.
- Ildikó Pilán. 2018. *Automatic proficiency level prediction for Intelligent Computer-Assisted Language Learning*. PhD Thesis, University of Gothenburg, Gothenburg, Sweden.
- Bojana Ranković, Sarah Smirnow, Martin Jaggi, and Martin J. Tomasik. 2020. Automated Essay Scoring in Foreign Language Students Based on Deep Contextualised Word Representations. In *LAK20-10th International Conference on Learning Analytics & Knowledge*. Issue: CONF.
- Rex Dajun Ruan. 2020. *Neural Network Based Automatic Essay Scoring for Swedish*. Master Thesis. Uppsala University.
- Veronica Juliana Schmalz and Alessio Brutti. 2021. Automatic assessment of English CEFR levels using BERT embeddings. In *Proceedings of the Eighth Italian Conference on Computational Linguistics*.
- Mark D. Shermis and Jill C. Burstein. 2003. *Automated essay scoring: A cross-disciplinary perspective*. Routledge.
- Jinnie Shin and Mark J. Gierl. 2021. More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing*, 38(2):247–272.
- Peter Skehan et al. 1998. *A cognitive approach to language learning*. Oxford University Press.
- Anders Søgaard. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019, Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.
- Svenska Regering Swedish Government. 2021. *Krav på kunskaper i svenska och samhällskunskap för svenskt medborgarskap*, sou 2021:2.
- Svenska Regering Swedish Government. 2023. *Kunskapskrav för permanent uppehållstillstånd*, sou 2023:25.
- Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.
- U.S. Citizenship and Immigration Services. 2023. *US-CIS Policy Manual: Chapter 2 - English and Civics Testing*. Accessed: 14-06-2024.
- Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 113–127, Uppsala, Sweden. LiU Electronic Press.
- Elena Volodina. 2024. *On two SweLL learner corpora – SweLL-pilot and SweLL-gold*. *Huminfra Conference*, pages 83–94.
- Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016a. SweLL on the rise: Swedish learner language corpus for European reference level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 206–212, Portorož, Slovenia. European Language Resources Association (ELRA).

- Elena Volodina, Ildikó Pilán, and David Alfter. 2016b. [Classification of Swedish learner essays by CEFR levels](#). In *CALL communities and culture – short papers from EUROCALL 2016*, pages 456–461. Research-publishing.net.
- Rodrigo Wilkens, Alice Pintard, David Alfter, Vincent Folny, and Thomas François. 2023. [TCFLE-8: a corpus of learner written productions for French as a foreign language and its application to automated essay scoring](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3465, Singapore. Association for Computational Linguistics.
- William Wresch. 1993. [The imminence of grading essays by computer—25 years later](#). *Computers and Composition*, 10(2):45–58.
- Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. [Rating short L2 essays on the CEFR scale with GPT-4](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584, Toronto, Canada. Association for Computational Linguistics.
- Helen Yannakoudakis, Øistein E. Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. [Developing an automated writing placement system for ESL learners](#). *Applied Measurement in Education*, 31(3):251–267. Publisher: Taylor & Francis.
- Jiaxin Yuan, Cunliang Kong, Chenhui Xie, Liner Yang, and Erhong Yang. 2022. [COMPILING: A benchmark dataset for Chinese complexity controllable definition generation](#). In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 921–931, Nanchang, China. Chinese Information Processing Society of China.
- Wajdi Zaghouani. 2002. [AUTO-ÉVAL : vers un modèle d'évaluation automatique des textes](#). In *Actes du colloque des étudiants en sciences du langage*, page 16, Montréal, Canada. Université du Québec à Montréal.
- Haichao Zhu, Zekun Wang, Heng Zhang, Ming Liu, Sendong Zhao, and Bing Qin. 2021. [Less is more: Domain adaptation with lottery ticket for reading comprehension](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1102–1113, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. [Aligning books and movies: Towards story-like visual explanations by watching movies and reading books](#). In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.



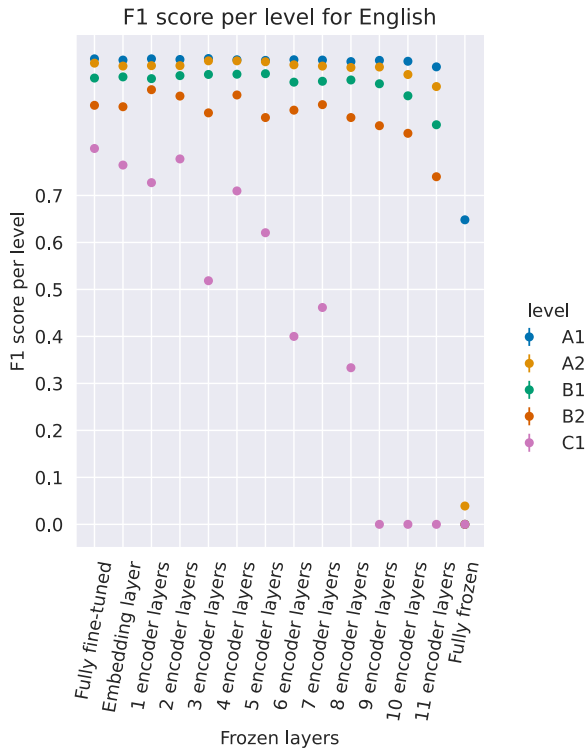


Figure 4: Performance per CEFR level when freezing different layers of the English model. Note that level A1 is the best performing one, while C1 is the worst.

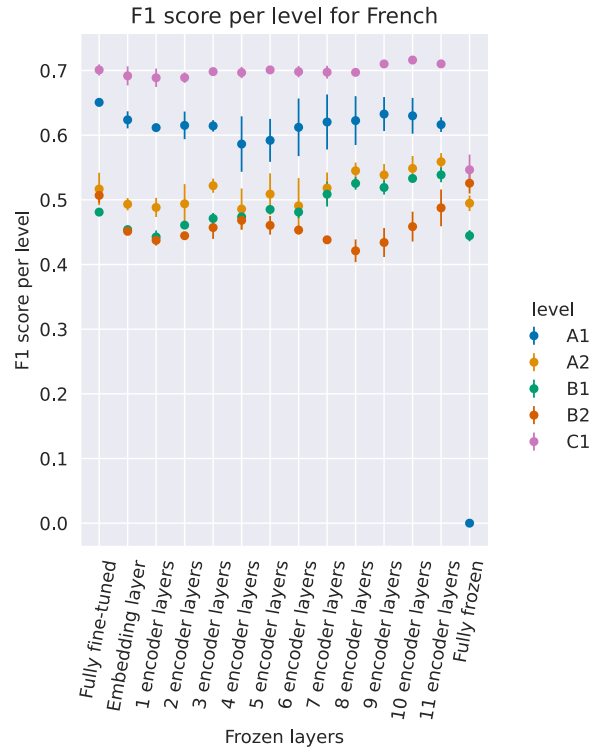


Figure 5: Performance per CEFR level when freezing different layers of the French model. Note that level C1 is the best performing one in general, followed by A1.

## A Performance Depending on the CEFR Level

In this appendix we present the figures for the F1 scores for the different languages. Figures 4, 5, and 6 show the effect of different degrees of fine-tuning of the BERT models across CEFR level in English, French, and Swedish, respectively.

## B Detailed Results per Language

In this appendix we present tables with the usual metrics for each language. The ones based on adjacent accuracy are in Appendix C. Thus, Tables 5, 6, and 7 show the performance of different degrees of fine-tuning of the BERT models in English, French, and Swedish, respectively.

## C Adjacent Metrics per Language

In this appendix we present tables with the metrics calculated using adjacent accuracy for each language. The ones based on standard accuracy are in Appendix B. Thus, Tables 8, 9, and 10 show the performance of different degrees of fine-tuning of the BERT models in English, French, and Swedish, respectively. Note that most of the experiments achieve very high results using these metrics.

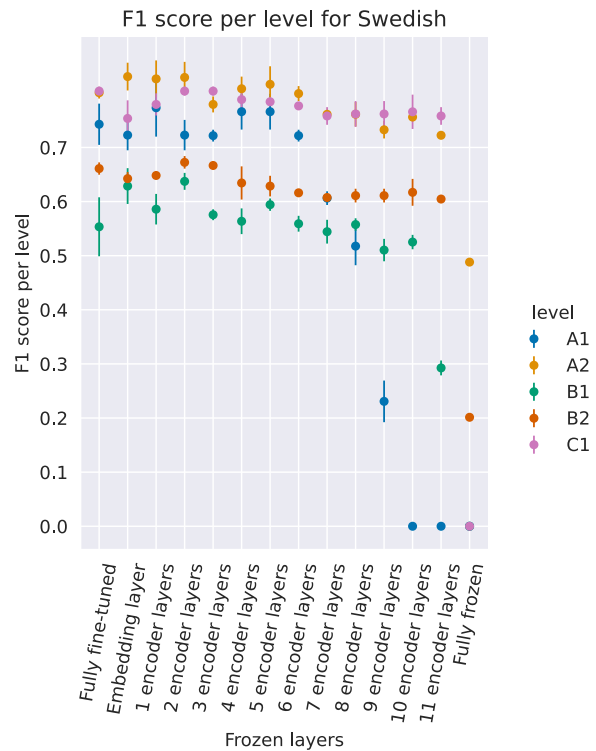


Figure 6: Performance per CEFR level when freezing different layers of the Swedish model. Note that levels A2 and C1 are the best performing ones in general, followed by A1.

Layers Frozen	Accuracy	F1 macro	F1 weighted
State-of-the-art (Schmalz and Brutti, 2021)	0.974	n/a	n/a
None	0.975 ± 0.000	0.923 ± 0.000	0.975 ± 0.000
All layers	0.475 ± 0.000	0.137 ± 0.000	0.319 ± 0.000
Embedding Layer	0.972 ± 0.000	0.914 ± 0.000	0.971 ± 0.000
1 Encoder Layer	0.974 ± 0.000	0.914 ± 0.000	0.974 ± 0.000
1 and 2	0.974 ± 0.000	0.922 ± 0.000	0.974 ± 0.000
1 to 3	0.975 ± 0.000	0.866 ± 0.000	0.974 ± 0.000
<b>1 to 4</b>	<b>0.977 ± 0.000</b>	<b>0.911 ± 0.000</b>	<b>0.977 ± 0.000</b>
1 to 5	0.973 ± 0.000	0.884 ± 0.000	0.972 ± 0.000
1 to 6	0.969 ± 0.000	0.838 ± 0.000	0.966 ± 0.000
1 to 7	0.969 ± 0.000	0.852 ± 0.000	0.967 ± 0.000
1 to 8	0.964 ± 0.000	0.820 ± 0.000	0.962 ± 0.000
1 to 9	0.962 ± 0.000	0.749 ± 0.000	0.957 ± 0.000
1 to 10	0.952 ± 0.000	0.737 ± 0.000	0.946 ± 0.000
1 to 11	0.924 ± 0.000	0.699 ± 0.000	0.919 ± 0.000

Table 5: Results of the various setups of English BERT model on the validation set using accuracy and macro and weighted F1. Note that the only result that outperforms a fully fine-tuned model was when freezing up to the fourth encoder layer. On top of that, the confidence interval was low enough for it to be considered practically zero.

Layers Frozen	Accuracy	F1 macro	F1 weighted
State-of-the-art (Wilkens et al., 2023)	0.57	n/a	0.56
None	0.560 ± 0.004	0.571 ± 0.003	0.555 ± 0.003
All layers	0.473 ± 0.005	0.402 ± 0.006	0.443 ± 0.005
Embedding Layer	0.533 ± 0.006	0.543 ± 0.004	0.526 ± 0.005
1 Encoder Layer	0.525 ± 0.011	0.534 ± 0.011	0.517 ± 0.011
1 and 2	0.533 ± 0.010	0.541 ± 0.012	0.524 ± 0.010
1 to 3	0.545 ± 0.003	0.553 ± 0.003	0.538 ± 0.002
1 to 4	0.538 ± 0.011	0.542 ± 0.014	0.529 ± 0.011
1 to 5	0.546 ± 0.008	0.549 ± 0.011	0.537 ± 0.008
1 to 6	0.542 ± 0.017	0.547 ± 0.020	0.532 ± 0.017
1 to 7	0.552 ± 0.018	0.557 ± 0.020	0.542 ± 0.018
1 to 8	0.559 ± 0.008	0.562 ± 0.010	0.548 ± 0.006
1 to 9	0.563 ± 0.006	0.567 ± 0.007	0.552 ± 0.004
1 to 10	0.573 ± 0.006	0.577 ± 0.007	0.564 ± 0.004
<b>1 to 11</b>	<b>0.578 ± 0.003</b>	<b>0.582 ± 0.002</b>	<b>0.572 ± 0.001</b>

Table 6: Results of the various setups of the French CamemBERT model on the validation set using accuracy and macro and weighted F1. Note that the best result on average is achieved when finetuning only the last encoder layer. More in general, finetuning the latter layers seems to lead to better results than also finetuning the earlier ones.

Layers Frozen	Accuracy	F1 macro	F1 weighted
State-of-the-art (Pilán et al., 2016)	0.18	0.16	0.23
None	0.727 ± 0.016	0.712 ± 0.021	0.722 ± 0.018
All layers	0.324 ± 0.004	0.138 ± 0.001	0.188 ± 0.001
Embedding Layer	0.731 ± 0.008	0.716 ± 0.008	0.727 ± 0.008
1 Encoder Layer	0.735 ± 0.020	0.723 ± 0.020	0.731 ± 0.019
<b>1 and 2</b>	<b>0.749 ± 0.012</b>	<b>0.733 ± 0.011</b>	<b>0.744 ± 0.011</b>
1 to 3	0.720 ± 0.008	0.710 ± 0.005	0.718 ± 0.006
1 to 4	0.724 ± 0.000	0.712 ± 0.003	0.720 ± 0.003
1 to 5	0.729 ± 0.012	0.718 ± 0.010	0.725 ± 0.010
1 to 6	0.710 ± 0.008	0.695 ± 0.005	0.705 ± 0.006
1 to 7	0.678 ± 0.008	0.656 ± 0.011	0.671 ± 0.009
1 to 8	0.673 ± 0.020	0.642 ± 0.021	0.664 ± 0.020
1 to 9	0.641 ± 0.016	0.569 ± 0.007	0.612 ± 0.011
1 to 10	0.649 ± 0.012	0.533 ± 0.014	0.596 ± 0.013
1 to 11	0.612 ± 0.000	0.476 ± 0.005	0.541 ± 0.004

Table 7: Results of the various setups of Swedish BERT model on the validation set using accuracy and macro and weighted F1. Note that the best result on average is achieved when finetuning the layers above the second encoder layer. Despite that, freezing some of the intermediate layers also leads to better results than those of the state-of-the-art.

Layers Frozen	Adj. Accuracy	F1 macro	F1 weighted
State-of-the-art (Schmalz and Brutti, 2021)	n/a	n/a	n/a
None	0.996 ± 0.000	0.997 ± 0.000	0.996 ± 0.000
All layers	0.799 ± 0.000	0.382 ± 0.000	0.721 ± 0.000
<b>Embedding Layer</b>	<b>0.998 ± 0.000</b>	<b>0.998 ± 0.000</b>	<b>0.998 ± 0.000</b>
1 Encoder Layer	0.996 ± 0.000	0.987 ± 0.000	0.996 ± 0.000
<b>1 and 2</b>	<b>0.998 ± 0.000</b>	<b>0.998 ± 0.000</b>	<b>0.998 ± 0.000</b>
1 to 3	0.996 ± 0.000	0.986 ± 0.000	0.996 ± 0.000
1 to 4	0.994 ± 0.000	0.984 ± 0.000	0.994 ± 0.000
1 to 5	0.994 ± 0.000	0.986 ± 0.000	0.994 ± 0.000
1 to 6	0.993 ± 0.000	0.964 ± 0.000	0.992 ± 0.000
1 to 7	0.993 ± 0.000	0.971 ± 0.000	0.993 ± 0.000
1 to 8	0.994 ± 0.000	0.988 ± 0.000	0.994 ± 0.000
1 to 9	0.995 ± 0.000	0.991 ± 0.000	0.995 ± 0.000
1 to 10	0.995 ± 0.000	0.990 ± 0.000	0.995 ± 0.000
1 to 11	0.996 ± 0.000	0.986 ± 0.000	0.996 ± 0.000

Table 8: Results of the various setups of English BERT model on the validation set using adjacent accuracy and the macro and weighted F1 scores that derive from it. Note that the best performance is achieved when freezing either just the embedding layer or by freezing up to the second encoder layer. This is the only model in which the best-performing does not match when using the usual accuracy and adjacent accuracy.

Layers Frozen	Adj. Accuracy	F1 macro	F1 weighted
State-of-the-art (Wilkins et al., 2023)	0.98	n/a	n/a
None	0.976 ± 0.002	0.976 ± 0.002	0.976 ± 0.002
All layers	0.952 ± 0.005	0.955 ± 0.005	0.952 ± 0.005
Embedding Layer	0.965 ± 0.001	0.966 ± 0.001	0.964 ± 0.001
1 Encoder Layer	0.958 ± 0.004	0.959 ± 0.003	0.958 ± 0.004
1 and 2	0.960 ± 0.002	0.961 ± 0.002	0.960 ± 0.002
1 to 3	0.965 ± 0.002	0.966 ± 0.002	0.965 ± 0.002
1 to 4	0.962 ± 0.001	0.963 ± 0.001	0.962 ± 0.001
1 to 5	0.960 ± 0.003	0.962 ± 0.002	0.960 ± 0.003
1 to 6	0.957 ± 0.004	0.958 ± 0.004	0.957 ± 0.004
1 to 7	0.960 ± 0.005	0.961 ± 0.005	0.960 ± 0.005
1 to 8	0.969 ± 0.002	0.970 ± 0.002	0.969 ± 0.002
1 to 9	0.972 ± 0.002	0.972 ± 0.002	0.972 ± 0.002
<b>1 to 10</b>	<b>0.976 ± 0.004</b>	<b>0.976 ± 0.003</b>	<b>0.976 ± 0.004</b>
<b>1 to 11</b>	<b>0.976 ± 0.002</b>	<b>0.976 ± 0.002</b>	<b>0.976 ± 0.002</b>

Table 9: Results of the various setups of French CamemBERT model on the validation set using adjacent accuracy and the macro and weighted F1 scores that derive from it. Note that the best result on average is achieved when finetuning either the final encoder layer or the final two.

Layers Frozen	Adj. Accuracy	F1 macro	F1 weighted
State-of-the-art (Pilán et al., 2016)	0.59	0.54	0.66
None	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
All layers	0.627 ± 0.012	0.585 ± 0.016	0.544 ± 0.015
Embedding Layer	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
1 - 10	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
1 to 11	0.992 ± 0.004	0.993 ± 0.003	0.992 ± 0.004

Table 10: Results of the various setups of Swedish BERT model on the validation set using adjacent accuracy and the macro and weighted F1 scores that derive from it. Note that the best result on average is achieved when finetuning the layer above the third encoder one.

# Generating Contexts for ESP Vocabulary Exercises with LLMs

Iglika Nikolova-Stoupak<sup>1\*</sup>, Serge Bibauw<sup>2</sup>, Amandine Dumont<sup>3</sup>,  
Françoise Stas<sup>3</sup>, Patrick Watrin<sup>1</sup>, Thomas François<sup>1</sup>

<sup>1</sup> CENTAL, Université catholique de Louvain, Belgium,

<sup>2</sup> GIRSEF, Université catholique de Louvain, Belgium

<sup>3</sup> ILV, Université catholique de Louvain, Belgium

\* [iglika.nikolova@uclouvain.be](mailto:iglika.nikolova@uclouvain.be)

## Abstract

The current paper addresses the need for language students and teachers to have access to a large number of pedagogically sound contexts for vocabulary acquisition and testing. We investigate the automatic derivation of contexts for a vocabulary list of English for Specific Purposes (ESP). The contexts are generated by contemporary Large Language Models (namely, Mistral-7B-Instruct and Gemini 1.0 Pro) in zero-shot and few-shot settings, or retrieved from a web-crawled repository of domain-relevant websites. The resulting contexts are compared to a professionally crafted reference corpus based on their textual characteristics (length, morphosyntactic, lexicosemantic, and discourse-related). In addition, we annotated the automatically derived contexts regarding their direct applicability, comprehensibility, and domain relevance. The 'Gemini, zero-shot' contexts are rated most highly by human annotators in terms of pedagogical usability, while the 'Mistral, few-shot' contexts are globally closest to the reference based on textual characteristics.

## 1 Introduction

The development of a wide vocabulary is a fundamental component of foreign language acquisition as it underpins the development of all other language skills (Ardasheva et al., 2019; Gorjian et al., 2011). To pursue this aim, learners are typically encouraged to exploit multiple strategies, such as studying from traditional mono- or bilingual vocabulary lists or making use of technology-based resources such as digital flashcards or vocabulary

learning apps (Restrepo Ramos, 2015).

Research shows that new vocabulary items are better acquired when encountered in authentic and informative contexts (Huckin and Coady, 1999; Restrepo Ramos, 2015; Godwin-Jones, 2018). However, looking for or coming up with high-quality contexts, especially more advanced and specialised ones, presents a serious challenge to teaching professionals in terms of time and effort. Therefore, the use of contemporary Natural Language Processing (NLP) techniques to come up with a large number of pedagogically sound contexts would present a significant benefit to both teachers and learners.

Against this backdrop, this paper presents our detailed experiments in deploying NLP methods to generate or retrieve contexts to help the acquisition of specialised English vocabulary by French-speaking university students reading science and agronomy. We used two Large Language Models (LLMs) of different sizes, namely Mistral-7B-Instruct and Gemini 1.0 Pro (in the context of both a zero-shot and a few-shot setting) and a custom-made web-based scientific corpus to produce context sentences for a predefined vocabulary list of 100 items belonging to CEFR levels B1-B2 in those two specialised domains. Our ultimate goal is to use the issuing contexts in the creation of exercises of the 'gapfill' and 'multiple-choice' types (see Fig. 1).

In this context, this paper addresses the three following research questions:

1. Which derivation method (web retrieval or LLM-generated) results in contexts for ESP vocabulary learning that are closer to professionally crafted ones in terms of textual char-

---

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

1. Climate models have traditionally shown considerable inaccuracy in their simulations of the Arctic. This **sh**..... is particularly troubling nowadays, because the Arctic is the region expected to undergo the most extreme climate changes in the future.
2. Climate models have traditionally shown considerable inaccuracy in their simulations of the Arctic. This ..... is particularly troubling nowadays, because the Arctic is the region expected to undergo the most extreme climate changes in the future.
  - a. cluster
  - b. shortcoming**
  - c. assertion
  - d. insight
  - e. endeavour

Figure 1: Examples of relevant 'gapfill' (1) and 'multiple-choice' (2) questions.

acteristics?

2. To what extent is it possible to guarantee the pedagogical quality of the issued contexts and their ready application in the classroom?
3. Is there a perceivable correlation between the contexts' textual characteristics and their pedagogical qualities as evaluated by teaching professionals?

The paper is organised as follows: Section 2 discusses related work regarding the pedagogical qualities of educational texts, automatic derivation of teaching materials, as well as their evaluation, with a particular focus on materials for the acquisition of EFL vocabulary. Section 3 explains our methodology for assembling and evaluating the examined corpora, and Section 4 presents the results of our experiments. We discuss our main findings in Section 5 and finally offer a conclusion and future directions in Section 6.

## 2 Background

### 2.1 Pedagogical Characteristics of Texts

There exists a variety of theories and perspectives when it comes to the definition of what makes a text suitable for a pedagogical setting, particularly in the context of foreign language learning. Siregar and Purbani (2024) draw attention to a number of narrow grammatical features as a guarantee for pedagogical suitability, such as the lack of nominalisations and extensive modifiers and the use of simpler patterns, such as *noun + preposition* or single clauses. Pedagogical qualities may also be dependent on the specific classroom addressed. Targeting younger learners, Morais and Neves (2010) underline the importance of interdisciplinarity in learning materials and tasks. Yet, most researchers agree that the essential prerequisite for any input in language acquisition is that

it should be "contextualised and comprehensible" (Tomlinson, 2012, 156).

Much emphasis has been placed on a text's *authenticity* as a pedagogical quality. A text is seen as authentic if it has been produced to serve a social purpose rather than a pedagogical one (Little et al., 1989). As a document's feature, authenticity has, hence, commonly been equated to a lack of adaptation, to the retaining of a text's original goal or context (Besse, 1981; Crossley et al., 2007). Yet, the superiority of authentic texts is still a subject of debate. Text simplification, for instance, has been shown to provide clear pedagogical advantages in textual characteristics (Crossley et al., 2007) and in comprehension and vocabulary learning effects (Rets and Rogaten, 2021). The emergence of generative AI also opens new debates on what qualifies as authentic, as such applications produce texts that are neither pedagogical nor the product of genuine human communication.

### 2.2 Automatic Derivation of Teaching Materials

The large amount of available data and the automatization opportunities that recent technology offers have been used extensively in the composition and presentation of teaching materials, particularly in the English as a Foreign Language (EFL) classroom. Various types of texts are derived from the web and typically adapted for use in a specific learning setting (Litman, 2016; Meurers et al., 2010). For instance, Heilman et al. (2008) gather a web-based textual corpus meant for vocabulary and reading practice as well as devise a user-friendly system (REAP Search) that enables the selection of elements from the corpus based on a list of relevant constraints.

In the past few years, LLMs have also been exploited in the language classroom due to their revolutionary ability to produce language based on personalised instructions. Expectedly, due to

its popularity and ease of access, ChatGPT has been receiving particular attention. A number of experimental studies have been conducted internationally in an attempt to define and estimate the chatbot’s potential to aid students in the ESL classroom. Following interaction with ChatGPT, learners of various age and proficiency levels are commonly discovered to have been motivated by the tool; furthermore, their academic results have been objectively improved, notably in the field of vocabulary acquisition, thanks to activities such as conversational practice and work with automatically generated text (Young and Shishido, 2023a,b; Shaikh et al., 2023; Songsingchai et al., 2023; Aktay and Uzunoglu, 2023; Lou, 2023).

### 2.3 Evaluation of Automatically Derived Teaching Materials

Jeon and Lee (2023) sum up LLMs’ applicability to language education as belonging to four discrete roles, namely interlocutor, content provider, teaching assistant, and evaluator. As per their last role, LLMs are claimed to be able to automatically evaluate the quality of student- and teacher-produced materials, as well as of automatically generated ones. Yet, such an evaluation by LLMs has not been substantially addressed due to its qualitative nature, and consequently, more traditional NLP techniques, especially related to readability or, otherwise, textual complexity in its different aspects, are typically applied to estimate textual quality and/or suitability. For instance, Loiseau et al. (2005) proposed an NLP-based system for pedagogical indexation where, upon insertion of a text or extract and indication of the intended learners’ level, its difficulty is estimated, and elements that may need to be adapted, such as complex grammatical tenses or vocabulary items, are highlighted. Aiming at consistent and large-scale evaluation of adapted internet materials, Hussin et al. (2010) performed a correlation analysis between the difficulty of texts as estimated by teachers and their readability characteristics, discovering statistical significance in relation to average sentence length, average word length and the coverage of the first 2000 high-frequency words.

Relevant human-based counterparts of generated materials have also been utilised as ground truth against which to evaluate them. For instance, Yunju et al. (2022) specifically addressed the evaluation of vocabulary exercises; more specif-

ically, in Chinese as a target language. They evaluated the quality of AI-generated distractors (non-correct answers) for multiple-choice questions based on a combination of semantic and visual similarity to the correct answer. Results and qualitative reflections of the test takers suggested that the automatically generated distractors are more complicated, possibly for reasons including the semantic similarity between them and their absence from textbooks used by the students. In a study related to the present one, Nikolova-Stoupak et al. (2024) generated and retrieved a number of contexts around ESP vocabulary list items and evaluated them based on their closeness to a gold standard of professionally crafted contexts in terms of a number of atomic readability-related features. Generated teaching materials for vocabulary acquisition have also been evaluated quantitatively in terms of compactness or informativeness. Paddags et al. (2024) generated sentences aimed at the teaching of Danish vocabulary using a few-shot LLM setting and consequently evaluated their quality based on their density in terms of the number of target words (based on a defined vocabulary list) that fit into a single sentence.

## 3 Methods

We conducted a series of experiments in deploying NLP methods to generate and retrieve contexts around a predefined vocabulary list of 100 items belonging to CEFR levels B1-B2 and the domains of general science and agronomy<sup>1</sup>. Each item is associated with a gold standard context as hand-picked by teaching professionals and previously used in a classroom for testing purposes (gapfill or multiple-choice questions). In particular, we generated contexts using two Large Language Models (LLMs) of different sizes, Mistral-7B-Instruct and Gemini 1.0 Pro, in both a zero-shot and a few-shot setting. In addition, we composed a corpus of scientific articles from relevant web sources and formulated a pipeline to extract relevant context sentences from them.

An important part of our work was to devise methods that guarantee that the derived contexts are of high educational quality and are thus directly applicable in an ESP classroom setting. Via

---

<sup>1</sup>the items were selected based on a larger pre-selection verified by teaching professionals; a balance between parts of speech was sought

hand-crafted rules, we ensured that the derived contexts resemble the gold standard defined by [Nikolova-Stoupak et al. \(2024\)](#) in terms of linguistic characteristics (as represented by common readability features). Additionally, we limited output to the appropriate scientific domain and CEFR level with the help of prompt engineering and classifier-based filters. The contexts issued from the different derivation methods were then manually annotated by experienced teachers of ESP from the Catholic University of Louvain in terms of their educational quality. Using insights from this human evaluation, we classified the contexts and, by extension, the methods behind their derivation, discussing their qualities and drawbacks and drawing conclusions about the interdependence between their automatable linguistic characteristics and their pedagogical qualities.

This section elaborates on the automatic derivation of the corpora (for an illustration of the process, see [Figure 2](#)) as well as on the methods applied in their evaluation.

### 3.1 Retrieval of Web-Crawled Contexts

Firstly, all accessible articles from a list of thematic websites as defined by a team of ESP teachers (see [Appendix 1: List of Crawled Websites](#)) were retrieved through web-crawling Python tools, such as *beautifulsoup*<sup>2</sup> and *newspaper*<sup>3</sup> and shaped into a database along with metadata including the textual format, date, the source webpage and its associated domain<sup>4</sup>. The derived text underwent a simple cleaning pipeline, such as the removal of non-alphanumeric symbols and non-English text. Context sentences associated with the predefined vocabulary list were then extracted from the database using a pipeline of hand-crafted rules. The articles were surveyed to determine the occurrence of the target vocabulary items or their alternative forms. When the search form was mapped, consistency was sought with the item’s domain and part of speech.

Several filters were then applied, ensuring that the target item is present in the sentence only a single time, that the sentence can be considered as scientific, that its CEFR level closely matches

<sup>2</sup>Version 4.12.3; <https://pypi.org/project/beautifulsoup4/>

<sup>3</sup>Version 0.2.8; <https://pypi.org/project/newspaper3k/>

<sup>4</sup>Among the three following domains: science, agronomy, and technology.

the intended level, and, eventually, that its linguistic characteristics<sup>5</sup> resemble those of a set of professionally crafted contexts sampled from the reference dataset of [Nikolova-Stoupak et al. \(2024\)](#). More precisely, this proximity was measured as Euclidean distance over the set of features and data points of up to two standard deviations from the values computed on the reference corpus were retained<sup>6</sup>.

#### 3.1.1 CEFR Level Classifier

It was important to guarantee that the contexts that were retrieved closely matched the CEFR level associated with the vocabulary list items that they were mapped with. Upon experimentation with established and readily available tools designed for estimation of the CEFR level of texts, such as *Textinspector*<sup>7</sup> and *English CEFR Level Predictor*<sup>8</sup>, it was observed that these solutions do not work well when faced with text that is a single sentence of length. Therefore, a custom classifier was trained to determine the extracted sentences’ CEFR levels. We built a corpus of sentences annotated with their CEFR level by concatenating [Arase et al. \(2022\)](#)’s WikiAuto- and SCORE-based corpora, which were annotated by two experienced teaching experts<sup>9</sup> and the sentences available through the *English Profile* website ([Salamoura and Saville, 2010](#)), which are originally taken from the Cambridge Learner Corpus<sup>10</sup> and exemplify discrete CEFR levels with their characteristics. We then used this corpus, which totalled 13,378 sentences, to finetune a BERT model

<sup>5</sup>The set of characteristics that we considered in this work is: the number of words, the number of letters per word, the number of punctuation signs, the number of noun phrases, the percentage of non-stem words, the number of first-person pronouns, the number of proper nouns, the number of pronouns, and the number of anaphora-denoting words. Please refer to [Appendix 3: Features Used in Corpus Comparison](#) for details about these characteristics.

<sup>6</sup>For some features, the value was increased or reduced based on observations. Sentence length was thus limited to 1.5 standard deviations from the reference, and the percentage of non-stem words was relaxed to 3 standard deviations. When present, negative values were rounded to 0.

<sup>7</sup><https://textinspector.com/api-developers/>

<sup>8</sup><https://github.com/AMontgomerie/CEFR-English-Level-Predictor>

<sup>9</sup>where the two annotators’ estimations differed, we took the higher CEFR level as it is less problematic for students to be provided with text that is slightly below their current level

<sup>10</sup><https://www.sketchengine.eu/cambridge-learner-corpus/>



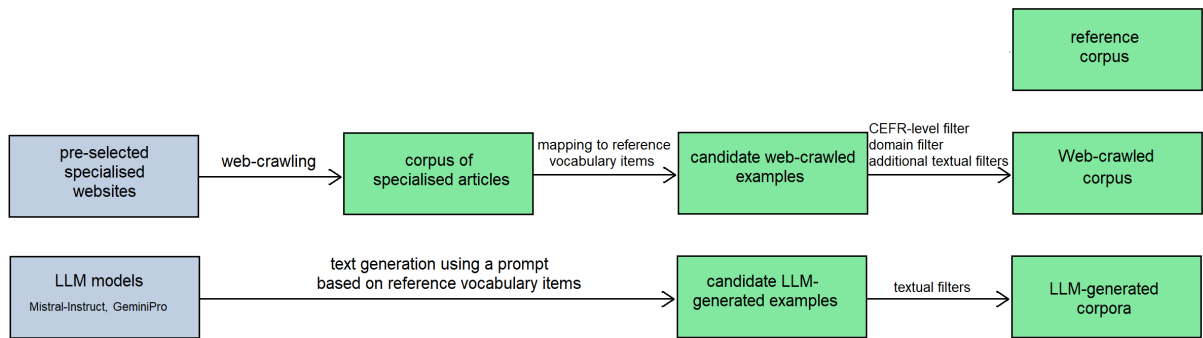


Figure 2: Collection procedure for the examined corpora

with a classification layer<sup>11</sup>. The derived classifier achieved 63% of accuracy<sup>12</sup>, the majority of mistakes being associated with the marginal A1 and C2 proficiency levels, which are absent from our reference vocabulary list. When adjacent levels were considered, the accuracy went up to 98%. Given the qualitative nature of CEFR levels and the lack of full agreement between the used corpus annotators, candidate web-crawled sentences were retained if they belonged to the associated course’s level or differed from it by a single level. Ultimately, only a small portion (around 10%) of the candidate sentences were discarded based on the CEFR-level filter.

### 3.1.2 Scientific Domain Classifier

As the web-crawled articles do not consist of scientific text in their entirety (e.g. there may be isolated informal sentences or even metadata within them), a binary SVM classifier model<sup>13</sup> was trained to label sentences that belong to a broad scientific domain. At first, the training corpus was composed of 2k scientific and 14k non-scientific sentences (2k ‘law’, 2k ‘business’, 2k ‘sports’, 2k ‘world news’, 2k ‘law’, 2k ‘informal communication’, and 2k ‘literature’), taken at random from the following sources: respectively, PubMed<sup>14</sup> (the ‘scientific’ label); the Caselaw Access Project<sup>15</sup>; the AG News Classification Dataset’s Business News, Sports News and World

News subcorpora<sup>16</sup>, Reddit’s API<sup>17</sup>, and an assembled corpus of full and abridged classical literary texts as freely available online. The classifier’s performance was then tested on a random 100-sentence sample extracted from our web-crawled corpus, and a bias toward complex sentences, as well as an underrepresentation of certain scientific fields, such as chemistry, were detected. In order to improve the classifier, 1000 sentences with a length of up to 2 standard deviations from the reference value for the feature (as defined by Nikolova-Stoupak et al. (2024)), which were also manually confirmed to be scientific, were added to the training corpus’s ‘scientific’ label. The newly derived classifier achieved 93% accuracy, and its performance was verified against the 100 manually labelled sentences and judged to act satisfactorily as a filter. The resulting classifier was used in the extraction of web-crawled context sentences, and non-scientific sentences (which turned out to be about one-third of the candidate sentences) were disregarded.

### 3.2 Generation of Contexts by LLMs

Two discrete contemporary LLMs were used for context generation: Mistral-7B-Instruct and Gemini 1.0 Pro. The former is a compact model whose performance compares to and occasionally surpasses that of LLaMA (Jiang et al., 2023), whilst the latter is a 600B variety of Gemini, a model that achieves state-of-the-art results in a number of key NLP tasks (Team et al., 2024) and is characterised with fast performance. For this experiment, Mistral was used through the ‘LM studio’ interface, and Gemini was accessed through the Google AI Studio developer tool, as freely available within a

<sup>11</sup>BERT was opted for in this task due to its strong language understanding and generation abilities

<sup>12</sup>Arase et al. (2022)’s associated classifier reaches a macro-F1 score of 84.5% as a result of elaborate techniques especially aimed at the correct recognition of sentences belonging to the rarer and marginal CEFR levels

<sup>13</sup>the choice of model was based on experiments with a few models that are strong in binary classification tasks

<sup>14</sup><https://pubmed.ncbi.nlm.nih.gov/>

<sup>15</sup><https://case.law/docs/>

<sup>16</sup><https://www.kaggle.com/datasets/amanandrai/ag-news-classification-dataset>

<sup>17</sup><https://www.reddit.com/wiki/api/>

given quota at the time of writing. Following experiments, Mistral’s temperature setting was adjusted to 0.8, as below this value, the output was highly homogeneous and commonly consisted of definitions of the target vocabulary. The experimental setup featured an 11th Gen Intel Core i7 CPU with 8 cores, and TigerLake-LP GT2 integrated GPU.

Both models were instructed to provide context examples based on the vocabulary list’s items, parts of speech and domains in both a zero-shot setting and a few-shot setting. Within the latter, five examples of paired vocabulary items of various parts of speech and corresponding reference contexts as provided by teaching professionals were added to the prompts. For the full prompts utilised, please refer to [Appendix 2: Prompts used for LLM Generation](#).

In addition, generation for a vocabulary item was iterated through until a number of conditions pertaining to the output were satisfied. As with the retrieval of web-crawled contexts, it was ensured that the target item was only present in the example a single time and that the example was of proximity to the gold standard sampled from [Nikolova-Stoupak et al. \(2024\)](#) as measured with Euclidean distance based on a selection of readability features<sup>18</sup>. We confirmed that the output was in English as well as the compatibility of the its part of speech and the absence of metatextual information (e.g. explanations of use) in addition to context examples.

### 3.3 Human Annotation

For the purpose of annotation, each of the five methods described above (generation with Mistral and Gemini in a zero-shot and few-shot setting and retrieval from a web-crawled database) was used to generate one context for each of the 100 words in our vocabulary list. This amounted to a total of 500 contexts, which were assigned unique IDs before being shuffled and information about their generation method being removed. Two ESP teachers with substantial experience were asked to evaluate the contexts for the following three pedagogical features: ‘ready to use’, ‘comprehensible’ and ‘in-domain’. A ‘ready to use’ context was defined as being directly applicable for classroom use and assessment purposes without editing; ‘comprehensibility’ referred to a context

being self-explanatory and understandable if encountered in isolation; finally, ‘in-domain’ meant that the contexts’ field of specialisation is appropriate for students in the intended specialisation (i.e. science or agronomy). A Likert scale from 1 to 5 was utilised for the annotation, 5 signifying maximal possession of the quality in question. The annotators were also invited to leave comments in free text in relation to each of the evaluated contexts.

Initially, the annotators were given the first 100 contexts to annotate independently of each other, following which the inter-rater agreement between them was calculated. The ‘ready to use’ and ‘comprehensible’ categories are marked with moderate agreement according to Cohen’s Quadratic Kappa but demonstrate good scores for exact agreement (respectively 65% and 87%). The ‘in-domain’ category comes with a low Cohen’s Kappa value in combination with 97% exact agreement, a phenomenon caused by the heavily skewed ratings towards a maximal number of points for the category ([Pontius Jr and Millones, 2011](#)). As a next step, the annotators gathered to adjust the annotation guidelines and agree on a gold value for the items where their initial annotation differed. The remaining 400 contexts were then split between the two teachers to annotate.

### 3.4 Context Evaluation

The 500 contexts, as well as the 100 reference ones, were evaluated based on readability-related textual characteristics (see [Appendix 3: Features Used in Corpus Comparison](#)). An analysis identical to the one defined by [Nikolova-Stoupak et al. \(2024\)](#) followed. That is to say, firstly, the non-parametric Mann-Whitney U test was used to measure the significance of the difference between the reference corpus and contexts for each of the five collection methods (retrieval from a web-crawled corpus and generation by Mistral and Gemini in a zero-shot and few-shot setting). In turn, statistical significance was assigned to one of three levels, corresponding to p-values of 0.001, 0.01, and 0.05. In addition, the global distance between the reference corpus and each of the five context corpora was determined through the use of Euclidean distance between the totality of characteristics as having undergone min-max normalisation. The five associated derivation methods were ranked based on their closeness to the reference corpus. As

<sup>18</sup>The same ones as referred to in section 3.1.

an additional experiment, the examined vocabulary items were divided into CEFR levels B1 (56 items) and B2 (44 items) and all textual characteristics were evaluated once again in an attempt to reveal the derivation methods' sensitivity to the CEFR level at hand.

The derived corpora were also ranked based on their pedagogical qualities, as estimated by the human evaluation. For this purpose, each corpus was given a percentage value representing the number of points received for all evaluated categories assembled ('ready to use,' 'comprehensible,' and 'in-domain') compared with the total number of points possible.

Ultimately, the two rankings were compared in an attempt to reveal a potential link between the contexts' linguistic and pedagogical qualities. It was assumed that the most highly rated method in the annotation process objectively has the highest pedagogical value.

## 4 Results

### 4.1 Automatic Evaluation

The corpus discovered to be globally closest to the reference one in terms of Euclidean distance based on all examined numeric textual characteristics is 'Mistral, few-shot' (3.82), followed by 'Gemini, few-shot' (4.86), 'Mistral, zero-shot' (4.99), 'Web-crawled' (5.46) and 'Gemini, zero-shot' (5.65). The 'Mistral, few-shot' model remains closest when the four categories of textual characteristics are considered separately, and the rest of the models mostly keep their place. The 'Mistral, zero-shot' model varies from second (for lexico-semantic and discourse-based characteristics) to fourth place (for length-based characteristics). The 'Web-crawled' corpus is closest to the reference in relation to length-based characteristics.

Table 1 shows a summary of the most relevant results of the corpus comparison based on atomic textual characteristics. For a comparison of all features, please refer to [Appendix 4: Detailed Results of the Comparison between Corpora based on Textual Features](#).

The reference corpus is generally associated with the highest ranges (i.e. distances between the maximal and minimal values) as well as the highest standard deviation for continuous characteristics. The 'Mistral, few-shot' corpus often comes closest to the reference in these aspects (e.g. in

relation to the number of words per sentence, the number of noun phrases per sentence, and the percentage of non-stem words per sentence).

The total number of words in the 'Mistral, few-shot' sample is closest to the reference and the only one larger. When length-based textual characteristics<sup>19</sup> as well as morphosyntactic characteristics<sup>20</sup> are considered, the 'Gemini, few-shot' corpus presents the least deviation from the reference. In the latter category, the 'Mistral, few-shot' corpus often comes closest to the reference, such as in terms of number of punctuation signs per sentence and the variety in end-of-sentence punctuation. The least statistical deviation when it comes to lexico-semantic characteristics is associated with the 'Mistral, zero-shot' corpus<sup>21</sup>. The most frequent words encountered in the 'Web-crawled' corpus strike as very generic and unrelated to the scientific domain compared to those in other corpora (e.g. 'would,' 'could,' 'said'). Within the 'Mistral, zero-shot' corpus, the personal pronoun 'I' is uniquely featured among the most frequent words when stop words are retained. Finally, discourse-related characteristics demonstrate little deviation from the reference, with the exception of those related to cosine distance, where statistical significance is smallest with the 'Web-crawled' sample. When subcorpora associated with CEFR level B1 are considered, the 'Mistral, few-shot' corpus demonstrates the lowest deviation from the reference corpus (only 4 features exhibiting statistical significance). In contrast, statistical significance is present in a minimum of 7 features for the others<sup>22</sup>. The two CEFR levels are also associated with different domains ('science' for B1 and 'agronomy' for B2), and this additional focus is reflected in the most used words for some of the corpora (e.g. the word 'scientists' is present for all LLM-based B1 subcorpora and the words 'crop' and 'soil' for the 'Gemini, zero-shot' and both Mistral B2 subcorpora).

### 4.2 Human Annotation

For a distribution of the values given to the corpora in the annotation in relation to the three character-

<sup>19</sup>The only (highly) significant deviation is for the average number of words per sentence.

<sup>20</sup>one significant deviation with moderate significance: the number of punctuation signs per sentence

<sup>21</sup>one instance of statistical significance of high value, for the number of proper nouns per sentence

<sup>22</sup>a number shared by the 'Web-crawled', 'Mistral, zero-shot' and 'Gemini, few-shot' corpora

Feature	Ref.	Web-crawled	Mistral, 0-shot	Mistral, f-shot	Gemini, 0-shot	Gemini, f-shot
words in sample	3787	2267	2823	<b>4091</b>	2345	2638
<i>words / sentence</i>	<i>13.33</i>	<i>11.11***</i>	<i>11.67***</i>	<b>13.73</b>	<i>11.17***</i>	<i>11.32***</i>
<i>letters / word</i>	<i>5.2</i>	<i>5.37</i>	<i>5.57***</i>	<b>5.4*</b>	<i>5.84***</i>	<i>5.21</i>
<i>noun phrases / sentence</i>	<i>5.76</i>	<i>6.34*</i>	<i>6.08</i>	<b>6.29**</b>	<i>6.26*</i>	<i>5.68</i>
<i>non-stem words / s-ce</i>	<i>31.91</i>	<i>34.2</i>	<i>38.06***</i>	<b>35.2**</b>	<i>40.09***</i>	<i>33.28</i>
<i>punctuation signs / s-ce</i>	<i>1.51</i>	<i>0.98*</i>	<i>1.06**</i>	<b>1.29</b>	<i>1.09</i>	<i>0.97**</i>
<i>verbs / sentence</i>	<i>2.45</i>	<i>2.92**</i>	<i>2.67</i>	<b>2.72*</b>	<i>2.72*</i>	<i>2.47</i>
<i>adj. and adv. / sentence</i>	<i>2.77</i>	<i>2.91</i>	<i>2.51</i>	<b>2.69</b>	<i>2.95</i>	<i>2.5</i>
<i>1st-person pron. / s-ce</i>	<i>0.11</i>	<i>0.01*</i>	<i>0.08</i>	<b>0.06</b>	<i>0.02*</i>	<i>0.02*</i>
<i>proper nouns / sentence</i>	<i>0.99</i>	<i>0.51</i>	<i>0.09***</i>	<b>0.32***</b>	<i>0.15***</i>	<i>0.23***</i>
hapax legomena	25.69	32.33	20.61	<b>19.3</b>	27.25	25.05
concreteness	2.48	2.42	2.46	<b>2.44</b>	2.37	2.4
<i>pronouns / sentence</i>	<i>0.95</i>	<i>0.64</i>	<i>0.87</i>	<b>0.88</b>	<i>0.66</i>	<i>0.73</i>
<i>anaphora words / s-ce</i>	<i>10.28</i>	<i>9.93</i>	<i>9.2</i>	<b>9.46</b>	<i>11.95</i>	<i>12.72</i>
<i>cos. distance btwn s-ces</i>	<i>0.12</i>	<i>0.1*</i>	<i>0.18***</i>	<b>0.15***</b>	<i>0.14***</i>	<i>0.14***</i>
Euclidean distance from ref.	-	5.46	4.99	<b>3.82</b>	5.65	4.86

Table 1: Comparison of the corpora based on a sample of textual features. The average values of continuous characteristics are indicated in *italics*, and the statistical significance of their divergence from the reference corpus is marked with \* (lowest), \*\* and \*\*\* (highest). The 'Mistral, few-shot' corpus is represented in **bold** to denote its highest global closeness to the reference.

istics, please refer to [Appendix 5: Distribution of Pedagogical Qualities per Corpus](#).

The corpus that is rated highest in the annotation process is 'Gemini, zero-shot', followed by 'Gemini, few-shot', 'Mistral, zero-shot', 'Mistral, few-shot' and 'Web-crawled' (see Table 2). The performance gap is largest between the web-crawled corpus (rated worst) and the second worst corpus, 'Mistral, few-shot', whilst the LLM-generated corpora exhibit higher similarity to one another. The figure in [Appendix 5: Distribution of Pedagogical Qualities per Corpus](#) clearly shows that the 'Web-crawled' corpus is the most frequent one to not receive the total number of point for all three investigated categories.

Interestingly, both corpora derived in zero-shot settings are rated more highly than their few-shot counterparts. The 'Gemini, few-shot' corpus is associated with the highest percentage of full points (71% of all contexts), followed by 'Gemini, zero-shot' (69%), 'Mistral, zero-shot' (61%), 'Mistral, few-shot' (60%) and 'Web-crawled' (29%). When the 'in-domain' characteristic is regarded in isolation, 'Gemini, few-shot' performs highest (by a small margin), and the rest of the classification remains the same. In turn, 'Mistral, few-shot' performs slightly better than 'Mistral, zero-shot' in

relation to the 'ready to use' characteristic. This is also the characteristic for which the models shows largest variance in terms of the attribution of the highest number of points (see [Appendix 5: Distribution of Pedagogical Qualities per Corpus](#)).

In the free text notes, Mistral-generated text was surprisingly judged to have negative qualities that were explicitly addressed during the generation and filtering process: contexts were judged as too long in 8 cases in the zero-shot setting and 5 in the few-shot setting, a definition or explanation was provided instead of or along with the context (6 vs 2 instances), the pronoun 'I' was mentioned to have been used extensively (in 4 vs 2 examples), and the target word was said to have been closely repeated in one example (in the zero-shot setting). Therefore, the robustness of the applied filters should be examined. Other problems linked with examples generated by the model include lack of clarity (4 vs 1 instance), lack of informativeness (3 instances in the zero-shot setting), scientifically unsound text (3 instances in the zero-shot setting) and different meanings of the target word addressed (2 instances in the zero-shot setting). Perceivably fewer problems are noted in relation to the few-shot setting. In contrast, the issues noted in relation to Gemini-generated text,

while smaller in number, are not clearly reduced by way of the few-shot setting. Some contexts are judged to be too long (2 vs 5 instances), too generic (4 vs 4 instances) or unclear (2 vs 1 instances). Also, definitions or explanations were featured (2 vs 2 instances), and target words were used with a different meaning to the intended one (1 vs 3 instances). Finally, web-crawled examples were criticised for including quotations (4 instances), containing textual processing mistakes (2 instances) and being unclear (2 instances).

## 5 Discussion

Human evaluation rates the 'Gemini, zero-shot' corpus highest, while automatic comparison ranks 'Mistral, few-shot' first. In the case of Mistral, the few-shot setting seems to be efficient in reducing problems that make contexts not directly applicable in a classroom setting. Thus, the different corpora and, by definition, the derivation methods behind them are associated with different qualities and drawbacks.

Table 3 shows a juxtaposition of the contexts derived through all five described methods for the same ESP vocabulary item. The only context that did not receive the maximal number of points in the annotation was the web-crawled one, which was evaluated as not being entirely ready to use. Possible reasons could be its beginning with 'and', instances of complex grammar ('and though', 'those cases that did occur'), and the use of the definite article ('the procedure') when the reference is unknown to the reader. The web-crawled context is the longest, the 'Gemini, few-shot' the shortest, and the other three display similar length (19-20 words), which is also equal or close to that of the reference context (20 words). In the 'Mistral, zero-shot', 'Gemini, zero-shot' and 'Gemini, few-shot' contexts, the target word appears very close to the sentence's beginning, which is not the case with the reference. One could assume, therefore, that the 'Mistral, few-shot' setting has benefited from the proposed professional examples. Another specificity in the latter is the presence of a named entity ('The Second Law of Motion'). In terms of qualitative characteristics, one can claim that the reference context is scientifically sound and can serve an interdisciplinary purpose, and the same can interestingly be said about the two zero-shot LLM settings, which offer surprisingly similar examples, implying at the

same time that the models' training suffices for a pedagogically apt formulation and that high similarity of output can be expected in the absence of narrow prompts and provided examples.

On the first research question, comparing web retrieval and generative AI, we observed that LLMs, when instructed using relevant prompt engineering and filtering techniques, are capable of providing contexts for the practice of ESP vocabulary that are evaluated by teaching experts as more pedagogically sound than counterparts retrieved from a corpus of scientific articles. In addition, examples of use generated by LLMs tend to share more textual characteristics with the ones hand-crafted by professionals. The second research question also receives a positive reply as a large number of automatically derived contexts (290 out of 500) score maximally in terms of their pedagogical qualities based on human evaluation. In particular, 435 contexts received the maximum Likert value for the 'ready to use' quality. Finally, no clear correlation can currently be established between automatically derived contexts' textual characteristics and their pedagogical qualities (research question 3), as the two methods led to fully different classifications of the derivation methods.

The presented experiments and analyses extend current findings pertaining to the ability of LLMs to generate pedagogical contexts for the learning of foreign language vocabulary (such as the ones exposed by Paddags et al. (2024)) through the exploration of the models' few-shot abilities and the juxtaposition of human-based (qualitative) and automatic (quantitative) evaluation.

## 6 Conclusion

In this study, we demonstrated that high-quality contexts for an ESP vocabulary list can be obtained through contemporary NLP methods, in particular via LLM-based generation with prompt engineering. A possible problem is the reduced range and standard deviations that are associated with the derived contexts' measurable textual characteristics, which in turn may relate to a limited textual variety. A simple mitigation method would be the application of a variety of LLMs and generation settings to different vocabulary items, as they show different degrees of variation and adaptability to instructions. Other future directions of improvement may include the further adaptation of textual filters, such as a mod-

Corpus	In-domain	Comprehensible	Ready to Use	Overall
Web-crawled	89.2%	88.4%	78.4%	85.33%
Mistral, zero-shot	<b>99.4%</b>	97.8%	87.0%	94.73%
Mistral, few-shot	97.0%	95.8%	87.6%	94.37%
Gemini, zero-shot	97.2%	<b>99.0%</b>	<b>94.4%</b>	<b>96.87%</b>
Gemini, few-shot	97.6%	98.6%	90.4%	95.53%

Table 2: Percentages given to the derived corpora based on the human annotation process (as a portion of the total number of points possible).

Corpus	Sample context	In-d.*	Compr.*	RTU*
Reference	The Second Law of Motion states that the <b>rate</b> of change of momentum is directly proportional to the force applied.	-	-	-
Web-crawled	And though the <b>rate</b> of deaths associated with the procedure remained statistically flat, those cases that did occur were found with older patients.	5	5	4
Mistral, zero-shot	The <b>rate</b> of photosynthesis in plants depends on many factors such as temperature, light intensity and carbon dioxide concentration.	5	5	5
Mistral, few-shot	When calculating population growth <b>rates</b> , scientists use statistics to estimate the number of births and deaths in a given region.	5	5	5
Gemini, zero-shot	The <b>rate</b> of photosynthesis is influenced by the intensity of light, the availability of carbon dioxide, and the temperature.	5	5	5
Gemini, few-shot	The <b>rate</b> at which the climate changes is affected by human activity.	5	5	5

Table 3: Contexts for the item 'rate' (CEFR level B1, domain 'science') from the reference corpus and all automatically derived ones as well as the points the latter received in the human annotation. \* Rating criteria: In-d. = In-domain; Compr. = Comprehensible; RTU = Ready to use.

ification of the permitted sentence lengths and domain filters that go beyond a binary classification of scientific vs non-scientific sentences. Finally, we are planning to make available a user-friendly online interface that facilitates the automatic generation of contexts based on selected ESP vocabulary items by teachers and students.

## References

- S. Aktay and G. G. D. Uzunoglu. 2023. Chatgpt in education: General applications as a dialogue agent and its impact on interaction, motivation, confidence, and vocabulary acquisition. *TAY Journal*, 7(2):378–406.
- Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. [CEFR-based sentence difficulty annotation and assessment](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yuliya Ardasheva, Tingting Hao, and Xiaobin Zhang. 2019. [Pedagogical implications of current SLA research for vocabulary skills](#). In Nihat Polat, Peter D. MacIntyre, and Tammy Gregersen, editors, *Research-driven pedagogy: Implications of L2A theory and research for the teaching of language skills*, pages 125–144. Routledge.
- Henri Besse. 1981. [The pedagogic authenticity of a text](#). In *The Teaching of Listening Comprehension. Papers presented at the Goethe Institut Colloquium held in Paris in 1979*, pages 20–29. British Council.
- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. [Concreteness ratings for 40 thousand generally known English word lemmas](#). *Behavior Research Methods*, 46(3):904–911.
- Scott A. Crossley, Max M. Louwerse, Philip M. McCarthy, and Danielle S. McNamara. 2007. [A linguistic analysis of simplified and authentic texts](#). *The Modern Language Journal*, 91(1):15–30.
- Robert Godwin-Jones. 2018. Evolving views on vocabulary development. *Language Learning & Technology*, 22(3):1–19.
- Bahman Gorjian, Seyed Rahim Moosavinia, Kamal Elahi Kavari, Parsa Asgari, and Abouzar Hydareh. 2011. [The impact of asynchronous computer-assisted language learning approaches on english as a foreign language high and low achievers' vocabulary retention and recall](#). *Computer Assisted Language Learning*, 24(5):383–391.
- Michael Heilman, Le Zhao, Juan Pino, and Maxine Eskenazi. 2008. [Retrieval of reading materials for vocabulary and reading practice](#). In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 80–88, Columbus, Ohio. Association for Computational Linguistics.
- Thomas Huckin and James Coady. 1999. [Incidental vocabulary acquisition in a second language: a review](#). *Studies in Second Language Acquisition*, 21(2):181–193.
- Anealka Hussin, Yuen Fook Chan, and Zubaidah Aliree. 2010. [Scientific structural changes within texts of adapted reading materials](#). *English Language Teaching*, 3.
- Jaeho Jeon and Seongyong Lee. 2023. [Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT](#). *Education and Information Technologies*, 28(12):15873–15892.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth ee Lacroix, and William El Sayed. 2023. [Mistral 7B](#).
- Diane J. Litman. 2016. [Natural language processing for enhancing teaching and learning](#). In *AAAI Conference on Artificial Intelligence*.
- David Little, Se an Devitt, and David Singleton. 1989. *Learning Foreign Languages from Authentic Texts: Theory and Practice*. Authentik.
- Mathieu Loiseau, Georges Antoniadis, and Claude Ponton. 2005. [Pedagogical text indexation and exploitation for language learning](#). In *Third international conference on multimedia and information and communication technologies in education (mICTE2005)*, volume 3 of *Recent Research Developments in Learning Technologies*, pages 984–994, Seville, Spain. Formatex.
- Yihan Lou. 2023. Exploring the application of ChatGPT to english teaching in a malaysian primary school. *Journal of Advanced Research in Education*, 2(4):47–54.
- Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. [Enhancing authentic web pages for language learners](#). In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18, Los Angeles, California. Association for Computational Linguistics.
- Ana M. Morais and Isabel P. Neves. 2010. [Educational texts and contexts that work discussing the optimization of a model of pedagogic practice](#). In

- Daniel Frandji and Philippe Vitale, editors, *Knowledge, Pedagogy and Society: International Perspectives on Basil Bernstein's Sociology of Education*, page 18. Routledge, London.
- Iglika Nikolova-Stoupak, Serge Bibauw, Amandine Dumont, Françoise Stas, Patrick Watrin, and Thomas François. 2024. [LLM-generated contexts to practice specialised vocabulary: Corpus presentation and comparison](#). In *Proceedings of TALN 2024*, Toulouse, France.
- Benjamin Paddags, Daniel Hershovich, and Valkyrie Savage. 2024. [Automated sentence generation for a spaced repetition software](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 351–364, Mexico City, Mexico. Association for Computational Linguistics.
- Robert Gilmore Pontius Jr and Marco Millones. 2011. [Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment](#). *International Journal of Remote Sensing*, 32(15):4407–4429.
- Falcon Dario Restrepo Ramos. 2015. [Incidental vocabulary learning in second language acquisition: A literature review](#). *Profile: Issues in Teachers' Professional Development*, 17(1):157–166.
- Irina Rets and Jekaterina Rogaten. 2021. [To simplify or not? Facilitating English L2 users' comprehension and processing of open educational resources in English using text simplification](#). *Journal of Computer Assisted Learning*, 37(3):705–717.
- Angeliki Salamoura and Nick Saville. 2010. [Exemplifying the CEFR: criterial features of written learner English from the english profile programme](#). In Inge Bartning, Maisa Martin, and Ineke Vedder, editors, *Communicative proficiency and linguistic development: intersections between SLA and language testing research*, number 1 in Eurosla Monographs Series, pages 101–132. Eurosla.
- Sarang Shaikh, Sule Yildirim Yayilgan, Blanka Klimova, and Marcel Pikhart. 2023. [Assessing the usability of ChatGPT for formal English language learning](#). *European Journal of Investigation in Health, Psychology and Education*, 13(9):1937–1960.
- Try Siregar and Widyastuti Purbani. 2024. [Prominent linguistic features of pedagogical texts to provide consideration for authentic text simplification](#). *Studies in English Language and Education*, 11:321–342.
- Saifon Songsiangchai, Bank-on Sereerat, and Wirot Watananimitgul. 2023. [Leveraging artificial intelligence \(ai\): Chatgpt for effective english language learning among thai students](#). *English Language Teaching*, 16(11):1–68.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, ..., and Oriol Vinyals. 2024. [Gemini: A family of highly capable multimodal models](#).
- Brian Tomlinson. 2012. [Materials development for language learning and teaching](#). *Language Teaching*, 45(2):143–179.
- Julio Christian Young and Makoto Shishido. 2023a. [Evaluation of the potential usage of ChatGPT for providing easier reading materials for ESL students](#). In *Proceedings of EdMedia and Innovate Learning*, pages 155–162. AACE.
- Julio Christian Young and Makoto Shishido. 2023b. [Investigating OpenAI's ChatGPT potentials in generating chatbot's dialogue for English as a foreign language learning](#). *International Journal of Advanced Computer Science and Applications*, 14(6).
- Luo Yunjiu, Wei Wei, and Ying Zheng. 2022. [Artificial intelligence-generated and human expert-designed vocabulary tests: A comparative study](#). *SAGE Open*, 12(1):21582440221082130.



## Appendix 1: List of Crawled Websites

[https://climate.ec.europa.eu/climate-change\\_en](https://climate.ec.europa.eu/climate-change_en)  
[https://climate.ec.europa.eu/eu-action\\_en](https://climate.ec.europa.eu/eu-action_en)  
[https://climate.ec.europa.eu/index\\_en](https://climate.ec.europa.eu/index_en)  
<https://climate.nasa.gov/>  
<https://engineeringdiscoveries.com/>  
<https://newatlas.com>  
<https://sciencedemonstrations.fas.harvard.edu/>  
<https://sustainability.stanford.edu/>  
<https://world-nuclear.org>  
<https://www.advancedsciencenews.com>  
<https://www.computerworld.com/>  
<https://www.eurekalert.org/>  
<https://www.green.earth/>  
<https://www.iea.org/>  
<https://www.ipcc.ch>  
<https://www.livescience.com/>  
<https://www.nationalgeographic.org/society/>  
<https://www.nature.com/>  
<https://www.ncbi.nlm.nih.gov/>  
<https://www.networkworld.com/>  
<https://www.newscientist.com/>  
<https://www.npr.org/sections/science/>  
<https://www.pcworld.com>  
<https://www.pewresearch.org/topic/internet-technology/>  
<https://www.pewresearch.org/topic/science/>  
<https://www.popularmechanics.com/>  
<https://www.science.org/>  
<https://www.sciencealert.com/>  
<https://www.sciencedaily.com/>  
<https://www.scienceopen.com/>  
<https://www.scientificamerican.com>

<https://www.triplepundit.com/>

<https://www.un.org/en/>

<https://www.un.org/en/climatechange>

<https://www.usgs.gov/programs/earthquake-hazards/>

<https://www.wwf.org.uk>

## Appendix 2: Prompts used for LLM Generation

### Zero-shot setting:

Here is a sentence<sup>23</sup> at CEFR level  $\{level\}$  showing how you use the  $\{pos\}$  if verb/noun/adverb/adjective; else 'word' or 'expression'  $\{item\}$  in the domain of  $\{domain\}$  ( $\{lower^{24}\}$ - $\{upper\}$  words):

### Few-shot setting, level B1:

Please provide an example at level **B1** showing how you use the  $\{pos\}$  ' $\{item\}$ ' ( $\{domain\}$ ). Please use between  $\{lower\}$  and  $\{upper\}$  words.

Examples:

the adjective 'scarce': "As the planet continues to warm, resources such as freshwater, land, and food are becoming increasingly scarce."

the noun 'poaching': "As rhino populations decline rapidly due to habitat loss and poaching, the challenges for conservationists to protect these endangered species have never been more important."

the noun 'rate': "The carbon cycle is a complex process, and changes in land use and deforestation can affect the rate at which carbon is exchanged between the atmosphere and terrestrial ecosystems."

the verb 'reclaim': "The Great Green Wall is both an initiative for ecological restoration, and part of the fight against hunger and food insecurity in Africa. In existence since 2007, the wall is above all part of an immense effort to reclaim land lost to desertification."

the noun 'strain': "Before an earthquake occurs, tectonic plates accumulate strain along fault lines, gradually building up stress until it is released in a sudden rupture."

### Few-shot setting, level B2:

Please provide an example at level **B2** showing how you use the  $\{pos\}$  ' $\{item\}$ ' ( $\{domain\}$ ). Please use between  $\{lower\}$  and  $\{upper\}$  words.

Examples:

the noun 'spore': "Why some mushrooms are bioluminescent remains uncertain, but a study using LED

<sup>23</sup>The reason for 'sentence' to be used rather than 'example', even though some of the gold standard examples consist of more than a single sentence, is that using 'example' tends to result in the rendition of extensive explanations instead of or in addition to an example of use. This problem does not persist with the few-shot setting, for which the word 'example' is used instead

<sup>24</sup>'Lower' and 'upper' denote a range of example lengths, which differs for the different CEFR levels (8 to 43 words for B1 and 20 to 87 words for B2). The ranges are defined as  $\pm 1.5$  standard deviations from the average value per level. This value as well as the addition of information about length itself was decided upon following a process of trial and error based on the behaviour of 20 sample examples in comparison to the reference's counterparts.

lights adds to the evidence they attract insects that help the fungus disperse its spores.”

the adjective 'bulbous': "Most of the evidence comes from soil fungi, many of which spend much of their life cycle as microorganisms, but also produce the bulbous fruiting bodies we know as mushrooms, toadstools, bracket fungi and the like. These are easy enough to spot, so they are often used as surrogates for the state of forest biodiversity, especially of the underground mycorrhizae – fungi that form symbiotic relationships with tree roots, taking sugars and supplying plants with water and mineral nutrients in return.”

the noun 'shrub': "More recently, botanists in Brazil discovered six previously unknown species of fungus growing on the leaves of a tropical shrub, *Coussapoa floccosa*, which until recently was thought to be extinct. If and when the last specimen dies, those fungi will disappear too.”

the verb 'undergo': "Nearly three-quarters of hammer coral colonies annually alternate between male and female. They are the only animal species known to undergo this change on such a regular schedule.”

the noun 'brood': "Two species of bird have been observed raising offspring together. Such cooperative breeding between different species has never been documented before, says Rosario Balestrieri at the Stazione Zoologica Anton Dohrn of Naples, Italy. "It is a very strange and rare situation, in which the brood is mixed between the two species," he says.”

### Appendix 3: Features Used in Corpus Comparison

Length-Based	total number of examples in the sample total number of words in the sample <i>average/min/max/SD number of words per sentence</i> average/min/max/SD number of syllables per sentence <i>average/min/max/SD number of letters per word</i> average/min/max/SD number of syllables per word
Morphosyntactic	average/min/max/SD number of noun phrases per sentence <i>average/min/max/SD percentage of non-stem words per s-ce</i> percentage of sentences ending in question mark percentage of sentences ending in exclamation mark <i>average/min/max/SD number of punctuation signs per s-ce (excluding end-of-s-ce punct.)</i> morphological richness
Lexico-Semantic	<i>average/min/max/SD number of verbs per sentence</i> <i>average/min/max/SD number of adj. and adv. per s-ce</i> <i>average/min/max/SD number of 1st-person pronouns per s-ce</i> <i>average/min/max/SD number of proper nouns per sentence</i> percentage of words not present in the Dale-Chall list percentage of hapax legomena type-to-token ratio (word-based) type-to-token ratio (lemma-based) average concreteness (as per Brysbaert et al. (2014)'s list of 40k English lemmas) 10 most frequent words (excluding stop words) 10 most frequent words (including stop words)
Discourse-Related	<i>average/min/max/SD number of pronouns per sentence</i> <i>average/min/max/SD % of anaphora-denoting words per sentence</i> <i>average/min/max/SD cosine distance between sentences</i>

Table 4: Description of the linguistic features used in corpus comparison. The features marked in *italics* are representative continuous ones used in filters at the automatic derivation of contexts.

## **Appendix 4: Detailed Results of the Comparison between Corpora based on Textual Features**

## Entire Sample

Feature	Reference	Web-Crawled	Mistral: zero-shot	Mistral: few-shot	Gemini: zero-shot	Gemini: few-shot
Total # examples in sample	100	100	100	<b>100</b>	100	100
Total # words in sample	3787	2267	2823	<b>4091</b>	2345	2638
Avg. # words / s-ce	13.33	11.11***	11.67***	<b>13.73</b>	11.17***	11.32***
Min.	4	1	9	<b>2</b>	10	8
Max.	55	33	34	<b>44</b>	32	34
SD	8.48	4.82	5.08	<b>5.79</b>	5.34	6.08
Avg. # syllables / s-ce	20.76	18.25***	19.99*	<b>22.54*</b>	19.89***	17.94
Min.	6	1	14	<b>4</b>	15	14
Max.	85	60	63	<b>64</b>	62	56
SD	14.29	9.39	9.53	<b>9.86</b>	10.53	10.05
Avg. # letters / word	5.2	5.37	5.57***	<b>5.4*</b>	5.84***	5.21
Min.	1	1	1	<b>1</b>	1	1
Max.	18	23	19	<b>16</b>	22	17
SD	2.8	3.02	3.06	<b>2.93</b>	3.2	2.87
Avg. # syllables / word	1.56	1.64***	1.71***	<b>1.64***</b>	1.78***	1.58
Min.	0	0	1	<b>1</b>	1	0
Max.	7	9	6	<b>5</b>	6	6
SD	0.88	0.99	1.0	<b>0.93</b>	1.04	0.92
Avg. # noun phrases / s-ce	5.76	6.34*	6.08	<b>6.29**</b>	6.26*	5.68
Min.	1	0	3	<b>1</b>	2	2
Max.	16	11	11	<b>14</b>	11	11
SD	2.75	1.95	1.85	<b>2.08</b>	1.93	1.96
Avg. % non-stem words / s-ce	31.91	34.2	38.06***	<b>35.2**</b>	40.09***	33.28
Min.	5.88	0.0	11.11	<b>7.14</b>	17.39	7.14
Max.	61.54	63.64	66.67	<b>72.22</b>	69.23	54.55
SD	10.69	11.26	9.28	<b>10.42</b>	10.1	9.07
% s-ces ending in “?”	0.54	0.0	0.0	<b>0.51</b>	0.0	0.0
% s-ces ending in “!”	0.54	0.0	0.0	<b>0.0</b>	0.0	0.0
Avg. # punct. signs / s-ce	1.51	0.98*	1.06**	<b>1.29</b>	1.09	0.97**
Min.	0	0	0	<b>0</b>	0	0
Max.	6	2	4	<b>6</b>	4	4
SD	0.54	0.35	0.43	<b>0.49</b>	0.4	0.4
Morphological richness	0.02	0.02	0.02	<b>0.02</b>	0.02	0.02
Avg. # verbs / s-ce	2.45	2.92***	2.67	<b>2.72*</b>	2.72*	2.47
Min.	0	0	0	<b>0</b>	0	0
Max.	7	5	8	<b>7</b>	6	6
SD	1.51	1.04	1.38	<b>1.36</b>	1.13	1.17
Avg. # adj. and adv. / s-ce	2.77	2.91	2.51	<b>2.69</b>	2.95	2.5
Min.	0	0	0	<b>0</b>	0	0
Max.	8	6	7	<b>7</b>	7	7
SD	1.83	1.56	1.47	<b>1.49</b>	1.57	1.5
Avg. # 1st-person pron. / s-ce	0.11	0.01*	0.08	<b>0.06</b>	0.02*	0.02*
Min.	0	0	0	<b>0</b>	0	0
Max.	2	1	1	<b>1</b>	2	1
SD	0.43	0.1	0.28	<b>0.23</b>	0.19	0.12
Avg. # proper nouns / s-ce	0.99	0.51	0.09***	<b>0.32***</b>	0.15***	0.23***
Min.	0	0	0	<b>0</b>	0	0
Max.	11	4	2	<b>6</b>	2	3
SD	1.96	0.8	0.33	<b>0.87</b>	0.45	0.61

% words not in Dale-Chall list	44.71	46.18	46.09	<b>45.61</b>	50.36	43.48
% hapax legomena	25.69	32.33	20.61	<b>19.3</b>	27.25	25.05
Type-to-token ratio (words)	0.37	0.45	0.32	<b>0.31</b>	0.39	0.36
Type-to-token ratio (lemmas)	0.35	0.42	0.3	<b>0.29</b>	0.37	0.34
Average concreteness	2.48	2.42	2.46	<b>2.44</b>	2.37	2.4
10 most frequent words (excl. stop words)	water, climate, change, species, world, people, new, plants, global, could	would, could, said, people, water, also, international, may, must, new	crop, soil, crops, farmers, scientists, agriculture, water, yields, new, order	<b>soil, crop, farmers, agriculture, crops, yields, practices, water, agricultural, climate</b>	crop, soil, new, scientists, farmers, crops, practices, yields, scientist, sustainable	water, soil, farmers, crops, crop, plant, species, food, practices, yields
10 most frequent words (incl. stop words)	the, of, and, to, in, a, is, that, are, for	the, of, and, to, in, a, that, for, be, is	the, to, of, and, in, a, that, I, for, can	<b>to, the, of, and, in, a, that, is, can, as</b>	the, of, to, a, and, in, for, crop, soil	the, of, to, and, in, a, for, is, that, are
Avg # pron. / s-ce	0.95	0.64	0.87	<b>0.88</b>	0.66	0.73
Min.	0	0	0	<b>0</b>	0	0
Max.	4	3	3	<b>3</b>	3	3
SD	1.1	0.7	0.79	<b>0.94</b>	0.75	0.76
Avg.% anaph. words / s-ce	10.28	9.93	9.2	<b>9.46</b>	11.95	12.72
Min.	0.0	0.0	0.0	<b>0.0</b>	3.23	0.0
Max.	27.27	22.73	23.81	<b>30.77</b>	25.0	25.0
SD	6.08	5.77	5.8	<b>5.99</b>	5.52	5.51
Avg.cos.d-ce btwn s-ces	0.12	0.10*	0.18***	<b>0.15***</b>	0.14***	0.14***
Min.	-0.18	-0.19	-0.17	<b>-0.17</b>	-0.20	-0.20
Max.	0.70	1.0	0.80	<b>0.79</b>	0.84	0.84
SD	0.12	0.11	0.16	<b>0.14</b>	0.15	0.13

Per Level: B1 (domain ‘Agronomy’)

Feature	Reference	Web-Crawled	Mistral: zero-shot	Mistral: few-shot	Gemini: zero-shot	Gemini: few-shot
Total # examples in sample	56	56	56	56	56	56
Total # words in sample	1382	1179	1030	1581	971	892
Avg. # words / s-ce	10.63	10.34***	8.8*	11.21	8.67***	7.82***
Min.	4	1	11	2	10	8
Max.	41	27	26	31	26	25
SD	7.49	4.63	3.34	5.06	3.36	4.32
Avg. # syllables / s-ce	16.65	16.88***	14.93	18.57	15.36	12.45*
Min.	6	1	14	4	15	14
Max.	76	60	43	54	48	44
SD	12.69	9.34	6.47	8.58	7.22	7.54
Avg. # letters / word	5.19	5.34	5.41	5.4	5.8***	5.26
Min.	1	1	1	1	1	1
Max.	16	23	15	15	22	17
SD	2.82	3.11	3.01	2.98	3.25	2.85
Avg. # syllables / word	1.57	1.63	1.7**	1.66*	1.77***	1.59
Min.	0	0	1	1	1	1
Max.	7	9	5	5	6	6
SD	0.89	1.01	1.0	0.94	1.04	0.91
Avg. # noun phrases / s-ce	5.08	5.75*	5.15	5.62*	5.05	4.41
Min.	1	0	3	1	2	2
Max.	12	10	9	9	9	8
SD	2.38	1.73	1.28	1.65	1.41	1.24
Avg. % non-stem words / s-ce	33.07	34.38	37.04*	35.49	39.57***	34.56
Min.	10.0	0.0	16.67	14.29	17.39	15.79
Max.	61.54	61.54	56.25	53.85	66.67	54.55
SD	10.2	10.69	8.73	9.2	9.89	9.54
% s-ces ending in “?”	0.0	0.0	0.0	1.16	0.0	0.0
% s-ces ending in “!”	1.33	0.0	0.0	0.0	0.0	0.0
Avg. # punct. signs / s-ce	1.16	0.86	0.57*	1.01	0.77	0.55*
Min.	0	0	0	0	0	0
Max.	6	2	3	3	2	2
SD	0.49	0.32	0.28	0.38	0.29	0.25
Morphological richness	0.02	0.01	0.02	0.02	0.02	0.02
Avg. # verbs / s-ce	2.15	2.67**	2.2	2.26	2.2	1.97
Min.	0	0	1	0	0	0
Max.	6	4	4	5	4	4
SD	1.33	1.06	0.96	1.08	0.96	1.03
Avg. # adj. and adv. / s-ce	2.48	2.75	2.15	2.43	2.27	2.09
Min.	0	0	0	0	0	0
Max.	6	6	6	6	5	5
SD	1.56	1.67	1.44	1.39	1.14	1.22
Avg. # 1st-person pron. / s-ce	0.15	0.02	0.15	0.13	0.0*	0.03
Min.	0	0	0	0	0	0
Max.	2	1	1	1	0	1
SD	0.51	0.13	0.36	0.34	0.0	0.18
Avg. # proper nouns / s-ce	0.77	0.49	0.15***	0.17***	0.14***	0.19**
Min.	0	0	0	0	0	0
Max.	6	3	2	2	2	3
SD	1.28	0.73	0.4	0.47	0.4	0.51

% words not in Dale-Chall list	44.21	44.7	43.79	45.86	49.02	43.83
% hapax legomena	34.9	38.59	32.06	29.77	38.35	37.51
Type-to-token ratio (words)	0.46	0.5	0.44	0.42	0.49	0.48
Type-to-token ratio (lemmas)	0.45	0.49	0.42	0.4	0.47	0.46
Average concreteness	2.5	2.37	2.4	2.39	2.33	2.43
10 most frequent words (excl. stop words)	climate, water, change, earth, new, tempera- ture, world, plants, two, greenhouse	development, may, next, many, eu- ropean, climate, resources, would, human, pos- sible	scientists, temperature, climate, change, growth, chemical, used, new, effects, plant	scientists, change, use, due, climate, sci- ence, world, around, uni- verse, new	scientists, research, experiment, new, study, researchers, temperature, significant, scientist, effects	scientists, climate, due, change, study, used, light, energy, earth, human
10 most frequent words (incl. stop words)	the, of, to, in, and, is, a, on, are, that	the, to, of, and, a, in, that, is, will, be	the, to, of, and, in, a, that, scien- tists, for, can	the, to, of, and, in, is, a, that, for, sci- entists	the, of, to, and, in, a, that, is, sci- entists, can	the, of, to, in, is, and, scientists, a, can, are
<i>Avg # pron. / s-ce</i>	0.8	0.61	0.8	0.86	0.52	0.55
Min.	0	0	0	0	0	0
Max.	4	2	3	3	2	2
SD	1.1	0.59	0.79	0.84	0.6	0.6
<i>Avg.% anaph. words / s-ce</i>	10.01	11.34*	11.42	9.7	13.65*	13.05
Min.	0.0	0.0	3.85	0.0	4.35	4.55
Max.	27.27	22.73	22.22	30.77	25.0	25.0
SD	6.86	5.4	5.38	5.65	5.99	5.25
<i>Avg.cos.d-ce btwn s-ces</i>	0.118	0.115**	0.151***	0.138***	0.136***	0.137***
Min.	-0.185	-0.145	-0.166	-0.167	-0.171	-0.200
Max.	0.701	1.000	0.765	0.751	0.584	0.758
SD	0.126	0.105	0.129	0.125	0.125	0.129



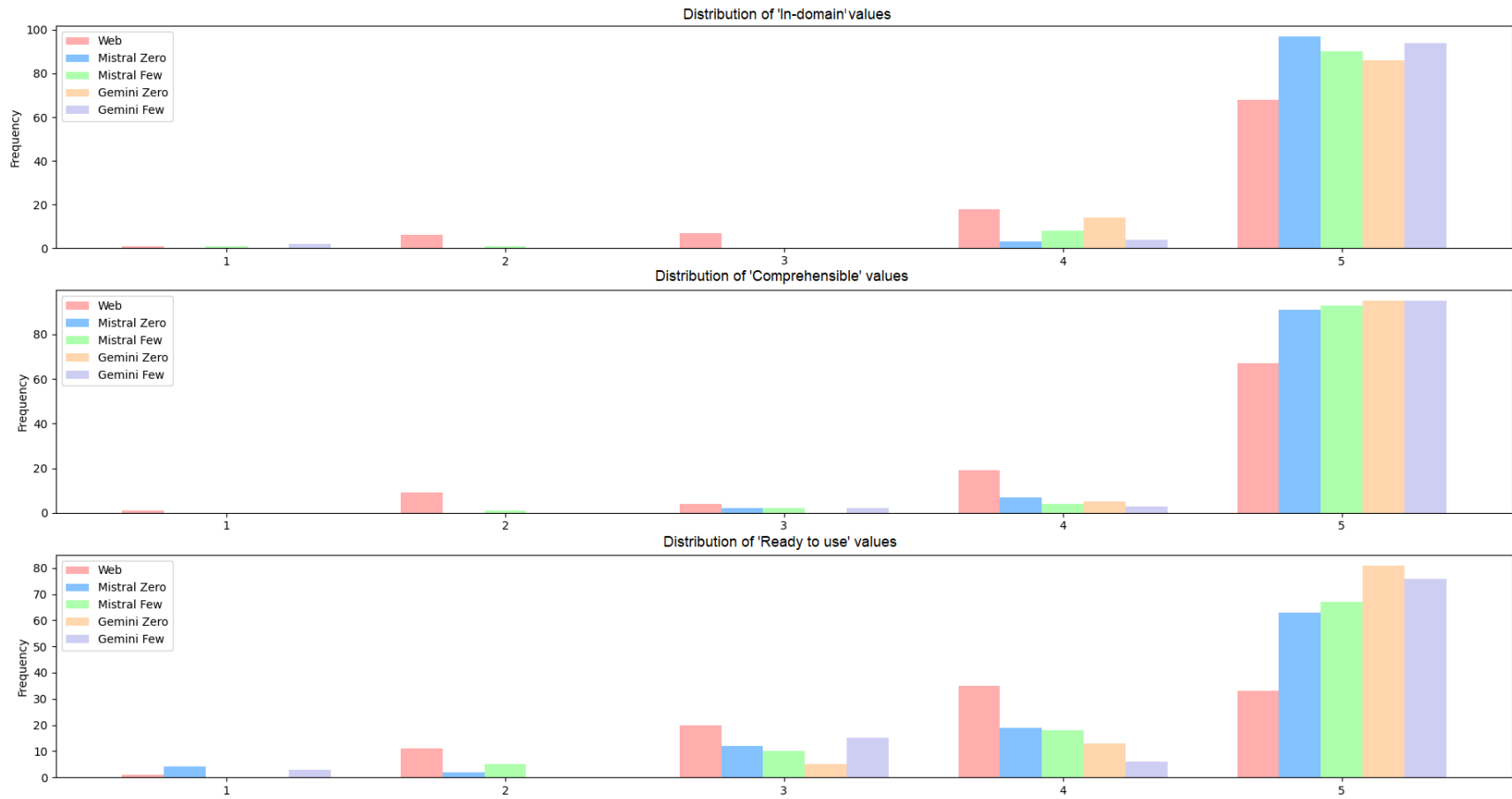
Per Level: B2 (domain ‘Science’)

Feature	Reference	Web-Crawled	Mistral: zero-shot	Mistral: few-shot	Gemini: zero-shot	Gemini: few-shot
Total # examples in sample	44	44	44	44	44	44
Total # words in sample	2405	1088	1793	2510	1374	1746
Avg. # words / s-ce	15.62	12.09***	14.34***	15.99	14.02***	14.67***
Min.	5	15	9	8	18	12
Max.	55	33	34	44	32	34
SD	8.85	4.09	5.02	5.74	3.59	4.9
Avg. # syllables / s-ce	24.23	19.98***	24.73**	26.11*	25.06***	23.19*
Min.	10	26	16	14	29	17
Max.	85	55	63	64	62	56
SD	15.0	7.89	9.46	9.98	7.8	8.31
Avg. # letters / word	5.21	5.39	5.67***	5.41*	5.87***	5.18
Min.	1	1	1	1	1	1
Max.	18	15	19	16	17	16
SD	2.79	2.94	3.09	2.91	3.16	2.89
Avg. # syllables / word	1.55	1.65***	1.72***	1.63*	1.79***	1.58
Min.	0	0	1	1	1	0
Max.	6	6	6	5	6	5
SD	0.87	0.96	0.99	0.92	1.04	0.93
Avg. # noun phrases / s-ce	6.22	7.09*	6.79	6.8	7.52***	6.65
Min.	1	3	3	2	4	2
Max.	16	11	11	14	11	11
SD	2.89	1.96	1.91	2.23	1.56	1.85
Avg. % non-stem words / s-ce	31.25	34.01	38.63***	35.01*	40.47***	32.62
Min.	5.88	9.09	11.11	7.14	25.0	7.14
Max.	58.06	63.64	66.67	72.22	69.23	53.85
SD	10.97	12.09	9.65	11.29	10.36	8.69
% s-ces ending in “?”	0.91	0.0	0.0	0.0	0.0	0.0
% s-ces ending in “!”	0.0	0.0	0.0	0.0	0.0	0.0
Avg. # punct. signs / s-ce	1.75	1.14*	1.42	1.51	1.43	1.29*
Min.	0	0	0	0	0	0
Max.	4	2	4	6	4	4
SD	0.57	0.39	0.51	0.55	0.49	0.48
Morphological richness	0.02	0.02	0.02	0.02	0.02	0.02
Avg. # verbs / s-ce	2.66	3.25**	3.02	3.07*	3.26**	2.85
Min.	0	1	0	0	1	0
Max.	7	5	8	7	6	6
SD	1.59	0.92	1.54	1.44	1.05	1.14
Avg. # adj. and adv. / s-ce	2.96	3.11	2.78	2.88	3.65**	2.83
Min.	0	1	0	0	0	0
Max.	8	5	7	7	7	7
SD	1.98	1.4	1.44	1.53	1.65	1.61
Avg. # 1st-person pron. / s-ce	0.09	0.0	0.04	0.0**	0.04	0.0*
Min.	0	0	0	0	0	0
Max.	2	0	1	0	2	0
SD	0.37	0.0	0.19	0.0	0.27	0.0
Avg. # proper nouns / s-ce	1.14	0.55	0.05***	0.43*	0.17**	0.25**
Min.	0	0	0	0	0	0
Max.	11	4	2	6	2	3
SD	2.31	0.87	0.27	1.07	0.5	0.68

% words not in Dale-Chall list	44.99	47.79	47.41	45.46	51.31	43.3
% hapax legomena	29.88	40.86	20.83	22.35	29.53	28.07
Type-to-token ratio (words)	0.42	0.53	0.33	0.34	0.41	0.39
Type-to-token ratio (lemmas)	0.4	0.51	0.31	0.32	0.39	0.37
Average concreteness	2.47	2.47	2.5	2.48	2.4	2.38
10 most frequent words (excl. stop words)	water, species, trees, carbon, could, wild, forests, researchers, reef, world	could, development, climate, change, international, benefits, health, people, responsibility, study	soil, crop, agriculture, farmers, crops, practices, use, sustainable, farming, growth	soil, crop, farmers, crops, water, use, growth, used, agriculture, levels	soil, crop, crop, soil, water, agricultural, farmers, growth, drought, conditions, practices, yields	species, soil, plant, plants, fungi, new, crop, crops, water, nutrients
10 most frequent words (incl. stop words)	the, of, and, to, in, a, is, that, are, for	to, the, and, of, a, in, with, that, for, or	and, the, to, of, soil, can, in, crop, a, agriculture	and, the, to, of, soil, can, in, crop, a, for	the, of, and, to, for, in, a, crop, as, their	the, of, and, to, a, in, that, is, for, as
<i>Avg # pron. / s-ce</i>	<i>1.05</i>	<i>0.68</i>	<i>0.91</i>	<i>0.9</i>	<i>0.81</i>	<i>0.87</i>
Min.	0	0	0	0	0	0
Max.	4	3	3	3	3	3
SD	1.1	0.83	0.79	1.0	0.85	0.84
<i>Avg.% anaph. words / s-ce</i>	<i>10.47</i>	<i>8.1</i>	<i>7.53</i>	<i>9.27</i>	<i>10.19</i>	<i>12.46**</i>
Min.	0.0	0.0	0.0	0.0	3.23	0.0
Max.	25.0	20.69	23.81	27.78	20.69	25.0
SD	5.52	5.78	5.57	6.26	4.38	5.72
<i>Avg.cos.d-ce btwn s-ces</i>	<i>0.14</i>	<i>0.11***</i>	<i>0.38***</i>	<i>0.24***</i>	<i>0.32***</i>	<i>0.24***</i>
Min.	-0.15	-0.17	-0.01	-0.10	-0.10	-0.08
Max.	0.63	1.00	0.80	0.79	0.84	0.84
SD	0.12	0.12	0.14	0.16	0.14	0.13

Features in *italics* have been tested for statistical significance, and the extent of the significance is marked with \*, \*\* and \*\*\* from lowest to highest. The few-shot Mistral corpus is marked with **bold** when the entire corpora are considered to denote its highest global similarity to the reference corpus.

## Appendix 5: Distribution of Pedagogical Qualities per Corpus



# Automatic Text Simplification with LLMs: A Comparative Study in Italian for Children with Language Disorders

Francesca Padovani<sup>1,2</sup>, Caterina Marchesi<sup>4</sup>, Eleonora Pasqua<sup>4,3</sup>, Martina Galletti<sup>1,3</sup> & Daniele Nardi<sup>3,5</sup>

<sup>1</sup>Sony Computer Science Laboratories - Paris, France

<sup>2</sup> University of Trento, Italy

<sup>3</sup> Sapienza University of Rome, Italy

<sup>4</sup>Centro Ricerca e Cura di Roma - Italy

<sup>5</sup>CINI-AIIS - Italy

francesca.padovani98@gmail.com, martina.galletti@sony.com

## Abstract

Text simplification aims to improve the readability of a text while maintaining its original meaning. Despite significant advancements in Automatic Text Simplification, particularly in English, other languages like Italian have received less attention due to limited high-quality data. Moreover, most Automatic Text Simplification systems produce a unique output, overlooking the potential benefits of customizing text to meet specific cognitive and linguistic requirements. These challenges hinder the integration of current Automatic Text Simplification systems into Computer-Assisted Language Learning environments or classrooms. This article presents a multifaceted output that highlights the potential of Automatic Text Simplification for Computer-Assisted Language Learning. First, we curated an enriched corpus of parallel complex-simple sentences in Italian. Second, we fine-tuned a transformer-based encoder-decoder model for sentences simplification. Third, we parameterized grammatical text features to facilitate adaptive simplifications tailored to specific target populations, achieving state-of-the-art results, with a SARI score of 60.12. Lastly, we conducted automatic and manual qualitative and quantitative evaluations to compare the performance of ChatGPT-3.5, and our fine-tuned transformer model. By demonstrating enhanced adaptability and performance through tailored simplifications in Italian, our findings underscore the pivotal role of ATS in Computer-Assisted Language Learning methodologies.

## 1 Introduction

The increasing access of digital information underscores the critical need to ensure universal access to knowledge, regardless of individuals' literacy levels or backgrounds. Automatic Text Simplification (ATS) is the Natural Language Processing (NLP) task aimed at reducing linguistic complexity of texts, while preserving their original meaning (Bott and Saggion, 2014; Shardlow, 2014b). ATS emerges as a promising solution to enhance text accessibility and readability, aiming to transform complex texts into versions that are more comprehensible, thus holding significant potential for fostering communication across diverse audiences and addressing gaps in information accessibility (Štajner, 2021). In recent years, research on ATS has focused on developing approaches to make texts simplified adapted for individuals facing cognitive disabilities or language impairment (Bott and Saggion, 2014; Rello et al., 2013; Aluisio et al., 2010). This development could have a significant impact on computer-assisted language learning (CALL), where adaptive learning technologies can personalize instruction based on individual learner progress and needs, ensuring a tailored and effective educational experience.

The emergence of large language models (LLMs) has significantly advanced automatic text simplification, among other NLP tasks. While their success in many benchmarks and challenges has been demonstrated (Anschütz et al., 2023; Sun et al., 2023; Engelmann et al., 2023; Shaib et al., 2023), it is imperative to ensure that the outputs of these models are truly suitable, especially before deployment in sensitive domains such as education or health (Kasneji et al., 2023). Fur-

---

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

thermore, there is limited research being conducted to investigate how LLMs can specifically be adapted to the needs of each user, including individuals with low literacy levels or cognitive and linguistic impairments, by providing adapted output (Demszky et al., 2023). The training data for large language models (LLMs) primarily comprises text created by individuals without language disabilities. This could potentially lead to a limited exposure to the varied linguistic patterns and communication styles exhibited by individuals with language impairments (Fiora et al., 2024; Guo et al., 2024).

Finally, most of the existing systems, focused on the English language. However, languages like Italian remain relatively under-explored in this domain, primarily due to data scarcity and poor data quality. Despite efforts to address this gap (Brunato et al., 2015, 2016, 2022; Tonelli et al., 2016, 2017), the availability of Italian simplification datasets remains limited, with only a few manually curated datasets and one large corpus assembled through a data-driven approach.

This paper aims to address these gaps by (I) creating a robust corpus by merging and cleaning existing resources (II) training a sequence-to-sequence neural model, (III) incorporating an adaptive component to control simplifications for specific target populations. Our most successful model achieves a SARI score of **60.12** and a BLEU score of **50.30** on the test set. Moreover, we present an experiment evaluating the suitability respectively of our fine-tuned model and Chat-GPT 3.5 for automatic text simplification specifically focused to the disability domain.

## 2 Related Work

ATS can occur at different levels of granularity: sentence-level, paragraph-level, or even at the level of entire documents and articles. In this work, we focus on a sentence-level automatic text simplification task. Consequently, our attention is solely directed towards existing work related to sentences.

Sentence-level simplification is often approached as a monolingual form of machine translation (MT). For years, attempts have been made to tackle this task using rule-based models capable of handling both lexical simplification and morpho-syntactic simplification. These techniques rely on manually crafted rules (Bott et al., 2012; Shardlow, 2014a; Siddharthan, 2011). Manually curated data offer several advantages. They en-

sure clear and consistent data labeling, non-redundant metadata recording, and structured presentation of contextual linguistic phenomena associated with text simplification. Nevertheless, constructing such models demands extensive investment of time and resources on experts in language knowledge. Moreover these systems suffer from a notable drawback: limited portability and scalability to new scenarios.

Authors	Description	Approach
Yatskar et al. (2010)	Context similarity to extract simplification rules.	DD
Siddharthan (2011).	Simplification and regeneration from typed dependencies	RB
Biran et al. (2011)	The first data-driven system available for English	DD
Bott et al. (2012).	First model and data for Spanish	RB
Shardlow (2014a).	Errors identification and classification scheme	RB
Glavaš and Štajner (2015)	Based on word vector representations, cased.	DD
Paetzold and Specia (2015)	Modeling words and POS tags.	DD
Nisioi et al. (2017)	Two LSTM layers incorporating global attention.	DD
Zhang and Lapata (2017)	Utilized LSTM, added lexical constraints, and combined with reinforcement learning.	DD
Scarton and Specia (2018)	Enhanced the encoder by incorporating external information.	DD
Zhao et al. (2018)	Transformer-based approach supplemented with a paraphrase database.	DD
Qiang et al. (2020)	Extension to BERT.	DD

Table 1: Models for Sentence Simplification from the least recent to the most recent, along with descriptions of the systems and an indication of whether it’s rule-based (RB) or data-driven (DD).

Most sentence simplification models are available for English, primarily due to the availability of extensive supervised training datasets containing pairs of complex and simple sentences that are aligned in structure and meaning (Wubben et al., 2012; Martin et al., 2020). However, efforts have also been made to explore languages beyond English, including Brazilian Portuguese (Aluísio et al., 2008), Spanish (Saggion et al., 2015), (Glavaš and Štajner, 2015), Italian (Brunato et al., 2015; Tonelli et al., 2016), Japanese (Goto et al., 2015; Kajiwara and Komachi, 2018; Katsuta and Yamamoto, 2019), and French (Gala et al., 2020).

Moreover, the emergence of LLMs and, particularly, GPT has brought about a revolution in the field of NLP. Its impressive text generation capabilities, supported by pre-trained knowledge and fine-tuning adaptability, make it a versatile tool for various NLP tasks, including automatic text simplification. Despite their success in many benchmarks and challenges (Anschütz et al., 2023; Sun et al., 2023;

Engelmann et al., 2023; Shaib et al., 2023), it’s important to verify that the outputs of these models can be suitable before deployment also in sensitive domains, such as for use with children who have language disabilities.

### 3 Dataset selection, curation and augmentation

Three main datasets are available for automatic sentence simplification in Italian: (1) Terence & Teacher (Brunato et al., 2015), (2) SIMPITIKI (Tonelli et al., 2016), (3) PaCCSS-IT (Brunato et al., 2016).

Terence & Teacher was introduced as the inaugural Italian Corpus for Text Simplification. Comprising around 1500 sentence pairs, it integrates two sub-corpora: Terence, consisting of 32 simplified children’s stories crafted by experts across three linguistic dimensions, and Teacher, which features 24 pairs of texts manually simplified by a teacher targeting L2 students.

In 2016, SIMPITIKI was created by gathering simplification pairs from Wikipedia edits designated as “simplified”. The pairs were then manually annotated and filtered, leading to a final set of 575 pairs out of the initially scraped 2,671 pairs. Additionally, employing a similar methodology, a second corpus was created by simplifying documents from the Trento Municipality pertaining to building permits and kindergarten admissions. This corpus, focused on public administration, adhered to the same annotation schema and encompassed an additional 591 pairs.

Finally PaCCSS-IT includes 63,000 pairs of sentences classified by their readability score. The corpus was constructed through monolingual sentence alignment techniques, aligning original sentences with their simplified counterparts using metrics like TF/IDF scores or similar methods assessing word similarity. Each pair includes the cosine similarity, accuracy of automatic classification for predicting sentence alignment, and readability level. Even though the dataset is quite large, the authors gathered a substantial amount of text from the web to initiate the process and reduce costs, which carried the risk of generating occasional errors, repetitions, and other issues.

For this reason, we propose an augmented dataset composed by PaCCSS-IT, SIMPITIKI and a translated one. The corpus creation pipeline can be seen in Figure 1. We started by cleaning the larger available corpus, PaCCSS-IT (Brunato et al., 2016), through a pre-

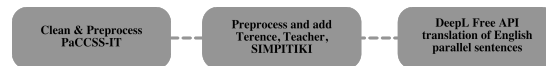


Figure 1: The steps we took to construct the Augmented Dataset

	COMPLEX	SIMPLE
PaCCSS-IT	Quale sarebbe allora la soluzione giusta?	È questa la soluzione giusta?
Teacher	I bei tempi finirono nel maggio 1940, prima la guerra, la capitolazione, l’invasione tedesca, poi cominciarono le sventure per noi ebrei.	I tempi felici finiscono nel maggio 1940, dopo la guerra, la sconfitta, e l’arrivo dei soldati tedeschi, cominciano i problemi per noi ebrei.
Terence	Ernesta Sparalesta è una bambina alta, più o meno, un metro e una noce.	Ernesta Sparalesta è una bambina alta poco più di un metro.
SIMPITIKI	Said spiega che questo processo è stato reso possibile attraverso una conoscenza superficiale di ciò che è in effetti l’Oriente.	Said spiega che questo processo si è realizzato mediante una conoscenza superficiale di ciò che è in effetti l’Oriente.
Translated English Sentences	L’orso bruno dell’Himalaya, noto anche come orso rosso dell’Himalaya, orso isabellino o Dzu-Teh, è una sottospecie dell’orso bruno.	L’orso bruno himalayano è una sottospecie dell’orso bruno.

Figure 2: Some examples of the composition of the Augmented Dataset

processing step similar to the one conducted in Palmero Aprosio et al. (2019). We deliberately retained both capital letters and punctuation within sentences to preserve meaning and convey grammatical and semantic cues. Punctuation was selectively removed primarily at the beginning and end of sentences, and identical pairs of parallel sentences were eliminated to prevent redundancy. However, we retained complex sentences that underwent distinct simplifications to ensure computational models learned the variability in simplifying the same sentence.

Additionally, we excluded complex sentences consisting of two tokens or fewer and those with low cosine similarity values compared to their simpler counterparts. More specifically, we disregarded sentences with cosine similarity less than 0.05. This value was chosen after a manual inspection which identified pairs of simple and complex sentences with significantly different meanings.

Lastly, we also addressed the issue of sentences containing numbers with no corresponding counterpart in the simple sentences. This adjustment ensured consistency not only in alphabetical tokens but also in numerical values. After cleaning, the curated version of the PaCCSS-IT corpus comprised 32,650 pairs of complex and simple sentences. Some examples of the sentences in the augmented corpus can be seen in Figure 2.

In a later stage, we integrated the Terence & Teacher (Brunato et al., 2015) and SIMPITIKI (Tonelli et al., 2016) datasets to the curated version of PaCCSS-IT, conducting spe-

cific parsing and pre-processing to allineate with the format in PaCCSS-IT. Our corpus at this stage consisted of 33,891 parallel sentences. The curated version incorporating the three datasets showed an increase in average sentence length due to the inclusion of sentences from Terence&Teacher, and SIMPITIKI datasets.

Finally, we augmented the curated versions by translating sentences from parallel English datasets. This was done with two main goals (I) enhance data variety and (II) improve the model generalization. For doing that, we used the DeepL API to translate around 5000 sentences pairs from a parallel English datasets. We decided to translate the *Human Simplification with Sentence Fusion Data Set* (Schwarzer et al., 2021) and few sentences translated by the first version of the *Wikipedia dataset* (Kauchak et al., 2022). The augmented version exhibited increased linguistic complexity in both complex and simplified sentences compared to the initial PaCCSS-IT dataset or its curated counterpart, as it can be seen in Table 2. The average sentence length slightly increased in the augmented version, with complex sentences averaging 9.14 words and simplified sentences 8.21 words. The use of conjunctions in simplified sentences showed a progressive increase from PaCCSS-IT to the curated and augmented datasets, suggesting greater cohesion in simplified constructs. Overall, both the curated and augmented datasets displayed higher linguistic detail and richer language use compared to the initial PaCCSS-IT dataset. The average length of complex sentences increased to 8.42, and that of simple sentences to 7.63 as it can be seen in Table 2.

Metric	PaCCSS-IT	Curated	Augmented
<i>AVG_words_complex</i>	8.26	8.42	9.14
<i>AVG_words_simplified</i>	7.34	7.63	8.21
<i>SVO_complex</i>	0.57	0.54	0.55
<i>SVO_simplified</i>	0.54	0.50	0.52
<i>CONJ_complex</i>	0.23	0.25	0.28
<i>CONJ_simplified</i>	0.26	0.27	0.29
<i>SUBJ_complex</i>	0.03	0.05	0.06
<i>SUBJ_simplified</i>	0.025	0.04	0.05
<i>stop_words_complex</i>	4.5	4.78	5.08
<i>stop_words_simplified</i>	2.76	3.02	3.25

Table 2: Normalized metrics for three dataset variations. The *Curated dataset* combines three existing distinct datasets, while the *Augmented Dataset* incorporates the three existing resources together with sentences translated from English parallel corpora. “AVG” stands for average. “SVO” for subject-verb-object. “SUBJ” for subordination conjunctions. “CONJ” for coordination conjunctions”.

## 4 Methods

In this section, we present the architecture details of the two models used in this study, respectively a BERT-based architecture fine-tuned for the task of sentence simplification for Italian and the details of the prompting to Chat-GPT 3.5. In Section 5, we detail the specifics of the BERT-based architecture’s fine-tuning and usage used in our experiments.

**Proprietary System architecture** Our model consists of both an encoder and a decoder component. We employ a BERT-based model fine-tuned for textual simplification tasks. The encoder checkpoints were initialized using pre-trained checkpoints tailored specifically for the Italian language<sup>1</sup> model available in the Hugging Face Hub repository. Conversely, the decoder checkpoints were initialized randomly. When making our architecture choice, it was crucial to consider our target language, namely Italian. At the time of implementing our model, the T5 pre-trained version (Sarti and Nissim, 2022) for Italian was not available. In a second version of our model, we integrated an adaptive component, enabling semi-supervised learning of the model by encoding five numerical values within complex sentences. Following the approach outlined in (Megna et al., 2021; Martin et al., 2019), we incorporated a discrete parametrization mechanism that allows explicit control of the generation. Additionally, we opted to include the Word Ratio parameter proposed by (Sheang and Saggion, 2021). As illustrated in Table 3, these features encompass sentence length (both in terms of characters and tokens), as well as lexical and syntactic complexity. We selected these five parameters because, as highlighted in previous studies, they significantly contribute to the comprehension challenges faced by individuals with reading comprehension deficits (Oakhill and Yuill, 1996; Nation and Snowling, 2000, 2004; Galletti et al., 2023).

**LLM architecture** To showcase the capabilities of Large Language Models (LLMs), we selected ChatGPT-3.5 (Madaan et al., 2022) due to its proficiency in zero-shot learning scenarios and user-friendly interface accessible through the OpenAI platform, which allows for easy integration and experimentation.

<sup>1</sup>namely the [bert-base-italian-xxl-cased](#)

Token	Value	Description
<i>Word_Ratio</i>	0.20	Ratio of words in the complex sentence to words in the simplified sentence.
<i>Character_Ratio</i>	0.20	Ratio of characters in the complex sentence to characters in the simplified sentence.
<i>Word_Rank</i>	0.90	Ranking of words based on frequency or importance.
<i>Lev_Similarity</i>	0.90	Levenshtein similarity between the complex and simplified sentences.
<i>Dependency_Tree</i>	1	Degree of similarity in dependency trees between the complex and simplified sentences.

Table 3: Description of parameters with values used in the adaptive component for simplification.

## 5 Experiment Settings

This section outlines the parameters for model fine-tuning (Subsection 5.2), and discusses the evaluation metrics (Subsection 5.3) used.

### 5.1 BERT-based model Fine-Tuning

For the fine-tuning process, we utilized *Optuna*, an open-source framework for hyperparameter optimization to dynamically build the search space for selecting the optimal parameters for our work. We configured a batch size of 4 for both training and evaluation loops, set a maximum token length of 300, established a learning rate of  $3e - 4$ , configured an Adam epsilon of  $1e - 8$ , implemented a warm-up ratio of 0.10, and conducted 20 epochs. The remaining parameters were kept at their default values from Transformers library. For dividing the three dataset into train and test we used a standard 0.80 split for training and 0.20 for testing. As explained in Section 6, we maintained this fine-tuning parameters for both the two version of our model —the one with the adaptive component and the one without.

### 5.2 GPT’s Prompting

We accessed the *ChatGPT-3.5* model through the [open-access model](#) available. For our experiment, we utilized GPT in zero-shot mode. At the time this work was conducted, ChatGPT-3.5 had only very recently been released. As a result, we couldn’t fully explore different prompt engineering techniques and we were constrained on relying solely on using -3.5 in a zero-shot mode. Specifically, we presented the model with a list of complex sentences and tasked ChatGPT 3.5 with simplifying them for school children aged 8 to 11 with a reading comprehension deficit. Subsequently, we computed our evaluation scores based on the simplified answers generated by

ChatGPT, comparing them to the ground truth provided in our annotated corpus.

### 5.3 Evaluation Metrics

For assessing the performance of both models, we employed well-established metrics for both automatic machine translation and text simplification evaluations, SARI (Xu et al., 2016) and BLEU (Papineni et al., 2002), on our test corpus. We qualitatively inspected the output data to examine the results from each model. Finally, we involved experts specialised in language disabilities to conduct a human evaluation.

SARI and BLEU were chosen for assessing the performance of both models, because of their use in previous work (Van den Bercken et al., 2019; Monteiro et al., 2022; Cardon and Grabar, 2020). SARI (System-level Automatic Reviewer for Machine Translation) is a metric designed to assess the quality of machine-generated sentences, particularly within the context of machine translation. It centers on evaluating the fluency and preservation of meaning in the generated sentences when compared to reference sentences. In contrast, BLEU (Bilingual Evaluation Understudy) is a widely used metric for evaluating machine-generated sentences, primarily within machine translation contexts. It quantifies the similarity between the generated sentence and one or more reference sentences through an n-gram overlap comparison.

These metrics however have several drawbacks to evaluate text simplification output, as pointed out in the literature (Sulem et al., 2018; Al-Thanyyan and Azmi, 2021). We thus also included qualitative human evaluation of the results by qualitatively inspecting the output data to examine the results from each model. We gathered a panel of experts specialized with domain-specific expertise, i.e. speech and language therapists at a partner specialised center in the rehabilitation of Neurodevelopment<sup>2</sup> to conduct a human evaluation, with a specific focus on young children diagnosed with language disabilities. The criteria for selection was their expertise in language learning and disabilities. All annotators were provided with detailed information regarding the study’s purpose, their role in the evaluation and the nature of the data that they were scoring. The experts were not reimbursed financially; however, their participation was voluntary and they were provided with

<sup>2</sup><https://www.crc-baluzie.it/>



informed consent before the beginning of the study.

The evaluation of the quality of the text simplification corpus was made possible through the utilisation of a Google Form available at this link <sup>3</sup>. The form evaluated scales from 0 to 5 (being 0, the lowest and 5, the highest values) concerning grammatical correctness, maintenance of meaning, and level of simplicity gained as similar work in the literature (Xu et al., 2016). We selected 10 sentences to represent both the highest and lowest cosine distances between the sentences generated by ChatGPT-3.5 and our model. Specifically, we selected five pairs with the highest cosine distances and five pairs with the lowest. These sentences have been put at disposal to the ten experts who participated in the users studies. Several considerations prompted this approach: firstly, we needed a manageable sample size feasible for evaluation within our available annotators. Secondly, by including both the most divergent and the most similar cases, we aimed to ensure robustness in extreme scenarios and reduce bias in our evaluation method.

## 6 Results

In this section we report results on the automatic and human evaluation conducted.

Dataset	SARI	BLEU
Palmero Aprosio et al. (2019)	49.49	N/A
(A) Fine-tuned + Original PaCCS-IT Dataset	57.10	46.00
(B) Fine-tuned + Merged and Cleaned Dataset	55.64	49.78
(C) Fine-tuned + Augmented Dataset	51.51	47.40
(D) Fine-tuned + Augmented + Adaptive Component	60.12	50.30
ChatGPT-3.5	40.51	15.00

Table 4: SARI and BLEU scores for all our fine-tuned models with the combinations of the different datasets.

### 6.1 Automatic Evaluation

In our work, we conducted three different fine-tuning runs using the same fine-tuned model and equivalent hyper-parameters using three different training data, as it can be seen in Table 4. These three models correspond to model (A), (B) and (C) in the table.

The first fine-tuning of the model, i.e. (A), was done using the original version of PaCCS-IT. It resulted in a SARI score of 57.10 when evaluated on the test corpus. This score was

higher than the current state-of-the-art for Italian language Automatic Text Simplification task (Palmero Aprosio et al., 2019). Given the errors manually noticed, it was hypothesized that the high SARI score achieved during fine-tuning resulted from over-fitting to poor-quality data, representing a learning fallacy. To investigate this hypothesis, we fine-tuned our model using the curated version of our dataset, i.e. (B). In this case, SARI fell by two points (55.64). This improvement may be attributed to the inclusion of three merged corpora (Teacher, Terence, and Simpitiiki), which provided the model with more diverse material to learn from and thus greater flexibility in the generative phase. The lower SARI value could precisely reflect this behavior and shed light on the previous over-fitting. Following the previous result, we conduct the final fine-tuning with the Augmented dataset, i.e. (C). At this stage, we note that SARI is another 4 percentage points lower than in the last training (51.51). Finally, we fine-tuned an additional model (D), adding the adaptive component detailed in section 4 and using our augmented dataset. Our model obtains a SARI score equal to 60.12 and a BLEU score of 50.30 on the same test set, achieving the best results over the four fine-tuned models.

GPT-3.5 exhibited notably lower performance, yielding a SARI score of 40.51 and a BLEU score of 15.00 on the same test set<sup>4</sup>. GPT-3.5’s notably lower performance in comparison to the fine-tuned transformer model within the context of text simplification can be attributed to several factors. Firstly, the fine-tuned transformer model likely benefited from training data and therefore domain adaptation while, GPT-3.5 was used in a zero-shot mode. Secondly, the fine-tuned transformer model’s architecture and parameters were carefully adapted to the nuances of text simplification, a step that might have been omitted in GPT’s preparation. Finally, the ability to control the level of simplification plays a vital role in effective text simplification (Nishihara et al., 2019; Xu et al., 2016). The nuanced interplay of these factors contributes to the observed performance disparities in the presented context and it pushed us to perform also a qualitative evaluation. All the results are presented in Table 4.

However, despite this good performance on the standard formal metric, a manual and visual evaluation of the generated simplifica-

<sup>3</sup>Click here to access the Google Form

<sup>4</sup>The test corpus is available for exploration via the following link

Input Test Set	Chat GPT	Best Performing Model
A cosa è dovuto ciò What is this due to	Perché è successo? Why did it happen?	A cosa è dovuto? What is it due to?
Beh, questo è tutto un altro discorso <i>Well, that's a whole other story</i>	Beh, questa è tutta un'altra storia <i>Well, that's a whole other story</i>	<i>É tutto un altro discorso</i> <i>That's a whole other speech</i>
Ma non ne farei un dramma <i>But I wouldn't make a drama out of it</i>	Non dovremmo preoccuparci troppo di questa cosa <i>We shouldn't worry too much about it</i>	Ma non faceva un dramma <i>But it wasn't a drama</i>
Tutti erano nell'atmosfera della festa <i>Everyone was in the party mood</i>	Tutti erano felici e festosi <i>Everyone was happy and festive</i>	Era nell'atmosfera della festa <i>It was in the party mood</i>
Ma questo è vero solo in teoria <i>But this is only true in theory</i>	Ma questa teoria è vera solo sulla carta <i>But this theory is only true on paper</i>	Questo è vero in teoria <i>This is true in theory</i>

Table 5: Example of some simplifications on the test corpus given by our model and GPT-3.5.

tions revealed several issues. The simplifications were found to be misleading and contained frequent gross errors. For instance, the reference sentences in the dataset were sometimes not very informative in terms of simplifications and appeared to be alternative versions of the complex sentence but not simplified ones, as shown in Table 6.

Simple	Complex
Questa sarebbe una cosa <i>positiva</i> This would be a <i>positive</i> thing	Questa è una cosa <i>gravissima</i> This is a <i>very serious</i> thing
Ma facciamo un passo più <i>avanti</i> But let's take a step <i>forward</i>	Ma facciamo un lungo passo <i>indietro</i> But let's take a long step <i>backward</i>

Table 6: The original complex sentences from the test dataset and simplifications produced by the fine-tuned model; highlighting mistakes in italics.

## 6.2 Human Evaluation

### 6.2.1 Qualitative Analysis

In a later stage, we inspected the generated simplified sentences given by our models. We found that while the simplification efforts undertaken by ChatGPT-3.5 are generally satisfactory upon close qualitative examination, there are instances where the simplifications verge on being abstract. The generated simplifications at times involve conceptual abstractions that could potentially introduce an unintended layer of complexity as it can be seen in Table 5. This paradoxical outcome could arise because the model simplify, yet occasionally employs abstract concepts that might prove too complex for the intended au-

dience, especially young children or individuals with specific clinical diagnoses. In fact, ChatGPT sometimes seems to capture greater nuances of cause-and-effect or context than an 8- to 11-year-old child who has limited experience of the world and thus may struggle to make such detailed connections, and as a result, the simplification proposed by ChatGPT can sometimes be difficult for children to interpret. For instance, ChatGPT-3.5 might attempt to convey a complex idea by substituting certain words or phrases with simpler alternatives. However, in doing so, it might inadvertently introduce terms that are not within the immediate vocabulary of the target audience or that require a certain level of background knowledge to be fully understood. This kind of simplification could lead to confusion or misinterpretation among individuals who require the content to be presented always in an easily accessible manner.

### 6.2.2 Experts Evaluation

To complete our qualitative analysis, we asked experts to evaluate the results given by the models. This evaluation yielded mixed results as it can be seen in Figure 3. When we compared the scores of the two models based on the chosen criteria (grammaticality, meaning preservation, and level of simplification), there was not a significant difference between them. This is in contrast to the results of the automatic evaluation, where our fine-tuned transformer model appeared to outperform ChatGPT-3.5 on our test set. This highlighted the fact that we are still lacking an evaluation mechanism that is both objective and aligns closely with human judgment. Without an accurate way to assess the quality of text generated by a simplification model, it becomes challenging to implement effective con-



Figure 3: The plots with the form’s results. The sentences were re-arranged and their order do not reflects their cosine distances.

trols. This underscores that research in this area is still very much in an experimental stage and is in its early phases.

## 7 Conclusions and future work

In this paper, we curated a comprehensive corpus by cleaning and combining existing resources, we fine-tuned an adaptive transformer model for the sentence simplification task in Italian, we integrated an adaptable component to tailor simplifications for specific target groups, we evaluated the model by comparing it to ChatGPT-3.5, through both quantitative and qualitative assessments, including expert and automatic evaluations of the simplified text. The automatic evaluation highlighted that the fine-tuned version of BERT model seem the better suited for the task. Moreover the adaptive component increase the State-Of-The-Art (SOTA) results by 11 points. Lastly, LLMs, particularly GPT-3.5, have shown significant advancements in the generation of coherent and fluently articulated text, but a substantial scope for improvement persists con-

cerning the crafting of textual content that aligns effectively with the requisites of individuals possessing particular diagnostic profiles or clinical conditions. This progress can hold promising implications for Computer-Assisted Language Learning, as it can facilitate the creation of tailored educational materials that accommodate the unique learning needs and abilities of diverse student populations. Finally, we believe that there is still much to do to improve the current evaluation metrics for automatic text simplification to understand the nuances and potential biases they may introduce and to make sure they align with human evaluation. Developing and refining new evaluation metrics tailored specifically for populations with diverse linguistic needs and clinical conditions could be a crucial step forward the use of NLP in clinical and educational contexts. Finally, more extensive and robust user studies are required to evaluate the effectiveness of GPT-3.5 in generating text for specific user groups.

## References

- Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.
- Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications*, pages 1–9.
- Sandra M Aluisio, Lucia Specia, Thiago AS Pardo, Erick G Maziero, and Renata PM Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*, pages 240–248.
- Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, and Georg Groh. 2023. Language models for german text simplification: Overcoming parallel data scarcity through style-specific pre-training. *arXiv preprint arXiv:2305.12908*.
- Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, pages 3286–3292.
- Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501.
- Stefan Bott, Luz Rello, Biljana Drndarević, and Horacio Saggion. 2012. Can spanish be simpler? lexis: Lexical simplification for spanish. In *Proceedings of COLING 2012*, pages 357–374.
- Stefan Bott and Horacio Saggion. 2014. Text simplification resources for spanish. *Language Resources and Evaluation*, 48(1):93–120.
- Dominique Brunato, Andrea Cimino, Felice Dell’Orletta, and Giulia Venturi. 2016. Pacss-it: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361.
- Dominique Brunato, Felice Dell’Orletta, and Giulia Venturi. 2022. Linguistically-based comparison of different approaches to building corpora for text simplification: A case study on italian. *Frontiers in Psychology*, 13:707630.
- Dominique Brunato, Felice Dell’Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41.
- Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716.
- Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margaret Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.
- Björn Engelmann, Fabian Haak, Christin Katharina Kreutz, Narjes Nikzad Khasmakhi, and Philipp Schaer. 2023. Text simplification of scientific texts for non-expert readers. *arXiv preprint arXiv:2307.03569*.
- A Fiora, F Piferi, P Crovari, and F Garzotto. 2024. Exploring large language models for the education of individuals with cognitive impairments. In *INTED2024 Proceedings*, pages 4479–4487. IATED.
- Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C Ziegler. 2020. Alector: A parallel corpus of simplified french texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1353–1361.
- Martina Galletti, Eleonora Pasqua, Francesca Bianchi, Manuela Calanca, Francesca Padovani, Daniele Nardi, and Donatella Tomaiuolo. 2023. A reading comprehension interface for students with learning disorders. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, pages 282–287.
- Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68.
- Isao Goto, Hideki Tanaka, and Tadashi Kumano. 2015. Japanese news simplification: tak design, data set construction, and analysis of simplified text. In *Proceedings of Machine Translation Summit XV: Papers*.
- Sichen Guo, François Leborgne, Jun Hu, and Walter Baets. 2024. Can ai bridge the literacy gap? developing a gpt-4 summarization tool for low literacy. *From User to Human*, page 52.
- Tomoyuki Kajiwara and Mamoru Komachi. 2018. Text simplification without simplified corpora. *The Journal of Natural Language Processing*, 25:223–249.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh,

- Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- Akihiro Katsuta and Kazuhide Yamamoto. 2019. Improving text simplification by corpus expansion with unsupervised learning. In *2019 International Conference on Asian Language Processing (IALP)*, pages 216–221. IEEE.
- David Kauchak, Jorge Apricio, and GONDY Leroy. 2022. [English Datasets resources](#).
- Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. *arXiv preprint arXiv:2210.07128*.
- Louis Martin, Angela Fan, Éric De La Clergerie, Antoine Bordes, and Benoît Sagot. 2020. Muss: multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.
- Louis Martin, Benoît Sagot, Eric de la Clergerie, and Antoine Bordes. 2019. Controllable sentence simplification. *arXiv preprint arXiv:1910.02677*.
- Angelo Luigi Megna, Daniele Schicchi, Giosué Lo Bosco, and Giovanni Pilato. 2021. A controllable text simplification system for the italian language. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 191–194. IEEE.
- José Monteiro, Micaela Aguiar, and Sílvia Araújo. 2022. Using a pre-trained simplet5 model for text simplification in a limited corpus. *Proceedings of the Working Notes of CLEF*.
- Kate Nation and Margaret J Snowling. 2000. Factors influencing syntactic awareness skills in normal readers and poor comprehenders. *Applied psycholinguistics*, 21(2):229–241.
- Kate Nation and Margaret J Snowling. 2004. Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of research in reading*, 27(4):342–356.
- Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. [Controllable text simplification with lexical constraint loss](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266, Florence, Italy. Association for Computational Linguistics.
- Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 85–91.
- JANE Oakhill and N Yuill. 1996. Reading comprehension difficulties. *Hillsdale, NJ*.
- Gustavo Paetzold and Lucia Specia. 2015. Lexenstein: A framework for lexical simplification. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 85–90.
- Alessio Palmero Aprosio, Sara Tonelli, Marco Turchi, Matteo Negri, and A Di Gangi Mattia. 2019. Neural text simplification in low-resource conditions using weak supervision. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen)*, pages 37–44. Association for Computational Linguistics (ACL).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8649–8656.
- Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or help? text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1–10.
- Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.
- Gabriele Sarti and Malvina Nissim. 2022. It5: Large-scale text-to-text pretraining for italian language understanding and generation. *arXiv preprint arXiv:2203.03759*.
- Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718.
- Max Schwarzer, Teerapaun Tanprasert, and David Kauchak. 2021. Improving human text simplification with sentence fusion. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 106–114.
- Chantal Shaib, Millicent L Li, Sebastian Joseph, Iain J Marshall, Junyi Jessy Li, and Byron C Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success). *arXiv preprint arXiv:2305.06299*.
- Matthew Shardlow. 2014a. Out in the open: Finding and categorising errors in the lexical

- simplification pipeline. In *LREC*, pages 1583–1590.
- Matthew Shardlow. 2014b. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.
- Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Advaith Siddharthan. 2011. Text simplification using typed dependencies: A comparison of the robustness of different generation strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 2–11.
- Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.
- Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*.
- Renliang Sun, Wei Xu, and Xiaojun Wan. 2023. Teaching the pre-trained model to generate simple texts for text simplification. *arXiv preprint arXiv:2305.12463*.
- Sara Tonelli, Alessio Palmero Aprosio, and Marco Mazzon. 2017. The impact of phrases on italian lexical simplification. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, pages 316–320.
- Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. Simpiti: a simplification corpus for italian. In *CLiC-it/EVALITA*.
- Sander Wubben, Antal Van Den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024.
- Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.
- Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. *arXiv preprint arXiv:1008.1986*.
- Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*.
- Sanqiang Zhao, Rui Meng, Daqing He, Saptono Andi, and Parmanto Bambang. 2018. Integrating transformer and paraphrase rules for sentence simplification. *arXiv preprint arXiv:1810.11193*.

# A Conversational Intelligent Tutoring System for Improving English Proficiency of Non-Native Speakers via Debriefing of Online Meeting Transcriptions

Juan Antonio Pérez-Ortiz<sup>\*,</sup> Miquel Esplà-Gomis<sup>°</sup>, Víctor M. Sánchez-Cartagena<sup>°</sup>, Felipe Sánchez-Martínez<sup>°</sup>, Roman Chernysh<sup>°</sup>, Gabriel Mora-Rodríguez<sup>°</sup>, Lev Berezhnoy<sup>°</sup>

<sup>°</sup> Universitat d'Alacant, Spain {japerez, mespla@ua.es}

<sup>\*</sup>Valencian Graduate School and Research Network of Artificial Intelligence, ValgrAI, Spain

## Abstract

This paper presents work-in-progress on developing a conversational tutoring system designed to enhance non-native English speakers' language skills through post-meeting analysis of the transcriptions of video conferences in which they have participated. Following recent advances in chatbots and agents based on large language models (LLMs), our system leverages pre-trained LLMs within an ecosystem that integrates different techniques, including in-context learning, external non-parametric memory retrieval, efficient parameter fine-tuning, grammatical error correction models, and error-preserving speech synthesis and recognition. While the system is still in development, a preliminary pilot evaluation of a prototype has been conducted with L2 English students.

## 1 Introduction

In an increasingly interconnected world, the ability to communicate effectively in English has become a vital skill, especially in professional settings where English has firmly established itself as the *lingua franca* (Nickerson, 2005; Shegebayev, 2023). However, this requirement often leads to challenging situations for many non-native speakers who, when participating in meetings, presentations, and discussions conducted in English, frequently find themselves navigating the complexities of the language under the potential scrutiny of more fluent colleagues. This dynamic can create a stressful environment, hindering effective communication and the free flow of ideas, leading to misunderstandings, and impacting the confidence and performance of less-proficient speakers (Aichhorn and Puck, 2017). These linguistic shortcomings are often silently noted by other participants, but rarely addressed in a constructive manner, and

the very settings where these individuals most frequently use English are not leveraged as opportunities for improvement.

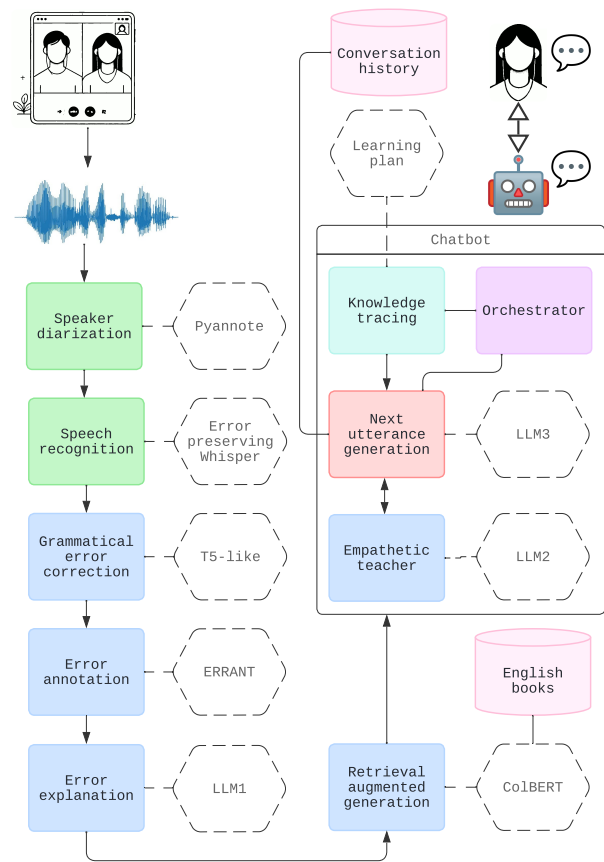


Figure 1: Main components of the DeMINT system. Section 3 describes each component thoroughly.

Although a human tutor could provide valuable feedback and guidance to help non-native speakers improve their language skills, this solution is often impractical due to logistical constraints, financial considerations, or the reluctance to introduce additional complexity into an already demanding professional life. To address this gap, we propose an automated language debriefing system that leverages the transcripts of online meetings to

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

Juan Antonio Pérez-Ortiz, Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Felipe Sánchez-Martínez, Roman Chernysh, Gabriel Mora-Rodríguez and Lev Berezhnoy. A Conversational Intelligent Tutoring System for Improving English Proficiency of Non-Native Speakers via Debriefing of Online Meeting Transcriptions. *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)*. Linköping Electronic Conference Proceedings 211: 187–198.

provide feedback and guidance to non-native English speakers (L2-English), thus mimicking the role of a language instructor. Our system, called DeMINT after the project in which it was developed, is implemented as an educational chatbot (Du and Daniel, 2024) that interacts with users in a conversational manner, thereby transforming everyday professional interactions into valuable opportunities for language improvement.

Conversational intelligent tutoring systems (ITS) are set to revolutionize the field of education, offering one-to-one, personalized, interactive, engaging, and inclusive learning experiences to students. Their application in computer-aided language learning (CALL) is particularly promising, as contemporary large language models (LLMs) show remarkable capabilities in language understanding and generation. While such systems were explored in the past (Jia, 2009; Bibauw et al., 2019), only with the advent of contemporary LLMs have functional implementations become feasible.

Our ITS aims to leverage LLMs in a CALL application to improve speakers’ language skills through interactive, personalized, and error-driven conversations. A functional prototype has been evaluated in a pilot study with L1 Spanish/L2 English learners. The source code, along with links to models and datasets, is available online.<sup>1</sup>

The rest of the paper is organized as follows. Section 2 reviews related work on chatbots in education. After that, Section 3 describes DeMINT’s design and its main components. Then, Section 4 outlines the human evaluation. After the conclusions and potential future work described in Section 5, the ethical considerations of the project and the main limitations are highlighted.

## 2 Related Work

The year 2022 marks a turning point where the capabilities of LLMs for conversational and educational tasks, in general, and ITS, in particular, became evident. Despite this, prior research had already demonstrated the potential benefits of using traditional chatbots within dialog-based CALL scenarios for L2 learners (Jia, 2009; Bibauw et al., 2019; Huang et al., 2022; Yang et al., 2022), identifying pedagogical, technological, and social affordances (Jeon, 2024).

<sup>1</sup><https://github.com/transducens/demint>

Shortly after the release of ChatGPT in November 2022, several studies explored its potential for L2 teaching. A survey among English-as-foreign-language faculty instructors by Mohamed (2024) highlighted ChatGPT’s ability to enhance proficiency and motivation, while also emphasizing the need to address limitations and ethical concerns. A meta-analysis by Zhang et al. (2023) of 18 articles on chatbot-assisted language learning concluded that “using chatbots for language learning has a positive impact, and the learning outcomes are better than those in non-CALL situations.” A more recent meta-study by Cisłowska and Acuña (2024) observed that “the use of chatbots can positively affect students’ attitudes toward learning a foreign language, enhancing motivation, interest, fun, proactivity, and learning commitment”; however, they also noted that “the novelty effect may decrease motivation over time, and lacking a human factor may fail to meet emotional needs and decrease motivation.” Several other recent reviews have reached similar conclusions (Labadze et al., 2024; Du and Daniel, 2024).

Additionally, the emergence of commercial AI-driven language learning assistants developed by companies like Duolingo,<sup>2</sup> Google,<sup>3</sup> or TalkPal<sup>4</sup> underscores the growing importance and effectiveness of LLM-based CALL systems. In spite of the potential of these systems, we are not aware of many open-source projects that implement a comprehensive conversational ITS for L2 learning as ours, especially one that leverages the transcripts of online meetings to provide feedback and guidance to L2-English speakers.

## 3 System Description

Our system design draws from recent chatbots like BlenderBot3 (Shuster et al., 2022) which are built as a pipeline of different modules that mainly consist of LLMs fine-tuned for specific tasks.<sup>5</sup>

A diagram of the main components of DeMINT is shown in Figure 1 on the first page. As can be seen, the system is composed of several modules that interact with each other to provide a compre-

<sup>2</sup><https://blog.duolingo.com/duolingo-max>

<sup>3</sup><https://research.google/blog/english-learners-can-now-practice-speaking-on-search>

<sup>4</sup><https://talkpal.ai>

<sup>5</sup>This also resonates, albeit on a smaller scale, with the revitalization of Minsky’s societies of mind theory (Minsky, 1986) in the form of natural language-based societies of LLMs and other machine learning models mindstorming together to solve a problem (Zhuge et al., 2023).



hensive tutoring experience. Some of them are based on pre-trained LLMs, pre-trained sequence-to-sequence models or ad-hoc models, while others rely on external resources such as textbooks on English grammar. Next sections describe each of these modules in detail. The pipeline of modules outside of the chatbot box in Figure 1 is run offline before the chatbot starts interacting with the user.

### 3.1 Diarization

The pipeline starts by processing the audio recordings of the target online meeting and identifying the segments corresponding to each speaker. This is done by using the library `pyannote.audio` (Bredin, 2023; Plaquet and Bredin, 2023), which relies on a neural speaker diarization model (Takashima et al., 2021). The diarization process returns the start and end times of each speaker turn, as well as the speaker ID.

The audio fragments corresponding to each speaker are then individually processed by the speech recognition system described in the next section. Remarkably, an alternative approach has been recently proposed where diarization and transcription are performed in parallel, and the outputs are subsequently combined.<sup>6</sup>

### 3.2 Speech Recognition

As our error analysis pipeline is performed on the written transcriptions of the online meetings, a speech-to-text (STT) model is needed to transcribe the utterances for each speaker. Our initial approach was to directly use open-weight pre-trained models such as Whisper (Radford et al., 2023), but preliminary tests showed that they were not entirely suitable for our purposes, due to the fact that their strong internal language model tends to correct some of the grammatical errors in the utterances. For example, the Whisper model would often transcribe “I \*doesn’t know” as “I don’t know”, which is unacceptable for our purposes as the original grammar errors need to be faithfully preserved. Consequently, our system includes a custom error-preserving STT model that retains more grammatical errors. This model is obtained by fine-tuning Whisper on a custom dataset of spoken sentences with grammatical errors that we specifically created for our system.<sup>7</sup>

<sup>6</sup><https://huggingface.co/blog/asr-diarization>

<sup>7</sup>Michot et al. (2024) recently demonstrated that certain CTC-based encoder models corrected slightly fewer errors

The ad-hoc dataset comprises both synthetic and natural texts containing grammatical errors. The natural texts are sourced from the COREFL dataset (Lozano et al., 2020), which contains essays by non-native students with varying levels of English proficiency.<sup>8</sup> COREFL includes some audio recordings of students reading their texts, as well as written compositions. However, since only a small percentage of the texts have corresponding audio recordings, we have also converted written texts into audio using the StyleTTS2 text-to-speech (TTS) model (Li et al., 2023), which allows us to synthesize each text with multiple voices, thereby increasing the diversity of the training data. On the other hand, the synthetic texts come from the C4<sub>200M</sub> dataset (Stahlberg and Kumar, 2021), which contains heterogeneous grammatically incorrect sentences synthetically generated via a corruption model.<sup>9</sup> We have converted these sentences into audio using the same StyleTTS2 model. The resulting dataset contains 32,000 speech training samples, 1,000 validation samples, and 1,000 test samples. The training set is composed of 28,592 utterances from C4<sub>200M</sub>, 814 audios directly obtained from COREFL, and 2,594 synthetic utterances generated from the COREFL written texts. The test and validation sets are similarly divided between the two sources. This dataset is then used to fine-tune Whisper, which is subsequently employed to transcribe the audio from the online meetings. Further details on the hyperparameters used for model fine-tuning can be found in the appendix. The two resulting models—one based on the original Whisper model and the other on its distilled version, which is the one we ultimately used—are available on the HuggingFace hub.<sup>10,11</sup>

than the encoder-decoder-based Whisper model, which they attributed to the reduced influence of the language model, but this came at the expense of degraded overall performance. As a result, we continue to use Whisper in our system. It is worth noting that their study addresses a similar challenge, aiming to develop error-preserving STT models. While our approach is primarily automatic, their work involves the collection and annotation of a corpus containing English grammatical errors from young learners.

<sup>8</sup>Given that our evaluation will primarily involve students whose mother tongue is Spanish, we use only the subset of COREFL produced by Spanish students.

<sup>9</sup>In order to avoid the fine-tuned model relying too much on ungrammatical utterances, we add clean utterances from the *correct side* of C4<sub>200M</sub> to the training dataset as well.

<sup>10</sup><https://huggingface.co/Transducens/error-p-reserving-whisper>

<sup>11</sup><https://huggingface.co/Transducens/error-p-reserving-whisper-distilled>

Both datasets complement as C4<sub>200M</sub> provides a wide range of sentences and errors, although with a limited repertoire of voices, while COREFL offers a more diverse set of voices, accents, and natural errors. The COREFL dataset has the additional advantage of allowing our system to adapt to the accents and errors typically made by L1-Spanish speakers, who are the users in our pilot study. Due to the licensing restrictions of COREFL, only the dataset portion based on the C4<sub>200M</sub> dataset has been released on the HuggingFace hub as the Synthesized English Speech with Grammatical Errors Dataset (SESGE).<sup>12</sup>

Each transcribed utterance is split into sentences<sup>13</sup> before proceeding to the next step, and each sentence is associated with the speaker ID.

### 3.3 Grammatical Error Correction

Core to our work, *grammatical error correction* (GEC) is a well-known NLP task that aims to correct grammatical errors in a given text (Bryant et al., 2023; Omelianchuk et al., 2024). There are established shared tasks (Bryant et al., 2019) and datasets such as FCE (Yannakoudakis et al., 2011), NUCLE (Dahlmeier et al., 2013), Lang-8 (Mizumoto et al., 2011; Tajiri et al., 2012), W&I+LOCNESS (Bryant et al., 2019), and JF-LEG (Napoles et al., 2017). A GEC model transforms a sentence with grammatical errors into a grammatically correct one.

Our system currently employs a relatively simple model obtained by fine-tuning the T5 encoder-decoder model (Raffel et al., 2020) on the JF-LEG dataset,<sup>14</sup> but we are considering using more advanced state-of-the-art models such as GRECO (Qorib and Ng, 2023) or the ensembles provided by Omelianchuk et al. (2024).<sup>15</sup>

### 3.4 Error Annotation

Given the original and the corrected version of each sentence, we use the ERRANT toolkit<sup>16</sup> (Felice et al., 2016; Bryant et al., 2017) to extract and annotate the edits necessary to transform one sentence version into the other. ER-

<sup>12</sup><https://huggingface.co/datasets/Transducers/sesge>

<sup>13</sup>Sentence splitting is achieved using the Python’s package `sentence-splitter`.

<sup>14</sup><https://huggingface.co/vennify/t5-base-grammar-correction>

<sup>15</sup><https://github.com/grammarly/pillars-of-gec>

<sup>16</sup><https://github.com/chrisjbryant/errant>

RANT accomplishes this by applying an extended, linguistically-motivated version of the classical Levenshtein distance (Levenshtein, 1966), followed by a rule-based labeling of the edits. The resulting annotations are stored in the M2 format and then integrated into the JSON schema used as the intermediate format between the different components of our system.

### 3.5 Error Explanation

As ERRANT provides high-level annotations such as R:VERB:SVA (error in subject-verb agreement) without additional details, an LLM is used to generate finer-grained natural-language explanations of these errors via few-shot in-context learning. This aligns with recent works on using LLMs to further explain corrections made by GEC models (Fei et al., 2023; Kaneko and Okazaki, 2024; Song et al., 2024). These explanations will later *inspire* the chatbot’s responses to the user.

Among all the open-weight, locally-installable LLMs available, we have found Llama-3.1-8B<sup>17</sup> to offer a good balance between speed and quality. Regarding the prompts used to query the model, we are considering using the DSPy framework (Khattab et al., 2023) to automatically generate them via DSPy’s principled search mechanism (Khattab et al., 2022).<sup>18</sup>

### 3.6 Retrieval from Textbooks

Another component of the pre-processing pipeline is a module that retrieves information from English learning textbooks based on the errors being analyzed. This information will be one of the inputs provided to the chatbot’s next-dialog-line generator at the end of the pipeline. We collected six PDF textbooks to be consulted under the retrieval-augmented generation (RAG, see below) approach, either open-licensed or available from `archive.org`. These English textbooks are valuable not only for their explanations of grammatical rules but also for the real examples of language usage they provide.<sup>19</sup>

<sup>17</sup><https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct>

<sup>18</sup>We have found that DSPy’s `compile` function is useful even for simple chains involving a single model, as it allows us to easily replace the model without manually rewriting the prompt, and also enforces a certain structure in the JSON output.

<sup>19</sup>We also plan to use an LLM as a source of this kind of grammatical information and examples, and to compare the results of both approaches.

*Retrieval-augmented generation* (RAG) encompasses a variety of techniques that integrate information from external documents into the generation process (Lewis et al., 2020). Naïve approaches to RAG involve segmenting documents into passages, computing an embedding for each passage, and storing both texts and embeddings in a vector database. Based on the current topic (each particular error in the use of English in our case), the most relevant passages are retrieved from the database by efficiently computing the similarity between an embedding of the topic and the embeddings of the passages. These selected passages are then provided to an LLM as a source of information for generating the output.

Our system employs the state-of-the-art ColBERTv2 model (Santhanam et al., 2022b), as implemented by the RAGatouille<sup>20</sup> library. ColBERTv2 computes token-level embeddings for passages and queries, making it more suited to our task than alternatives that compute a single dense embedding for each paragraph, as books’ passages are likely to contain heterogeneous information such as grammar rules, open-domain examples, and exercises. ColBERTv2 is combined with a technique called *performance-optimized late interaction driver* (PLAID) (Santhanam et al., 2022a), which replaces conventional vector databases such as FAISS (Douze et al., 2024) with a more efficient and scalable approach based on using centroids of clusters of embeddings instead of the embeddings themselves. Additionally, the RAGatouille documentation states that its implementation of ColBERTv2 is robust in new domains and includes strong default settings, thereby eliminating the need for fine-tuning.

The above modules run offline prior to the debriefing session. Next, we describe those actively engaging in the chatbot’s interaction with the user.

### 3.7 Empathetic Teacher

Another *ingredient* fed to the next-dialog-line generator comes from an LLM fine-tuned with real-life, ideally-empathetic teacher-student conversations. This model processes the recent conversation history and provides guidance on how a teacher might respond to the student’s utterance. In order to obtain this model, we fine-tuned the Llama-3.1-8B model<sup>21</sup> with the following

<sup>20</sup><https://github.com/bclavie/RAGatouille>

<sup>21</sup><https://huggingface.co/meta-llama/Meta-Llama-3.1-8B>

datasets: the Teacher-Student Chatroom Corpus, TSCCV2 (Caines et al., 2022), CIMA (Stasaski et al., 2020), the Multicultural Classroom Discourse Dataset (Rapanta et al., 2021), MathDial (Macina et al., 2023), and Conversational Uptake (Demszky et al., 2021). Some of the datasets were preprocessed in order to split very long conversations resulting in the figures shown in Table 1. The resulting collection of 6 503 conversation turns was split into 5 859 for training, 322 for validation, and 322 for testing, with each dataset contributing proportionally the same across these splits. Further details on the training hyperparameters are provided in the appendix. The fine-tuned teacher model is available on the HuggingFace hub.<sup>22</sup>

Dataset	Original		Split turns	
	Turns	Words	Turns	Words
TSCC v2	570	788k	1 074	786k
CIMA	1 135	44k	1 135	38k
MathDial	2 861	923k	2 876	879k
Multicultural	5	614k	643	614k
Uptake	774	35k	775	34k
<b>Total</b>	5 345	2 404k	6 503	2 351k

Table 1: Datasets used to train the empathetic teacher. Number of conversation turns and words in the original datasets and after splitting long conversations.

### 3.8 Orchestrator

The orchestrator is a simple Python program that iterates through the different errors and sentences to discuss them with the user during the debriefing session. For the current target error and sentence, the orchestrator prepares a complex prompt that includes the original sentence, the corrected sentence, the current error to review, the error annotation, the explanation of the error, the related information extracted from textbooks, the hints of the empathetic teacher’s response, the short-term conversation history and a summary of the mid-to-long-term most relevant topics discussed with the user. Note that many of these items are by-products of the components described above. The orchestrator takes also into consideration the directives of the knowledge tracing module (see below) as regards the current state of the conversation flow and the user’s understanding level. This long prompt will then be fed to the next-dialog-line generator.

<sup>22</sup><https://huggingface.co/Transducens/empathetic-teacher>

The prompt consists of several key parts. First, the chatbot receives instructions on guiding the user through explanations, examples, and exercises to address the errors. It also outlines a set of user *intentions* for different interaction stages, with specific actions the chatbot must take for each. The prompt also includes the items generated in the pre-processing pipeline. Finally, concise instructions guide the chatbot to identify the user’s intention, generate a suitable response, and output both following a JSON template. Additional guidelines ensure that responses are brief, engaging, and flexible beyond the provided data.

### 3.9 Knowledge Tracing

In order to guide the conversational flow, we first considered a traditional strategy based on a state transition model, with states representing the user’s position in the learning path. However, we later found that the language model could manage the conversation flow autonomously via in-context learning and intention detection, without the need for extensive external intervention to track the dialog state. Transitions are therefore naturally handled by the LLM, based on the user’s responses. Errors are prioritized based on frequency, according to the ERRANT’s classes.

### 3.10 Next-Dialog-Line Generator

Although all LLMs used in the previously discussed components of DeMINT are implemented as local open-weight models, our preliminary experiments show that the best results are achieved when the next-dialog-line generator in particular is a more powerful LLM. Currently, GPT-4 accessed via the OpenAI API<sup>23</sup> is our preferred choice for this task.<sup>24</sup> This component faces the challenging task of generating the next line of the conversation based on the informative prompt prepared by the orchestrator. The output of this generator is then presented to the user as the chatbot’s response.

### 3.11 Chatbot Interface

The interface is a simple web app built with *gradio*.<sup>25</sup> It shows the chatbot conversation in one column and the transcription, centered on the

<sup>23</sup><https://openai.com/api>

<sup>24</sup>In particular, we use the `gpt-4o-2024-08-06` model, which, in addition to being one of the most powerful LLMs available today, includes the built-in feature *structured outputs* that enforces the generation of outputs in a specific JSON schema, thereby simplifying the ensuing parsing.

<sup>25</sup><https://github.com/gradio-app/gradio>

current sentence, in another. The user types their input, and the machine responds accordingly on the screen.

## 4 Human Evaluation

A preliminary evaluation<sup>26</sup> has been conducted through interactions between the chatbot and L1-Spanish/L2-English students. These students have been recruited through the Languages Service of our university, which maintains a pool of students registered for activities related to multilingualism promotion. This service retains information regarding the students’ backgrounds, native languages, proficiency levels in languages, etc. Among the students willing to participate in this evaluation, 7 participants were selected, each dedicating approximately 10 hours to evaluation activities. We targeted students with B2/C1 levels of English according to the Common European Framework of Reference for Languages, and aimed to create a balanced group in terms of gender and diversity of backgrounds.

Fifteen video calls of approximately 10 minutes were organized among the selected students, with two or three participants per call. We employed role-playing games, specifically designed to engage students in English conversations. Role-playing games help avoid the difficulties associated with anonymizing real online meetings and allow us to control the topics and complexity of the conversations. Specifically, we have used the materials designed by Pitts (2015), which provide the context for the role-playing games, as well as preparatory questions to help students familiarize themselves with the topic. Students were given time to prepare for the online meeting. These video calls were recorded, and students then participated in a debriefing session with our chatbot to analyze errors in their use of English during the online meeting. Finally, students completed a survey to evaluate their interaction with the chatbot.

Feedback from the human evaluation addressed two main areas: overall user experience and the chatbot’s effectiveness as an English tutor, with responses rated on a Likert scale from 1 to 5. Regarding the first aspect, participants were generally satisfied with the tool’s performance and response time. In response to the question, “*Did you enjoy interacting with the chatbot?*”, all participants gave positive feedback, with a score of 4

<sup>26</sup>The empathetic teacher was disabled during evaluation.

or 5. However, fluency emerged as the system’s main area requiring improvement, with an average score of 3. In terms of the chatbot’s performance as an intelligent English tutor, the overall evaluation was positive, though some areas still require enhancement. The main concern of the participants in this evaluation was the accuracy of the chatbot in identifying speech errors, which received an average of 3. Other aspects, such as the chatbot’s ability to understand their queries, or the usefulness of examples and resources provided by the chatbot, were rated with an average score of 3.3. The clarity of the chatbot’s error explanations received a slightly higher average score of 3.4. Notably, most participants agreed that the chatbot helped improve certain aspects of their English, with five out of seven giving a score of 4 for this question. Additionally, when asked whether they would be interested in using a similar chatbot in future video conferences, all participants but one gave scores of 4 or 5, demonstrating a general interest in this kind of tools.

The audio recordings from the online meetings, descriptions of the role-playing activities, and the corresponding transcriptions are available as part of the English Learners Role-Playing Dialogue Dataset (ELRD), released under a CC license.<sup>27</sup>

Although we do not plan to involve human English teachers in the near future to evaluate the system’s error detection capabilities or the interaction between chatbot and students from a teacher’s perspective, we are considering this for later stages.

## 5 Conclusions

In this paper, DeMINT, an innovative conversational intelligent tutoring system designed to enhance English proficiency of non-native speakers through the analysis of online meeting transcriptions, has been presented. Our system leverages the latest advancements in LLMs and integrates various techniques such as in-context learning, retrieval augmented generation, grammatical error correction, and error-preserving speech synthesis and generation. We have provided a comprehensive overview of the system’s architecture, including modules for diarization, speech recognition, error correction and annotation, error explanation, knowledge tracing and chatbot orchestration. A pilot evaluation of the system’s effectiveness through controlled interactions with

L2-English students has been carried out utilizing role-playing games to simulate real-life conversations. Our ultimate goal is to create a scalable, accessible tool that mimics the guidance of a human tutor, providing personalized and context-aware feedback to help non-native speakers improve their language skills by conveniently leveraging their everyday interactions in English. The code, data, and models developed for this project have been openly released across various repositories to promote further research in the field. The central code repository<sup>28</sup> contains links to the additional datasets and models.

Despite being a work-in-progress, we already foresee some future developments. Potential enhancements include supporting voice cloning with tools such as XTTS-v2 (Casanova et al., 2024) so that the error-preserving STT model can be fine-tuned with each user’s voice before the debriefing session. Another line of future research involves integrating conversational interaction with users through speech, thus helping students to improve not only their grammatical skills but also their pronunciation. Most components of the system will likely benefit from new emerging models and techniques; for example, for the error explanation module, very recent end-to-end systems that provide error explanations such as xTower (Treviso et al., 2024) are worth exploring. Additionally, multimodal models could be investigated to integrate the non-verbal aspects of online meetings, such as facial expressions and body language. Finally, another area of future work is to conduct an ablation study to determine the relevance of each component within the overall system and explore their potential replacement by a more advanced prompting strategy on the final LLM model.

**Acknowledgments.** DeMINT (Automated Language Debriefing for English Learners via AI Chatbot Analysis of Meeting Transcripts) is a project funded via FSTP (financial support to third parties), a mechanism by the European Union to support smaller projects through grants provided by larger, EU-funded initiatives. DEMINT is funded under the UTTER<sup>29</sup> (Unified Transcription and Translation for Extended Reality) project, a collaborative Research and Innovation project under Horizon Europe, grant agreement 101070631.

<sup>27</sup><https://github.com/transducens/elrd>

<sup>28</sup><https://github.com/transducens/demint>

<sup>29</sup><https://he-utter.eu/>

## Ethics

Since the human evaluation involves collecting and distributing data from participants, special care has been taken to adhere to relevant ethical guidelines<sup>30</sup> and applicable data protection laws. Specifically, the research ethics committee of our university has overseen the experimental process. Each participant was informed about how their interaction with the model would be used and disseminated, and they signed a consent form. Additionally, participants’ personal information has been pseudonymized in the released data.

## Limitations

Our current system has several limitations that we are aware of. First, the system is currently designed to work with L1-Spanish/L2-English students. Although the system could be adapted for other languages, this would require additional fine-tuning of the models and the incorporation of language-specific resources. Additionally, the system is currently designed to provide feedback on grammar errors and language usage, but it does not address other aspects of language learning such as vocabulary acquisition or pronunciation. Achieving fluency in the communication with the chatbot poses a significant challenge, and the system may fall short of reaching the spontaneity of a human tutor. Finally, some users may prefer reviewing a report over interacting with a chatbot, as non-native speakers are often aware of many errors caused by the improvisation required during conversation, which they would not make in writing.

## A Fine-tuning hyperparameters

**Empathetic teacher.** To fine-tune Llama-3.1-8B to function as the generic teacher described in Section 3.7, we employed the parameter-efficient 8-bit QLoRA method (Dettmers et al., 2023) using a single A100 GPU with 80 GB of VRAM and the LLaMA-Factory toolkit.<sup>31</sup> The LoRA configuration was set to  $r = 8$ ,  $\alpha = 16$ , with no dropout applied, and targeting all linear modules. Flash Attention version 2 was used (Dao, 2023), and the sequence length was limited to 4 096 tokens. The learning rate was set to  $10^{-4}$  and then adjusted with a linear learning rate scheduler with 10 warmup steps. The training batch size was 12,

<sup>30</sup><https://www.acm.org/code-of-ethics>

<sup>31</sup><https://github.com/hiyouga/LLaMA-Factory>

and weights were updated after each minibatch. We used the AdamW optimizer with hyperparameters  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 10^{-8}$ , while capping the maximum gradient norm at 1.0. The best model was obtained after 2 900 training steps, achieving a cross-entropy loss of 1.83.

	FT	D-FT	W	D-W
Our test set	31.47	38.81	41.82	39.48
Peoples Speech	47.05	30.77	39.45	40.02
Parler-tts	13.70	15.93	26.26	8.63
mlls-eng-10k	13.89	15.37	7.34	8.11
Fleurs	13.12	14.98	16.83	17.43

Table 2: WER results on test sets for the best fine-tuned models and original Whisper models. FT: fine-tuned Whisper, D-FT: fine-tuned distilled Whisper, W: original Whisper, D-W: distilled Whisper.

**Error-preserving speech-to-text model.** As regards the error-preserving speech-to-text model discussed in Section 3.2, we employed a fine-tuning approach using LoRA (Hu et al., 2022) and some specific training arguments to fine-tune the original Whisper model<sup>32</sup> and one distilled version.<sup>33</sup> The configuration for LoRA was set with  $r = 16$ ,  $\alpha = 32$ , targeting the modules `q_proj` and `v_proj`. Additionally, no dropout was applied, and no bias was included. We fine-tuned the models on one GPU RTX A6000 with 48 GB of VRAM. For the training arguments, the training batch size was set to 8 for the original model and 28 for the distilled one (its smaller size allowed for a larger batch size). Parameters were updated after each minibatch. We used the Adam optimizer with hyperparameters  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  and  $\epsilon = 1e - 8$ , and a linear learning rate scheduler with 50 warmup steps. The learning rate was set to  $10^{-5}$ . The fine-tuning was run for 7 500 steps in the case the original Whisper model, and 7 000 steps in the case of the distilled one. The model parameters were saved every 500 steps, and evaluations were also conducted every 500 steps. At the end of the training, the best model was chosen based on the lowest word error rate (WER) upon the validation set.<sup>34</sup> Table 2 shows the scores of the best models on different test sets.

<sup>32</sup><https://huggingface.co/openai/whisper-large-v3>

<sup>33</sup><https://huggingface.co/distil-whisper/distil-large-v3>

<sup>34</sup>The selected models had a WER of 12.14 for the original Whisper model and 18.10 for the distilled one.

## References

- Nathalie Aichhorn and Jonas Puck. 2017. “I just don’t feel comfortable speaking English”: Foreign language anxiety as a catalyst for spoken-language barriers in MNCs. *International Business Review*, 26(4):749–763.
- Serge Bibauw, Thomas François, and Piet Desmet. 2019. Discussing with a computer to practice a foreign language: research synthesis and conceptual framework of dialogue-based CALL. *Computer Assisted Language Learning*, 32(8):827–877.
- Hervé Bredin. 2023. `pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe`. In *Proc. INTERSPEECH 2023*.
- Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.
- Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.
- Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical Error Correction: A Survey of the State of the Art. *Computational Linguistics*, 49(3):643–701.
- Andrew Caines, Helen Yannakoudakis, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2022. The Teacher-Student Chatroom Corpus version 2: more lessons, new annotation, automatic detection of sequence shifts. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 23–35, Louvain-la-Neuve, Belgium. LiU Electronic Press.
- Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Gökner, Iulian Gulea, Logan Hart, Aya Al-jafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: a massively multilingual zero-shot text-to-speech model. `arXiv:2406.04904 [eess.AS]`.
- Anna Izabela Cisłowska and Beatriz Peña Acuña. 2024. Integration of chatbots in additional language education: A systematic review. *European Journal of Educational Research*, 13(4):1607–1625.
- Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.
- Tri Dao. 2023. `Flashattention-2: Faster attention with better parallelism and work partitioning`. `arXiv:2307.08691 [cs.LG]`.
- Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring Conversational Uptake: A Case Study on Student-Teacher Interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient fine-tuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. `arXiv:2401.08281 [cs.LG]`.
- Jinming Du and Ben Kei Daniel. 2024. Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice. *Computers and Education: Artificial Intelligence*, 6:100230.
- Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. Enhancing grammatical error correction systems with explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7489–7501, Toronto, Canada. Association for Computational Linguistics.
- Mariano Felice, Christopher Bryant, and Ted Briscoe. 2016. Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835, Osaka, Japan. The COLING 2016 Organizing Committee.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Weijiao Huang, Khe Foon Hew, and Luke K. Fryer. 2022. Chatbots for language learning: Are they really useful? a systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1):237–257.
- Jaeho Jeon. 2024. Exploring AI chatbot affordances in the EFL classroom: young learners’ experiences and perspectives. *Computer Assisted Language Learning*, 37(1-2):1–26.

- Jiyou Jia. 2009. CSIEC: A computer assisted English learning chatbot based on textual knowledge and reasoning. *Knowledge-Based Systems*, 22(4):249–255.
- Masahiro Kaneko and Naoaki Okazaki. 2024. Controlled generation with prompt insertion for natural language explanations in grammatical error correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3955–3961, Torino, Italia. ELRA and ICCL.
- Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. [arXiv:2212.14024](https://arxiv.org/abs/2212.14024) [cs.CL].
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. DSPy: Compiling declarative language model calls into self-improving pipelines. [arXiv:2310.03714](https://arxiv.org/abs/2310.03714) [cs.CL].
- Lasha Labadze, Maya Grigolia, and Lela Machaidze. 2024. Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education*, 21(1):28.
- Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 19594–19621. Curran Associates, Inc.
- Cristóbal Lozano, Ana Díaz-Negrillo, and Marcus Callies. 2020. Designing and compiling a learner corpus of written and spoken narratives: COREFL. In Christiane Bongartz and Jacopo Torregrossa, editors, *What’s in a Narrative? Variation in Story-Telling at the Interface between Language and Literacy*, pages 21–46. Peter Lang.
- Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.
- Janick Michot, Manuela Hürlimann, Jan Deriu, Luzia Sauer, Katsiaryna Mlynchyk, and Mark Cieliebak. 2024. Error-preserving Automatic Speech Recognition of Young English Learners’ Language. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6444–6454, Bangkok, Thailand. Association for Computational Linguistics.
- Marvin Minsky. 1986. *The Society of Mind*. Simon & Schuster, New York.
- Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning sns for automated Japanese error correction of second language learners. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 147–155.
- Amr M. Mohamed. 2024. Exploring the potential of an AI-based chatbot (ChatGPT) in enhancing English as a foreign language (EFL) teaching: perceptions of EFL faculty members. *Education and Information Technologies*, 29(3):3195–3217.
- Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.
- Catherine Nickerson. 2005. English as a lingua franca in international business contexts. *English for Specific Purposes*, 24(4):367–380.
- Kostiantyn Omelianchuk, Andrii Liubonko, Oleksandr Skurzhanskyi, Artem Chernodub, Oleksandr Kornienko, and Igor Samokhin. 2024. Pillars of grammatical error correction: Comprehensive inspection of contemporary approaches in the era of large language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 17–33, Mexico City, Mexico. Association for Computational Linguistics.
- Larry Pitts. 2015. *ESL Role Plays: 50 Engaging Role Plays for ESL and EFL Classes*. ECQ Publishing.
- Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.



- Muhammad Reza Qorib and Hwee Tou Ng. 2023. [System combination via quality estimation for grammatical error correction](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12746–12759, Singapore. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Chrysi Rapanta, Cláudia Gonçalves, João Rui Pereira, Dilar Cascalheira, Beatriz Gil, Rita Morais, Anna Čermáková, Julia Peck, Benjamin Brummernhenrich, Regina Jucks, Mercè Garcia-Milà, Andrea Miralda-Banda, José Luna, Maria Vrikkki, Maria Evagorou, and Fabrizio Macagno. 2021. [Multi-cultural classroom discourse dataset on teachers’ and students’ dialogic empathy](#). *Data in Brief*, 39:107518.
- Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. 2022a. [PLAID: An efficient engine for late interaction retrieval](#). In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM ’22*, page 1747–1756, New York, NY, USA. Association for Computing Machinery.
- Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022b. [ColBERTv2: Effective and efficient retrieval via lightweight late interaction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.
- Maganat Shegebayev. 2023. [Rise of English as business lingua franca at the turn of the century: An overview](#). In Stanley D. Brunn and Roland Kehrein, editors, *Language, Society and the State in a Changing World*, pages 357–365. Springer International Publishing, Cham.
- Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. [BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage](#).
- Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2024. [GEE! grammar error explanation with large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 754–781, Mexico City, Mexico. Association for Computational Linguistics.
- Felix Stahlberg and Shankar Kumar. 2021. [Synthetic data generation for grammatical error correction with tagged corruption models](#). In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47. Online. Association for Computational Linguistics.
- Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020. [CIMA: A large open access dialogue dataset for tutoring](#). In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64, Seattle, WA, USA → Online. Association for Computational Linguistics.
- Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. [Tense and aspect error correction for esl learners using global context](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 198–202.
- Yuki Takashima, Yusuke Fujita, Shinji Watanabe, Shota Horiguchi, Paola García, and Kenji Nagamatsu. 2021. [End-to-end speaker diarization conditioned on speech activity and overlap detection](#). In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 849–856.
- Marcos Treviso, Nuno M. Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan van Stigt, and André F. T. Martins. 2024. [xTower: A multilingual LLM for explaining and correcting translation errors](#). arXiv:2406.19482 [cs.CL].
- Hyejin Yang, Heyoung Kim, Jang Ho Lee, and Dongkwang Shin. 2022. [Implementation of an AI chatbot as an English conversation partner in EFL speaking classes](#). *ReCALL*, 34(3):327–343.
- Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. [A new dataset and method for automatically grading ESOL texts](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.
- Shunan Zhang, Cheng Shan, John Sie Yuen Lee, ShaoPeng Che, and Jang Hyun Kim. 2023. [Effect of chatbot-assisted language learning: A meta-analysis](#). *Education and Information Technologies*, 28(11):15223–15243.
- Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader

Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piękos, Aditya Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanić, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. 2023. [Mindstorms in natural language-based societies of mind](#). In *RO-FoMo: Robustness of Few-shot and Zero-shot Learning in Foundation Models*, *NeurIPS 2023*.

# Evaluating the Generalisation of an Artificial Learner

Bernardo Stearns<sup>1</sup> and Nicolas Ballier<sup>2</sup> and Thomas Gaillat<sup>3</sup>  
and Andrew Simpkin<sup>4</sup> and John P. McCrae<sup>1</sup>

<sup>1</sup> Insight Centre for Data Analytics, Data Science Institute, University of Galway, Ireland

<sup>2</sup> LLF & CLILLAC-ARP / Université Paris Cité, rue Thomas Mann, 75013 PARIS, France

<sup>3</sup> LIDILE / Université de Rennes 2, 35000 Rennes, FRANCE

<sup>4</sup> School of Mathematical and Statistical Sciences, University of Galway, University Road, Galway, Ireland

Contact: [bernardo.stearns@insight-centre.org](mailto:bernardo.stearns@insight-centre.org)

## Abstract

This paper focused on the creation of LLM-based artificial learners. Motivated by the capability of language models to encode language representation, we evaluated such models for predicting masked tokens in learner corpora.

We domain-adapted the BERT model, pre-trained on native English, by further pre-training two learner models on learner corpora: a natural learner model on the EFCAM-DAT dataset and a synthetic learner model on the C4200m dataset. We evaluated the two artificial learner models alongside the baseline native model using an external English-for-specific-purposes corpus from French undergraduates.

We evaluated metrics related to accuracy, consistency, and divergence. While the native model performed reasonably well, the natural learner pre-trained model showed improvements in recall-at-k. We analysed error patterns, showing that the native model made “overconfident” errors by assigning high probabilities to incorrect predictions, while the artificial learners distributed probabilities more evenly when wrong. Finally, we showed that the general token choices from the native model diverged from the natural learner model and this divergence was higher at lower proficiency levels.

## 1 Introduction

Over the last 20 years, learner corpora have significantly benefited research in applied linguistics and NLP by providing insights into how sec-

ond language (L2) learners improve their proficiency. This understanding has led to enhanced course material design, improved teacher training, and greater awareness of students’ linguistic abilities. Additionally, when combined with NLP technologies, learner corpora have proven valuable for CALL applications like grammar error detection and proficiency classification (Bryant and Briscoe, 2018; Tetreault et al., 2018). This paper explores the potential of leveraging Large Language Models (LLMs) with learner corpora, which have traditionally been used to test specific research hypotheses. Instead of relying on diverse corpora with relevant metadata for testing various hypotheses, we explore the possibility of a single model that simulates learner behavior across different contexts. Such artificial learners could respond to new stimuli, providing a testbed for linguistic hypotheses, with outputs from a generic English learner model compared to those from a native model. By training an LLM on learner data, it may be possible to create an artificial English learner that captures the idiosyncrasies of actual learners.

This research explored the creation of an Artificial L2 Learner (ALL) model by pre-training it on second language learner corpora, leveraging domain-adaptive pre-training. We also believe that modelling learners’ knowledge and their use of words and linguistic skills is crucial for Intelligent Tutoring Systems (ITS) and digital learning platforms in second language teaching and learning. For an ITS focused on language learning, modelling word usage and language skills of learners is essential. This is why any simulation of learner behavior, as a key goal for an ITS, should be accurate and reliable. Motivated by the capability of

---

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

language models to represent linguistic concepts, this research explored the domain-adaptive pre-training of large language models (LLMs) to simulate the behavior of English learners, which we call Artificial Learner models. Creating an artificial learner raises at least three questions:

1. How accurate is the artificial learner in predicting what learners would actually say?
2. How confident is the learner in its predictions?
3. How divergent is the AL compared with a generic native model?

The rest of the paper is structured as follows: Section 2 presents related research. Section 3 explains the training data and the procedures used to create our artificial learners. Section 4 delves into our results, and Section 5 provides a discussion of these results.

## 2 Background Research

Research in second language acquisition has been explored from many different perspectives, resulting in different models for each aspect of the learning process. For example, [Whitehill and Movellan \(2017\)](#) models learners taking into account how a learner infers and updates vocabulary knowledge after doing exercises in a specific ITS for foreign language learning. The SLAM shared task ([Settles et al., 2018](#)) models the history of a learner’s mistakes in Duolingo, predicting if a learner is likely to make a mistake given their past history of mistakes. There are also models that are complementary to modelling the second language acquisition process, such as spaced repetition practice models ([Settles and Meeder, 2016](#)) and efficient grammatical error correction ([Omelianchuk et al., 2020](#)). Despite the success of such diverse tasks in their specific modelling objectives, the usage of their models is tied to the specific case of their system or language learning task. This restricts the capability of such models to simulate the general behavior of language learners.

There is another set of language-learning tasks that explicitly model learners’ behavior and knowledge however, they are still tied to a single task depending on handcrafted features. Examples include [Whitehill and Movellan \(2017\)](#), which models vocabulary learning from concepts;

[Knowles et al. \(2016\)](#), which models noun understanding from the context of the native language; and [Zylich and Lan \(2021\)](#), which models retrieval practice performance for SLA based on linguistic and memory-based features. Other similar modeling tasks include [Avdiu et al. \(2019\)](#); [Renduchintala et al. \(2016\)](#). In a similar fashion, corpus linguists have also developed single tasks aimed at predicting specific outcomes in the form of linguistic constructions. [Bresnan and Nikitina \(2009\)](#) modelled the dative alternation, where learners hesitate between the prepositional dative structure or the double object structure. [Gries et al. \(2020\)](#) approaches in corpus linguistics also reflect this method by modeling the genitive vs. *noun\_of\_noun* construction. Modelling construction outcomes in learner texts helps understand the contexts, triggering constructions. Nevertheless, these models cannot handle different sets of constructions, which appears to be a limitation if one wants to analyze many different linguistic systems at the same time. In contrast, large language models (LLMs) are capable of accommodating diverse constructions and analyzing multiple linguistic systems simultaneously, offering a more flexible approach to understanding language patterns.

In the broader field of Natural Language Processing, language models have been effectively adapted to multiple domains and tasks using a single generic model, in a similar scenario we see in the Second Language Acquisition domain. [Gururangan et al. \(2020a\)](#) examines the effectiveness of adapting pre-trained language models to multiple domains and tasks with a single model. They test how well a task-specific fine-tuned model transfers to different types of other tasks, showing a large gain in task performance using an overall multi-phase domain and a task-adaptive pre-trained model. Though we see an underutilization of language models in learner modelling tasks, many other diverse areas have successfully adapted language models to their tasks.

To the best of our knowledge, two tasks analysed the potential of language models in SLA. [Palenzuela et al. \(2022\)](#) explored native pre-trained language models to predict language mistakes in the SLAM shared task. [Kim \(2024\)](#) investigated the use of language models as “artificial English learners” with a model called Bidirectional Encoder Representations from Transform-

ers (BERT). They specifically tested BERT’s ability to simulate English learners’ usage of prepositions. Notably, BERT was domain-adaptively pre-trained on the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2013). The study focused on how this artificial learner utilized four English prepositions: *at*, *for*, *in*, and *on*.

Our work proposes a generalized analysis of artificial English learners, which expands the scope of previous analysis by introducing a broader range of metrics, including accuracy, consistency, and behavior validation. The goal is to establish trust in the trained models before exploring their capabilities in specific tasks.

### 3 Material and Methods

#### 3.1 Data

##### 3.1.1 Training data

**EFCAMDAT corpus** - We trained two artificial learner models. The first model was trained on the EFCAMDAT. We used the refined version of the EFCAMDAT corpus texts (Shatz, 2020). It includes 723,282 writings from *Englishtown* language schools (Shatz, 2020).

The learners wrote texts following prompts such as “introducing yourself by email”. Students gradually moved from one level to the next based on language teachers’ grades. The writings span across 16 proficiency levels, which were mapped to the first five CEFR levels. The CEFR levels of the texts correspond to the successful completion of the coursework levels at *Englishtown*.

**C4200M corpus** - The second model was trained on the C4200M corpus (Stahlberg and Kumar, 2021). It is a corpus of synthetically generated ungrammatical sentences used in neural grammatical error correction. This model produces an ungrammatical sentence given a clean sentence and an error type tag following the tags defined in the ERRANT automatic annotation tool (Bryant et al., 2017). The generated ungrammatical sentences follow the distribution of error tags in the BEA-dev dataset (Bryant et al., 2019). They argue for the utility of the generated ungrammatical data by pre-training grammar error correction models with it, outperforming genuine parallel data on the CONLL-2014 and JFLEG-test.

We chose the C4200m with the goal of analysing a common trade-off in the training process of large language models: balancing the qual-

ity of authentic texts versus the quantity of augmented texts, similar to works surveyed in Feng et al. (2021). We aimed to understand how this trade-off affects the performance of artificial English learners. By using the C4200m dataset, we wanted to see how different amounts of high-quality and lower-quality texts impact the learning results of our models. This would help us understand the best balance between text quality and quantity for training large language models. Our approach aligns with other NLP research, providing a comparative view that adds to the relevance and usefulness of our findings.

##### 3.1.2 Testing data

The external test set (see Table 1) is made up of learner writings from the CELVA-SP (Mallart et al., 2023) a corpus of French undergraduates using English for specific purposes (ESP). Learners answered one of three question prompts as part of a 45-minute in-class writing task. For instance, they had to describe and share their opinion on the most important invention in their field. All their writings were subsequently annotated with the writing competence scale of the CEFR (Council of Europe, 2018, Appendix 4, p.187-189) by four expert raters. Pairwise inter-rater agreement was computed on the basis of 60 writings, yielding Cohen’s kappa values ranging from .52 to .72. The rest of the writings were then annotated independently. Table 1 presents the distribution of the levels in CELVA-SP data.

#### 3.2 Data processing

Processing the learner texts for our analysis involved two types of data processing. First, for the model training, we simply passed the raw texts as input to a masked language modelling collator, following the standard masking strategy used in the training process of BERT (Devlin et al., 2019). The collator dynamically generates batches of masked sentences, which the BERT tokenizer processes into WordPiece tokens for use in the training loop.

Second, for prediction analysis, we used a Universal Dependency (UD) tokenizer (Nivre et al., 2016) to represent “human” learner tokens. We masked each token in the text one at a time, creating a unique masked sentence for every UD token. These sentences with a single masked token were then fed to our artificial learners and the baseline native model to predict token usage. We annotated

Table 1: Distribution of levels and essays in the CELVA-SP data (Mallart et al., 2023)

Writings	# of writings	% of writings	av # of words	SD
A1	85	8.70	126.78	76.67
A2	311	31.83	182.02	87.21
B1	335	34.28	231.34	111.70
B2	198	20.26	285.84	126.75
C1	48	4.91	347.93	144.69
Total	977	100	224.11	120.64

the part of speech for each UD token using UD-Pipe (Straka, 2018) implementation in spaCy<sup>1</sup>. It allowed us to visualize the distribution of probability scores across different parts of speech for the natural learner model. Since our experiment focused on the BERT base model and its limitation of 512 WordPiece tokens, we filtered out texts with more than 512 such tokens.

### 3.2.1 Domain-Adaptive Pre-training

The main step in developing the two proposed artificial learner models was the domain-adaptive pre-training of an already pre-trained baseline BERT model. We used the EFCAMDAT as a training set for the natural learner model, and the C4200m as a training set for the synthetic learner model. We trained both artificial learners on a masked language modelling task. In Devlin et al. (2019) they refer to pre-training as training a model on unlabeled data across various tasks, such as masked language modelling, where fine-tuning involves initializing a pre-trained model’s weights and updating them using labeled data. We initialized a baseline BERT model weights and further pre-trained them in learner corpus in an unsupervised masked language modelling task. This is referred in (Gururangan et al., 2020b) as domain-adaptive pre-training.

We used the same masked language modelling pre-training task described in Devlin et al. (2019). Specifically, we masked 15% of WordPiece tokens in each sentence of the training set, allowing the model to learn contextual representations by predicting the masked tokens.

## 3.3 Evaluation

To evaluate the predictions of the two artificial learner models and the native baseline model, we

<sup>1</sup>You can find the repository at <https://github.com/TakeLab/spacy-udpipe>.

used three types of metrics: recall-at-k, KL divergence, and calibration. We calculated the metrics on the CELVA-SP dataset.

### 3.3.1 Accuracy with recall-at-k

We used the recall-at-k metric as our accuracy measure. It naturally extends the concept of accuracy by taking into account the model’s top-k potential responses and explicitly consider a criteria for relevant responses that could be easily extended. In essence, we measured on average how many of the top-k token predictions recommended by a given model were relevant for the target masked token used by the learner.

The recall-at-k metric evaluates the top-k responses of a model that generates a list of potential responses  $\hat{y}$  to a given query  $q$ , ranked by their likelihood of being correct according to the model. In our experiment, for a given masked token sentence the query  $q$  is the actual masked token used by the learner, and the list of potential responses  $\hat{y}$  is the list of tokens predicted by a model ranked by probability in the softmax layer of BERT vocabulary.

For a target masked token  $q_i$  and a top-k token  $t_j$  predicted by the model,  $t_j$  is considered relevant to  $q_i$  simply if  $t_j$  is in the set of relevant items for  $q_i$ . In our experiment, the only relevant item was the target masked token itself, so this is equivalent to verifying if  $q_i$  is in the top-k predictions but this would not be the case in more complex scenarios.

We report the average recall@k over all masked tokens in the CELVA-SP for each of the three evaluated models.

$$\text{AVG Recall}@k = \frac{\sum_{q_i \in \text{masked\_tokens}} 1[q_i \in \text{top-k}(\hat{y})]}{\# \text{ of masked tokens}}$$

We report recall for  $k = [1, 5, 10]$ .

### 3.3.2 Kullback–Leibler metric

The Kullback-Leibler divergence is rooted in information theory and provides a general approach for quantifying how two probability distributions differ. We framed each of our models’ output probabilities for a given masked token as a discrete probability distribution over BERT’s vocabulary tokens. Within that frame, we interpreted the KL metric for two given models as if their token choices generally diverged. We implemented the element-wise KL metric with a small epsilon value perturbation,  $\epsilon = 10^{-6}$ , to avoid the scenario where probabilities are zero. We calculated the KL element-wise metric for each masked token, and we grouped them by their text CEFR level with the intuition to find differences between CEFR levels.

$$KL(p_t, q_t) = p_t \log \left( \frac{p_t + \epsilon}{q_t + \epsilon} \right)$$

### 3.3.3 Calibration Curves

To foster trustworthiness in our models, high accuracy is the immediate desired property of our models, assigning high probabilities to correct tokens. A second desirable property is that our models do not overconfidently make mistakes, assigning high probabilities to incorrect predictions.

One approach for such analysis is through the “calibration curve” method. Initially employed in analysing weather forecasts (Brier, 1950; DeGroot and Fienberg, 1983), this technique has since been applied to neural networks (Guo et al., 2017; Minderer et al., 2021) and recently to evaluate Large Language Models (LLMs) from a semantic perspective (Levinstein and Herrmann, 2024). For example, (Levinstein and Herrmann, 2024) utilizes calibration curves to assess the veracity of LLM statements on specific datasets and asserts that “calibration offers another metric for evaluating the quality of probes’ forecasts.” Calibration analyses have been utilized in neural networks and language models (Minderer et al., 2021; Chen et al., 2024), allowing researchers to assess the relationship between a model’s prediction confidence and success rate.

Calibration curves help us analyze how well a model performs when it is confident or unconfident about its prediction. In our experiment, our calibration curves correspond to how many successful predictions (event rate) we observe across different probability scores of the top-1 prediction of each model.

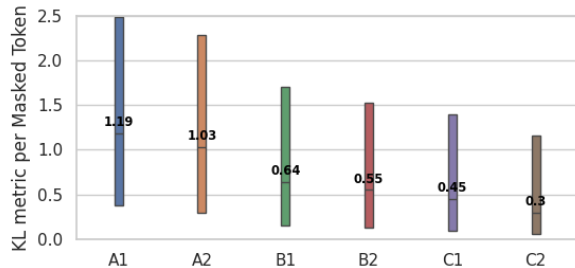


Figure 1: Interquartile range plot of KL metric between natural learner and native model per masked token sentence grouped by CEFR level in the CELVA-SP dataset as described in 3.3.2

$$\text{Event Rate} = \frac{\text{Number of Successful Predictions}}{\text{Total Number of Predictions}}$$

## 4 Results

### 4.1 Recall-at-k

We evaluated the accuracy of our models with recall-at-k metrics. We found a slight difference in accuracy between the Learner Models and the native model in the external CELVA-SP test set. We noticed a slow increase in recall as k increases. A slow increase in the values of top-k recall may indicate that the token vocabulary of the language model is not adequate for the task. We believe it is unlikely that the model is confused when choosing among 10 or more tokens; instead, the correct token is likely represented by multiple word-piece tokens in the model’s vocabulary.

model	recall@1	recall@5	recall@10
bert-native (baseline)	0.600	0.622	0.635
bert-efcamdat	0.648	0.670	0.684
bert-c4200m	0.586	0.610	0.623

Table 2: Average recall-at-k in the CELVA-SP for each evaluated model as described in section 3.3.1

### 4.2 KL Distance

The KL metric interquartile plot in Figure 1 presents the KL metric between native BERT and the natural learner model. It allowed us to analyse the intuition that a learner model will generally differ from a native model in terms of token usage and that this difference is higher in beginner texts. The figure indicates that the learner model exhibits greater disagreement in token choice for masked sentences at lower proficiency levels, with a monotonic decrease in disagreement as proficiency increases.

### 4.3 Calibration Curves

The calibration curve in Figure 2 illustrates the relationship between the predicted probabilities of the top candidate token and the success rate at which these tokens correctly predict the true token. The three models follow a linear trend, showing that all of them classify more accurately as their top-1 token probability increases, suggesting that they are well-calibrated overall. However, the EFCAMDAT curve shows a discrepancy for probabilities around 0.6. Specifically, the natural artificial learner demonstrates underperformance in this range, as candidates predicted with a 60% probability only successfully predict the true token 40% of the time but increase and become slightly higher for probabilities close to 1. In general, the natural learner model outperforms the native model in the range of higher top-1 probabilities. This analysis can be further supported by Figure 3 where we noticed that the native model (on the right side of the figure) very frequently assign high probabilities to its top-1 prediction where the two artificial learners assign lower probabilities. Even though the native model assigns higher top-1 probabilities more frequently, it has a lower success rate than the natural learner model. One possible explanation for the learner model’s underperformance in the 60% probability range is that the masked tokens in this range likely come from advanced learners’ texts, whereas the EFCAMDAT dataset primarily consists of beginner learners. This motivates a detailed analysis of the performance of such models across CEFR levels as future work.

## 5 Discussion

### 5.1 Role of Part of Speech

Parts of speech (POS) provide a way to filter out the prediction distribution. It is possible to analyse the behaviour and success rate of the artificial learners according to linguistic properties related to not only the lexicon but also grammar. For instance, filtering out probabilities per auxiliary gives an insight into a closed class. This helps characterize the impact of universal part-of-speech (UPOS) on the probability distributions of the probability scores for the first three predictions (rank) across the three models. For example, Table 3 shows the average probability score assigned by a given model to its top-3 predictions, as well as the respective success rate for masked prepositions. We observe a similar pattern, where the na-

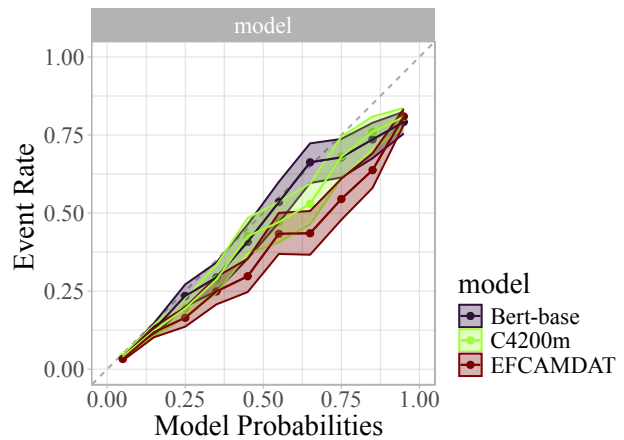


Figure 2: Success event rate across top-1 token model probabilities for all 3 models across all masked tokens in the CELVA-SP data

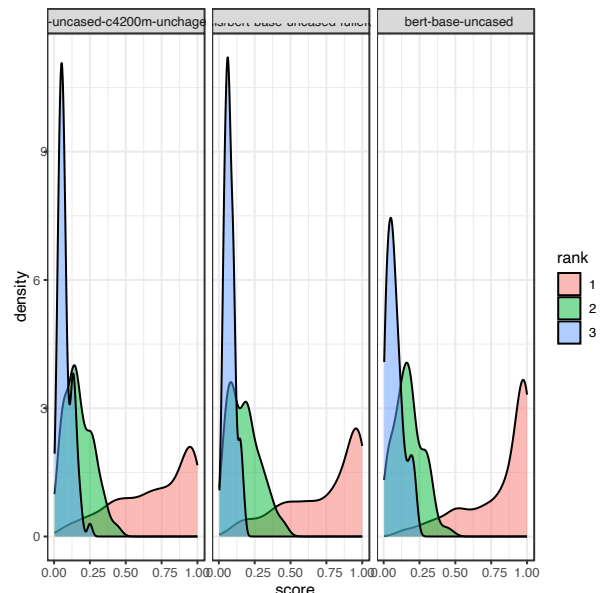


Figure 3: Probability Density Distribution per Models (Synthetic Learner Model, natural learner Model and Native model from left to right) for the top rank predictions from 1 to 3.



tive model, on average, assigns higher probability scores to its top-1 prediction, yet, has a lower success rate compared to the natural learner model.

Figure 4 displays the probability density distribution of words across different Universal Part-of-Speech (UPOS) for the first prediction (rank = 1). The x-axis represents the probability assigned by the natural learner model for each UPOS, while the y-axis shows the probability density. This visualization allows for a quick comparison of the relative frequencies of different UPOS across the dataset. It indicates how the model makes use of tokens of a certain type across levels.

Open-class categories such as adjectives (ADJ), nouns (NOUN), and verbs (VERB) have bimodal distributions, but the prominent mode reflects the uncertainty of the prediction (probability around 0.2 for ADJ). However, a closed class like prepositions (ADP) also has a bimodal distribution, but the prominent mode is around 0.9. This suggests that the model is more confident with some closed classes than open classes.

## 5.2 Domain Effects for ESP

We conducted a chi-squared test, which demonstrated that the difference between the domains was significant ( $X^2 = 45.04, df = 6, p < 0.001$ ). Our data indicated that masked tokens were easier to predict in essays written for Communication Studies compared to those for Pharmacy, as illustrated in Table 4. This is some indication to further take into consideration domain and tasks effects.

## 5.3 Training Limitations

A significant limitation in our training process is the imbalance in the distribution of proficiency levels within the EFCAMDAT dataset. Specifically, there is a disproportionately higher number of beginner-level texts (A1, A2) compared to advanced-level texts (C1, C2). This imbalance may affect our KL plot 1. While the result aligns with expectations for lower proficiency levels, it may exhibit a training artifact effect where the model’s contextual representation seems to be coherent towards the characteristics of beginner-level texts since it was exposed to a large amount of such texts, whereas for higher proficiency levels, the model’s token choices simply follow the native BERT distribution.

This artifact impacts the model’s ability to generalize across proficiency levels. For higher proficiency levels, the model’s token choices tend to

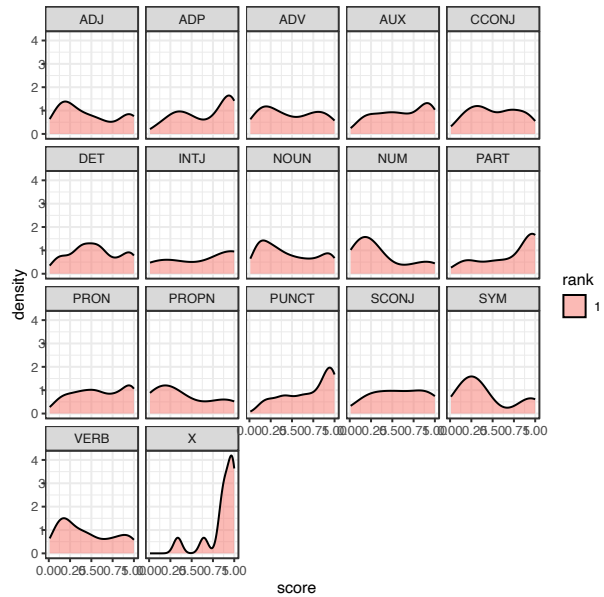


Figure 4: Probability Density Distribution of top-1 prediction of natural learner model per UPOS

align more closely with the original pre-training distribution, primarily because the advanced-level data is underrepresented. This limitation suggests that the model might not be equally effective across all proficiency levels, potentially underperforming for more advanced learners.

## 5.4 Perspectives for Future ITS Implementations

If our artificial learners manage to be sufficiently trustworthy for the emulation of what a learner would say, one can compare the prediction or the use of a given learner with each model pre-trained with a given CEFR level. Our experiment is only a prototype of our global undertaking. We will extend the pre-training to other areas, such as pre-training on different sub-levels of the CEFR scale. We have seen the reliability of the results, and we have also suggested that the models created were not too data-dependent in the sense that they could be generalized to other types of data.

## 6 Conclusion

In this paper, we have compared two artificial learners against a native language model in predicting tokens produced by learners. Our primary goal was to propose a masked language modelling task in learner corpora and analyse the accuracy, consistency, and divergence of such artificial learners. We explicitly chose a large synthetic ungrammatical dataset and an authen-

model	success_rate	score_mean	rank
bert-c4200m	0.53	0.60	1
bert-c4200m	0.08	0.10	2
bert-c4200m	0.04	0.04	3
bert-fullefcamdat	0.58	0.64	1
bert-fullefcamdat	0.09	0.09	2
bert-fullefcamdat	0.02	0.04	3
bert-base-uncased	0.55	0.71	1
bert-base-uncased	0.08	0.09	2
bert-base-uncased	0.03	0.04	3

Table 3: Model success rate and average probability score per rank (top-k position) for prepositions

	Communication	Electronics	Medicine	Pharmacy	Education	Environment	Physics
Success	1278	219	249	85	265	1139	749
Total	4284	933	1002	401	1157	4873	3301

Table 4: Contingency table of correct predictions per ESP domain (all models)

tic learner corpus to analyse the trade-off between the quality of authentic texts and the quantity of augmented texts. Even compared to the native BERT model, pre-training BERT in the synthetic C4200m dataset decreased accuracy, while training BERT on authentic texts increased accuracy. Accuracy is greater for closed classes, and the previous study on artificial learners rightly focused on a subset of a closed class, prepositions. Through analysing predicted probabilities against success rates, we investigated indications of calibrations and overconfident mistakes of our models, where native BERT showed a wider gap between its success rate and predicted probability. We finally compared native BERT with our natural artificial learner in relation to their choice of tokens, where the KL metric exhibit to be a coherent metric to generally measure the choice of tokens between language models. Since we pre-trained our artificial learner on a dataset containing more texts from beginner learners than those from advanced learners, we expect that it will simulate better beginner learners. Future work could address multiple aspects of the training process to enhance performance. We believe that merely increasing computational power and training time could still improve our artificial learners. Additionally, we believe that more specific masking strategies, such as masking incorrect tokens, and architectures that can personalize the artificial learner to a specific individual, could further enhance performance. In the direction of personalization, there are opportunities for training more specific artificial learners,

such as nationality or proficiency based artificial learners.

## Limitations

There are several limitations to our work that need to be acknowledged. One significant limitation is the high training cost associated with using deep learning models for natural language processing tasks. Training these models requires substantial computational resources, which can be expensive and time-consuming. In our study, although we aimed to mitigate these costs by using "small" encoder models such as BERT, the training costs were still considerably higher compared to traditional language modelling methods.

Furthermore, we expect to make our model available in accordance with the EFCAMDAT corpus curators, which provides a significant advantage in terms of cost-effectiveness and collaborative potential. Researchers and practitioners can leverage our pre-trained models and fine-tune them for their specific applications without incurring the high costs associated with training a model from scratch. This open-source approach promotes transparency and encourages further innovation and experimentation within the community.

## Ethics Statement

In accordance with the curators of the EFCAMDAT corpus, we have planned to make our models pre-trained on the EFCAMDAT accessible on the

web server hosting the EFCAMDAT data.

## Acknowledgments

This work has been supported by Science Foundation Ireland under Grant Number SFI12RC2289\_P2 Insight\_2, Insight SFI Centre for Data Analytics and by the French ANR under ANR grant ANR-22-CE38-0015 for the A4LL project.

## References

- Drilon Avdiu, Vanessa Bui, and Klára Ptačinová Klimčíková. 2019. [Predicting learner knowledge of individual words using machine learning](#). In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 1–9, Turku, Finland. LiU Electronic Press.
- Joan Bresnan and Tatiana Nikitina. 2009. On the Gradience of the Dative Alternation. In Karuvannur Puthanveetil Mohanan, Linda Uyechi, and Lian-Hee Wee, editors, *Reality Exploration and Discovery: Pattern Interaction in Language & Life*, pages 161–184. Center for the Study of Language and Information, Stanford, CA.
- Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.
- Christopher Bryant and Ted Briscoe. 2018. [Language model based grammatical error correction without annotated training data](#). In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 247–253.
- Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. [The BEA-2019 shared task on grammatical error correction](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.
- Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. [Automatic annotation and evaluation of error types for grammatical error correction](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.
- Lihu Chen, Alexandre Perez-Lebel, Fabian M Suchanek, and Gaël Varoquaux. 2024. Reconfiguring LLMs from the Grouping Loss Perspective. *arXiv preprint arXiv:2402.04957*.
- Council of Europe. 2018. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Companion Volume with New Descriptors*. Council of Europe.
- Morris H DeGroot and Stephen E Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. *arXiv preprint arXiv:2105.03075*.
- Stefan Th Gries, Justus Liebig, and Sandra C Deshors. 2020. There’s more to alternations than the main diagonal of a  $2 \times 2$  confusion matrix: Improvements of MuPDAR and other classificatory alternation studies. *ICAME Journal*, 44(1):69–96.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020a. [Don’t stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020b. [Don’t stop pretraining: Adapt language models to domains and tasks](#). *arXiv preprint arXiv:2004.10964*.
- Shin’ichiro Ishikawa. 2013. The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the world*, 1(1):91–118.
- Wonbin Kim. 2024. [Let’s make an artificial learner to analyze learners’ language!](#) (*Language Sciences*), (70):167–193.
- Rebecca Knowles, Adithya Renduchintala, Philipp Koehn, and Jason Eisner. 2016. [Analyzing learner understanding of novel L2 vocabulary](#). In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 126–135.
- Benjamin A Levinstein and Daniel A Herrmann. 2024. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, pages 1–27.

- Cyriel Mallart, Andrew Simpkin, Rémi Venant, Nicolas Ballier, Bernardo Stearns, Jen Yu Li, and Thomas Gaillat. 2023. [A new learner language data set for the study of English for Specific Purposes at university level](#). In *Proceedings of the 4th Conference on Language, Data and Knowledge - LDK 2023*, volume 1, pages 281–287, Vienna, Austria.
- Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694.
- Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.
- Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. [GECToR – grammatical error correction: Tag, not rewrite](#). In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA. Association for Computational Linguistics.
- Álvaro J Jiménez Palenzuela, Flavius Frasinca, and Maria Mihaela Truşcă. 2022. [Modeling second language acquisition with pre-trained neural language models](#). *Expert Systems with Applications*, 207:117871.
- Adithya Renduchintala, Rebecca Knowles, Philipp Koehn, and Jason Eisner. 2016. [User modeling in language learning with macaronic texts](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1859–1869, Berlin, Germany. Association for Computational Linguistics.
- Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. [Second language acquisition modeling](#). In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana. Association for Computational Linguistics.
- Burr Settles and Brendan Meeder. 2016. [A trainable spaced repetition model for language learning](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1848–1858, Berlin, Germany. Association for Computational Linguistics.
- Itamar Shatz. 2020. Refining and modifying the efcamdat: Lessons from creating a new corpus from an existing large-scale English learner language database. *International Journal of Learner Corpus Research*, 6(2):220–236.
- Felix Stahlberg and Shankar Kumar. 2021. [Synthetic data generation for grammatical error correction with tagged corruption models](#). In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.
- Milan Straka. 2018. [UDPipe 2.0 prototype at CoNLL 2018 UD shared task](#). In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.
- Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis, editors. 2018. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.
- Jacob Whitehill and Javier Movellan. 2017. [Approximately optimal teaching of approximately optimal learners](#). *IEEE Transactions on Learning Technologies*, 11(2):152–164.
- Brian Zylich and Andrew Lan. 2021. [Linguistic skill modeling for second language acquisition](#). In *LAK21: 11th International Learning Analytics and Knowledge Conference, LAK21*, page 141–150, New York, NY, USA. Association for Computing Machinery.

# Semantic Error Prediction: Estimating Word Production Complexity\*

**David Strohmaier**

ALTA Institute,  
University of Cambridge  
david.strohmaier@cl.cam.ac.uk

**Paula Buttery**

ALTA Institute,  
University of Cambridge  
paula.buttery@cl.cam.ac.uk

## Abstract

Estimating word complexity is a well-established task in computer-assisted language learning. So far, however, complexity estimation has been largely limited to comprehension. This neglects words that are easy to comprehend, but hard to produce. We introduce semantic error prediction (SEP) as a novel task that assesses the *production* complexity of content words. Given the corrected version of a learner-produced text, a system has to predict which content words are replacements for word choice errors in the original text. We present and analyse one example of such a semantic error prediction dataset, which we generate from an error correction dataset. As neural baselines, we use BERT, RoBERTa, and LLAMA2 embeddings for SEP. We show that our models can already improve downstream applications, such as predicting essay vocabulary scores.

## 1 Introduction

Automatically estimating complexity of a word is a core task for computer-assisted language learning (CALL). This literature uses “complexity” to refer to the difficulty of processing a word (cf. North et al., 2023). But words can be difficult to process in multiple ways, leading to varieties of complexity. So far, the focus in NLP has been largely on complexity in comprehension. We fill a gap left by this focus and investigate the overlap of two varieties of complexity:

1. Lexical Semantic Complexity: The difficulty of a word due to its meaning.
2. Production Complexity: The difficulty of producing a word.

---

\*This paper reports on research supported by Cambridge University Press & Assessment. We thank Chris Bryant for comments, advice, and provision of code, and all anonymous reviewers for their comments.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

In the next section, we will discuss these types of complexity and their overlap in more detail, establishing their nature and relevance for CALL. To investigate the overlap, i.e. *lexical semantic production complexity*, we propose the task of semantic error prediction (SEP) and create an SEP dataset from an error correction dataset. Our method can be applied to other error correction datasets.

After describing the creation method for our dataset, we perform a Bayesian logistic regression analysis of candidate features for predicting semantic errors. We then provide SEP baseline results using BERT, RoBERTa, and LLAMA2 embeddings and compare them with the performance of the feature-based regressions. Finally, we use scores from the LLAMA2-based model for predicting the vocabulary scores of L2 learner essays with a Bayesian linear regression.

Our contributions are as follows:

1. We propose a new CALL task, semantic error prediction, which offers access to lexical semantic production complexity.
2. We present a method for creating SEP datasets from error correction datasets as well as a dataset created that way.
3. We provide results from transformer-based models for the SEP task.
4. We showcase the use of SEP models for the downstream application of predicting essay vocabulary scores.

The scripts required for creating the dataset are available online at [https://github.com/dstrohmaier/semantic\\_error\\_prediction](https://github.com/dstrohmaier/semantic_error_prediction).

## 2 Types of Complexity

Word complexity, understood here as the difficulty of a word in processing, has many varieties. We develop the two overlapping types of complexity investigated in this paper and why they are important for CALL.

## 2.1 Lexical Semantic Complexity

Lexical semantic complexity is the complexity of a word due to its meaning. It can be distinguished from e.g. the syntactic complexity of a sentence or the orthographic complexity of a word form. Lexical semantic complexity has long been recognised as one of the main forms of lexical complexity, although its exact nature has been heavily debated (Cutler, 1983).

The notion of lexical semantic complexity can be compared with that of lexical sophistication, which is often understood as the use of low frequency vocabulary items (Laufer and Nation, 1995), although more detailed analyses have been put forward (Kim et al., 2018). We will find that, in some contexts, frequent words are difficult to produce, suggesting a difference between lexical sophistication and *contextual* lexical semantic complexity.

Similarly, lexical semantic complexity can be distinguished from other aspects of lexical complexity, such as morphological complexity. “abjure” is morphologically simpler but arguably semantically more complex than “theatergoer”. In fact, we find in section 5 that character length appears negatively related to contextual semantic complexity.

Lexical semantic complexity poses deep challenges for CALL applications, as semantic nuances can be subtle and, thus, identifying semantically challenging words can be difficult. At the same time, semantic correctness is especially important for communication, more so than e.g. word order and subject-verb-agreement. We can understand other speakers even when their sentences violate multiple grammatical rules, but when they produce multiple semantically incorrect words, communication tends to break down.<sup>1</sup>

One reason that lexical semantic complexity has been difficult for CALL applications is that few ways exist to estimate it on the word-level. Compared to morphological and syntactic complexity, for which syllable count and depth of the syntactic graph serve as easily accessible features, features to predict the semantic complexity of a word token are harder to engineer. Many measures for assessing lexical complexity, such as the type-token ratio, op-

<sup>1</sup>See Olsson (1972) and Khalil (1985) for support of the thesis that semantic errors impede communication more than grammatical errors. The research by Nushi et al. (2022) suggests that formal errors can reduce intelligibility more than lexical semantic errors, but, in their discussion, formal errors include e.g. the choice of the wrong suffix, which could arguably be treated as a semantic issue.

erate on the document rather than the word-token-level (consider the features in table 1 of Bulté and Housen, 2012, p. 31).

There exist word-type-level features commonly associated with lexical complexity, such as:

- word frequency,
- age-of-acquisition (for first language speakers), and
- concreteness of the word.

As word-type-level features, they suffer from three shortcomings:

1. They ignore the contextual aspect of lexical complexity.
2. They typically fail to account for homonymy and polysemy, i.e. most data for them are only available on the word form level.
3. They cannot cover the entire vocabulary, as it rapidly evolves, e.g. how does the complexity of “rizz” compare to that of “mid”?

Hence, there is a need for another way to measure lexical semantic complexity in context, which we will meet.

## 2.2 Production Complexity

Production complexity, which we distinguish from comprehension complexity, is the difficulty of producing a linguistic unit either in speech or writing. For the purposes of the present investigation, production will be limited to writing.<sup>2</sup>

The distinction between comprehension and production complexity is related to the distinction between passive and active vocabulary, i.e. the recognition-recall difference, because production typically requires recall. Research into second language learning has investigated the difference, finding that even advanced learners show a large gap between passive and active vocabulary (Laufer, 1998; Fan, 2000).

Production complexity can diverge from comprehension complexity, because a semantic difference might be important for word choice without being important for word recognition. One example for this is the mass-count distinction. A language user might very well understand a sentence such as “He drank much milk.” and yet erroneously produce sentences such as “He drank many milk.”. That is, the mass-count distinction might play a bigger role in production than comprehension complexity.

<sup>2</sup>For a survey of psycholinguistic research into task complexity and its interactions with other forms of complexity for L2 writing, see Johnson (2017).

For an example closely linked to word form, consider the case of “price”/“prize”. In English, these two words differ in form and meaning. In German, however, the neargraph “Preis” is ambiguous between the two meanings. This might lead an L2 learner of English with German L1 to be able to comprehend the English words, while erroneously producing “prize” instead of “price”.

Multiple data sources exist for assessing word complexity in general, with a tendency towards comprehension (Shardlow, 2013; Shardlow et al., 2020), while production is under-resourced.<sup>3</sup> One reason for this neglect is that much work on word complexity was intended to improve readability (see North et al. 2023, and, for an example, see the introduction of Gooding et al. 2021). Complex word identification was, thus, conceived of as a step in a pipeline for adapting text to a specific set of learners for *comprehension* (cf. North et al., 2023).

However, systems able to predict which words learners struggle to *produce* are also of use for adaptive teaching systems. Three such use cases are:

1. Content calibration: When learners are prompted to produce a particular word, the complexity of the word should be at the intended level for the task. For example, cloze tasks require learners to produce words that can fill a gap in a text. Knowing the production complexity of the target word would be of value for calibrating the item.
2. Assessment: Production complexity scores can serve to assess learner produced text, even though the relationship is not simple, as we will see in section 7.
3. Highlighting during learning: Words in a text read by a learner might be flagged to make the learner aware that they are harder to produce.

A further reason for the lack of resources on production complexity is that such datasets are harder to create. Eliciting complexity judgements from annotators reading a text is relatively simple. There does not appear to exist a simple equivalent for production, as it is challenging to ask annotators to rate the complexity of words while producing them at the same time.

To address this problem, we are using an error correction dataset based on learner-written texts for creating our SEP dataset.<sup>4</sup> Our method can be

<sup>3</sup>One resource specifically for production is the SweLLex word list (Volodina et al., 2016) for Swedish as an L2.

<sup>4</sup>Other options for estimating production complexity

applied to any error correction dataset providing appropriate error annotations and corrections.

### 2.3 The Overlap: Lexical Semantic Production Complexity

We are interested in cases where a word is difficult to produce due to its semantics in a specific sentential context. This overlap gives rise to its own dynamics, because, in production, the conceptual information is typically activated prior to the word form information, rather than the reverse, as in the case of comprehension (see Jiang 2000 for an example of this). As a result of this reversal, we expect different complexity patterns in production than in comprehension.

Specifically, the patterns might show a different type of contextual effect: Language learners might inadvertently create contexts that require a certain word choice and as a result the learners might select the wrong word. Thus, a word that might be easy to comprehend and frequently selected in one context might be difficult to produce in another context, even though both contexts are created by the language user.

That lexical semantic production complexity is impacted by contextual effects is backed up by the empirical literature on English second language acquisition, which documents a sizeable number of semantic errors resulting from collocational phenomena (Al-Shormani and Al-Sohbani, 2012; Jęptarus and Ngene, 2016). Our approach and dataset provide a way for CALL applications to account for such phenomena specific to lexical semantic production complexity.

## 3 Related Work in NLP

Our work builds upon the NLP literature for both word complexity and error detection.

### 3.1 Word Complexity

The complex word identification (CWI) task, which has been investigated in multiple shared tasks (Paetzold and Specia, 2016; Yimam et al., 2017; Shardlow et al., 2021), aims to identify complex words in context. Recently, it has been extended under the name “lexical complexity prediction” (LCP) to a continuous task of predicting the complexity of a

would include key-stroke or eye-tracking data. We thank an anonymous reviewer for these suggestions. These behavioural trace data, however, render it difficult to differentiate the semantic component of production complexity from other aspects.

word (Shardlow et al., 2021, 2020). For an in-depth review of the CWI/LCP literature, see North et al. (2023).

While feature-based machine learning systems were state-of-the-art for many years (Gooding and Kochmar, 2018), by now end-to-end neural systems dominate the area (Shardlow et al., 2021). These models often use BERT-style transformers as their basis (Devlin et al., 2019; Liu et al., 2019). In the CWI literature, it has also been shown that backgrounds of language learners, e.g. their overall proficiency level, matter for whether a word is complex or not (Gooding et al., 2021).

As mentioned above, datasets in the CWI/LCP literature are generally more appropriate for capturing comprehension rather than production complexity. This tendency is due to the annotation process: annotators are presented with text for which they assign complexity labels. Effectively, the annotators engage in comprehension when deciding on a label.

Furthermore, complexity annotation is an artificial way of engaging with text, which raises the question of external validity. Even when the annotations are provided by L2 learners, these learners are not trying to communicate with another human language user in a natural manner. By predicting errors in text production, our approach is closer to natural engagement with text and, therefore, addresses this issue.

There also exists a literature on predicting the CEFR levels of words (Alfter and Volodina, 2018; Pintard and François, 2020), which is less comprehension focused. This literature tends to consider words or word senses in isolation, rather than in the context of use (but see Aleksandrova and Pouliot, 2023).

### 3.2 Error Detection

Semantic errors are covered by the error detection literature, but much of this literature is focused on morpho-syntactic errors. Similar to complexity, error detection and closely related problems have been the subject of multiple shared tasks (Ng et al., 2014; Bryant et al., 2019; Volodina et al., 2023). Similar to CWI/LCP, this field is dominated by transformer-based models, often combined to increase performance (Qorib et al., 2022; Qorib and Ng, 2023). For a recent survey of error correction, see Bryant et al. (2023).

While closely related to SEP, error detection and

correction systems are not designed for the purpose of assessing the lexical complexity of content words, but rather their correctness. Correctness, however, can be due to the learner avoiding more difficult terms and resorting to simpler expressions. By contrast, our approach is able to distinguish two correct words with regard to which one was more complex to produce.

## 4 Dataset

We present a SEP dataset that can be constructed from existing resources.<sup>5</sup> Our dataset uses error correction as the starting point for determining production complexity. Using learner texts as the source of the dataset ensures high external validity: The learners are engaged in a naturalistic task and patterns of their output are used to assess the lexical semantic complexity.

In constructing our dataset, we only predict the corrections of word choice errors. That is, we focus on the word tokens learners *should* have produced, but failed to do so. This production failure is taken as a direct indicator of production complexity.

Our approach only considers the corrected token, not the erroneously produced words. That is, when an annotator tags replaces “work” by “job” in a sentence, this is taken as evidence that the “job”-token in this sentence is complex, without any further inference regarding “work”.

The reason for this choice is that we are interested in the complexity of word tokens in a specific context. It is unclear what we learn from an erroneously produced token. When a token is produced, it was evidently feasible to wrongly produce “work”, even if it was semantically impossible to produce this word token correctly. Due to these conceptual problems, our dataset construction will focus on the context-appropriate words that learners fail to produce.

Our dataset concerns both the breadth and depth of lexical knowledge (Bulté and Housen, 2012): Errors occur both when learners lack items in the vocabulary, an issue of breadth, and when learners lack the lexical knowledge to correctly integrate words into sentences, an issue of depth. Thus, our research cuts across the theoretical constructs of lexical complexity presented by Bulté and Housen (2012, figure 3).

<sup>5</sup>The scripts required for doing so are made available at: [https://github.com/dstrohmaier/semantic\\_error\\_pr\\_ediction](https://github.com/dstrohmaier/semantic_error_pr_ediction).



## 4.1 Dataset Construction

Our starting point is the dataset published as part of the 2019 BEA shared task on grammatical error correction (Bryant et al., 2019). This dataset provides error annotations for sequences, taken primarily from texts written by second language learners of English, although the evaluation data also includes some native speakers. The annotations follow the scheme of the ERRANT tool (Bryant et al., 2017). In addition, the dataset provides CEFR levels for the texts (CoE, 2020).

Since we are interested in semantic word choice and such word choice can be evaluated only in a semantic context, we use whole paragraphs extracted from the dataset as input.<sup>6</sup> We then invert the dataset so as to move from error detection to error prediction.<sup>7</sup>

Error Code	Meaning	Example
<b>R:VERB</b>	Verb replacement	<i>order</i> → <i>book</i>
<b>R:NOUN</b>	Noun r.	<i>base</i> → <i>foundation</i>
<b>R:ADJ</b>	Adjective r.	<i>low</i> → <i>poor</i>
<b>R:ADV</b>	Adverb r.	<i>graciously</i> → <i>gracefully</i>

Table 1: Selected error types (cf. Bryant et al., 2017).

In the next step, we select the relevant error types: word replacement errors in which content words, i.e. nouns, verbs, adjectives, and adverbs have been replaced by the annotators.<sup>8</sup> Orthographic, morphological, tense, subject-verb agreement and similar are thus excluded from the prediction task to focus on semantic complexity. They are also corrected, however.

In addition, we render the labels binary: Each token in the dataset is annotated for whether it has been corrected using one of the selected error tags.<sup>9</sup>

<sup>6</sup>Extremely short paragraphs, for example best wishes at the end of a letter, are merged into larger paragraphs when possible. In a small number of cases, the sub-tokenized paragraphs are longer than the maximal sequence length (512). 7 texts are affected, only one of which belongs to the split used for evaluation.

<sup>7</sup>We thank Chris Bryant, one of the original organizers of the 2019 BEA shared task, for providing code.

<sup>8</sup>When the replacement crosses part of speech, e.g. a verb is replaced by a noun or an adverb by an adjective, Errant typically treats this as an R:OTHER error, which is not used by us. We assume that when such errors occur, usually more has gone wrong than just the choice of a wrong word due to its semantics.

<sup>9</sup>We only label word tokens with the spaCy POS-tags: VERB, NOUN, ADJ, ADV. As a result, we exclude a small number of positive labels for other POS. The largest block of these positive labels are auxiliary verbs. 344 out of 49118 are labelled positively, most of which are in their turn forms of “be”, “have”, and “do”. We use the spaCy en-core-web-sm

For evaluation of these binary tags, we use the  $F_1$ -score and the area under the curve (AUC) of the Receiver Operating Characteristic Curve (ROC).

The original dataset provides a public train- and a dev-split.<sup>10</sup> We use the dev-split as an eval-split and split the train-split into a new train- and new dev-split. We apply our method to these public splits, with the new dev-split being primarily used for development purposes prior to evaluation (e.g. checking code correctness).

	# sequences	# tokens	# content t.	% errors
train	10523	577892	239156	2.45
dev	1170	63420	26409	2.43
eval	1419	88580	36923	2.04

Table 2: Descriptive dataset metrics. “content t.” stands for tokens with content word POS tags. Percentages indicate the percentage of content word tokens corresponding to replacement errors.

## 4.2 Descriptive Metrics

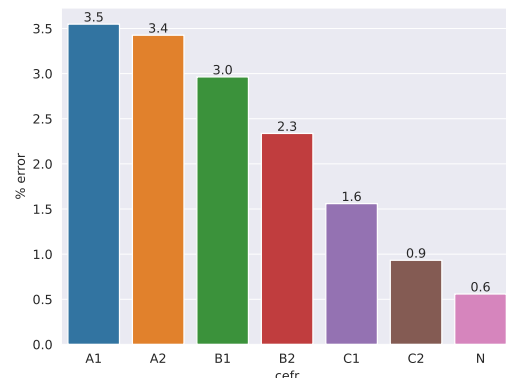


Figure 1: Replacement error percentages for content word tokens across CEFR levels (N=ative speakers).

The training-split contains more than half a million tokens, slightly less than half of which are content word tokens (see table 2). The dev- and eval-split are  $> 10\%$  of that size.

Across splits, around 2.4% of content word tokens correspond to semantic errors,<sup>11</sup>. These overall numbers, however, mask considerable differences in the error percentages across CEFR levels: The lower the CEFR level, the more content word

model for POS-tagging (Honnibal et al., 2020).

<sup>10</sup>The test split is not public and therefore not used by us.

<sup>11</sup>The number for the eval split in table 2 is lower, because the dev split also includes native speakers.

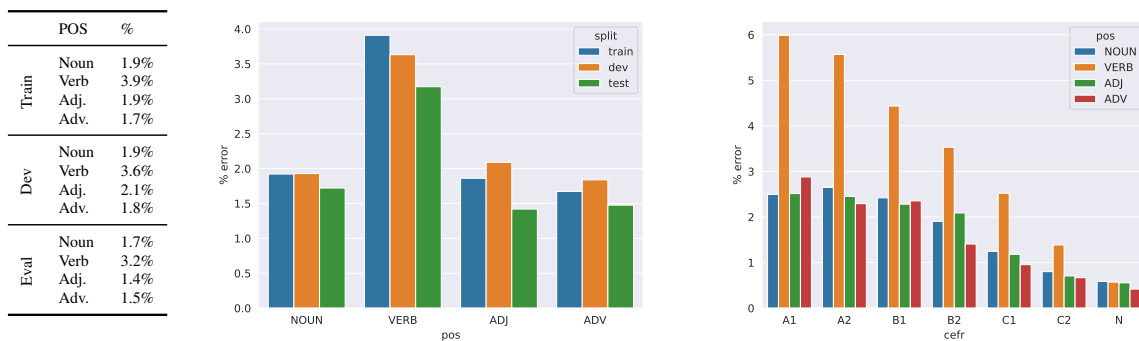


Figure 2: Percentages of errors for different POS across splits and CEFR levels (N=Native).

replacement errors are committed by a learner (see figure 1).

Across POS-tags, verbs are particularly likely to have been corrected, around 3.5% of the time (see figure 2). Verbs are a greater source of errors for second language learners of English, even at the C2 level, but the effect disappears for native learners of English, which are included in the eval split of the dataset. We speculate that this might be due to the higher context dependence of verbs, at least when compared to nouns (Gentner and France, 1988; Kersten and Earles, 2004; Earles and Kersten, 2017).<sup>12</sup> The context-dependence might take a language specific form, leading to L1-interference for L2 learners.

### 4.3 Qualitative Inspection

In this section, we consider hand-picked examples of replacement errors from our new train split.

The qualitative inspection suggests that learners often replace words with neargraphs, e.g. using “aspects” instead of “respects” or “affection” instead of “infections”.<sup>13</sup> That being said, mistaken words and their corrections are also semantically related, with learners using “blame” instead of “guilt” and “contaminated” instead of “polluted”.

Some mistaken tokens exhibit a lack of specificity. For example, in the corrected sentence “And some buses drive at night to transport [take → transport] passengers.” “transport” replaces “take”.

<sup>12</sup>The claim that verbs are more context-dependent is related to the idea that a verb predicates something of something else, thus being constrained both by what it predicates and the subject of its predication. The idea that verbs play this connecting role might be tracked back at least to Aristotle, who in *De Interpretatione* (3.16b6–7) asserts that “it [a verb] is a sign of things said of something else” (Aristotele, 1994, p. 44).

<sup>13</sup>Errant provides a separate tag for orthographic error (R:ORTH), which we do not use.

While the meaning of the sentence can be understood without this correction, *transport* is more specific than *take*. It would be too simplistic, however, to think that learners always use less specific words.

Adjectives provide evidence that the words learners fail to produce are not necessarily highly specific or generally lacking from their vocabulary: “good” is one of the adjectives most often inserted by annotators. It typically replaces more specific adjectives such as “suitable” and “healthy” that fail to be contextually appropriate. This observation underlines the difference between lexical semantic complexity in production and comprehension: a learner producing “suitable” instead of “good” is likely able to comprehend the word “good”.

A lack of idiomaticity can also lead to corrections. For example, annotators changed “big enterprises” to “big businesses”. Similarly, annotators replace “main friend” with “best friend”. In these cases, a reader will be able to comprehend the sentences with either word choice, but the corrected formulation is more idiomatic.

Errors due to a lack of idiomaticity are one reason why semantic error prediction is a highly challenging. Consider the following sentence:

“I began doing this sport three years ago when I lost my job [work → job].”

In this sentence, the annotators replaced “work” with “job”, but this is a very nuanced correction, that arguably involves collocational preference as well as semantic detail.

## 5 Bayesian Regression Analysis

To analyse the dataset, we perform a Bayesian logistic regression using Bambi (Capretto et al., 2022), a package for Bayesian regression models based on PyMC (Oriol et al., 2023). We fit the regression on

the combined training and evaluation splits of the data using only tokens that were tagged as content words using spaCy (Honnibal et al., 2020). Since we are also primarily interested in interpreting the features, we drop rows that lack a feature required for any of the regression models.

We estimate two models. The first is a base model which has as input features (see next section for details):

- length of the word in characters,
- word frequency,
- age of acquisition, and
- whether the token is a verb.

The second models adds an interaction effect between being a verb and the frequency. The equations are described in appendix figure 6.

## 5.1 Observed Variables for Regression

Except for being a verb, all the explanatory variables were selected based on their general usage in the complex word identification literature (e.g. Gooding and Kochmar, 2018). However, in SEP the features are for the *corrected* learner text.

**Length in characters.** Provides the length of the token as counted in characters.

**Frequency.** We use the wordfreq package for Python,<sup>14</sup> specifically the Zipf frequency estimate. The package uses 0 as the default value of words not found in the word list.<sup>15</sup>

**Age of acquisition (AoA).** While the age of acquisition is a metric for L1 acquisition, it can also be applied to L2 acquisition under the simplifying assumption that both acquisition processes proceed from simpler to more complex words. While this assumption is probably not correct in all cases due to vocabulary transfer from L1 to L2, it offers a sufficiently close approximation of learning order (as our results show; see also the correlation of learning order documented by Flor et al. 2024). The AoA values are taken from the dataset presented in Kuperman et al. (2012).<sup>16</sup> The coverage by this dataset is incomplete and tokens for which no AoA is available are dropped from the dataset. Other tokens from the same sentence are still used for

<sup>14</sup><https://github.com/rspeer/wordfreq>. The package uses the ExquisiteCorpus (<https://github.com/LuminosoInsight/exquisite-corpus>).

<sup>15</sup>0 does not correspond to zero occurrences due to the zipfian transformation.

<sup>16</sup>Downloaded from <https://osf.io/kz2px/>.

training and evaluation. We scale the age of acquisition to a mean of 0 and variance of 1 to make it comparable to the CEFR-j.

**CEFR-j.** The CEFR-j project provides the CEFR level of word types based on the word lists provided by Open Language Profiles and Octanove.<sup>17</sup> We convert and scale the CEFR-j data to make it comparable with the AoA.

**Is verb.** The spaCy tags were used for this feature. It was motivated by our previous analysis, suggesting that verbs are much more likely to be semantic errors (see section 4.2).

**CEFR.** The underlying dataset provides the CEFR level for the submissions. We treat this as a categorical variable.<sup>18</sup>

## 5.2 Results and Interpretation

A Bayesian logistic regression produces a probability distribution over the parameters of interest. For the estimated parameters, we report the highest density interval (HDI), i.e. the interval of minimum width containing the parameter with a certain probability. As is the standard for Bambi, we consider the 94% HDI credible interval (i.e. the interval spanning from 3% to 97%). HDIs are often treated analogously to frequentist confidence intervals, but have the straightforward interpretation that, given observed data,<sup>19</sup> the effect has a 94% probability of falling within the interval.

The results for the base model can be seen in table 3 and figure 3. In line with the expectations from the CWI literature and the previous analysis, we find that;

- more frequent content words are less likely to be semantic errors (HDI:  $[-0.21, -0.11]$ ),
- content words with a higher CEFR level are more likely to be semantic errors (HDI:  $[0.09, 0.16]$ ),

<sup>17</sup>The lists were downloaded from <https://github.com/openlanguageprofiles/olp-en-cefrj/>. The CEFR-J Wordlist Version 1.5 was compiled by Yukio Tono, Tokyo University of Foreign Studies (Negishi et al., 2013). We use the CEFR-j list over others, because it is on the level of word form + POS rather than word sense. For example, the online EVP lists CEFR levels A1 and C2 among others for different senses of the noun “head”, while CEFR-j only provides A1 for the noun. Furthermore, CEFR-j provides a permissive license and easy access.

<sup>18</sup>In contrast to CEFR-j, we do not convert and scale the CEFR-level to be able to compare it to the AoA. The reason for this difference is that we do not intend a comparison with AoA, because it is a student-level rather than token-level variable.

<sup>19</sup>More rigorously, given the model specification, the prior, and the observed data.

	mean	sd	hdi <sub>3%</sub>	hdi <sub>97%</sub>
is verb	0.80	0.03	0.75	0.85
cefr[C2]	-1.49	0.08	-1.6	-1.35
cefr[C1]	-0.91	0.06	-1.02	-0.8
cefr[B2]	-0.5	0.05	-0.59	-0.4
cefr[B1]	-0.25	0.05	-0.33	-0.15
cefr[A2]	-0.07	0.05	-0.15	0.03
scale(cefr-j)	0.12	0.02	0.09	0.16
scale(aoa)	0.09	0.02	0.05	0.13
frequency	-0.16	0.03	-0.21	-0.11
character length	-0.06	0.01	-0.07	-0.04
intercept	-2.39	0.16	-2.69	-2.08

Table 3: Estimated parameters of base model. Mean, standard deviation, and HDI boundaries of the estimated posterior are provided.

- content words with a higher AoA are more likely to be semantic errors (HDI: [0.02, 0.05]), and
- students with higher CEFR level are generally less likely to commit errors,
- verbs are more likely to be semantic errors (HDI: [0.75, 0.85]) compared to other content words.

Contrary to what one might expect from the CWI literature, longer words appear less likely to be replacements for semantic errors (HDI: [-0.07, -0.04]). That is, a word in the corrected sentence being longer is not an indicator of it corresponding to a semantic error. This could be a result of human annotators avoiding complex corrections for pedagogical reasons, or an effect of learners rarely intending to write long words. The finding is in line with “good” being one of the most frequent corrections for adjectives, replacing words like “suitable” and “healthy”.

The second model, which introduces an interaction between frequency and being a verb, complicates the picture considerably (see table 4 and figure 4). Being a verb stops being a strongly positive predictor for semantic errors (HDI: [-0.64, 0.06]), while the interaction between frequency and being a verb is positive (HDI: [0.15, 0.18]). This additional analysis suggests that the effect of verbs being more likely to be semantic errors is due to *frequent* verbs. This is line with our speculation that verbs enjoy a special status due to their contextual dependence: learners struggle with verbs because they are heavily constrained by context, not because they are rare.

One practical implication of this result is that a focus on the correct use of frequent verbs could be beneficial to support learners in production.

	mean	sd	hdi <sub>3%</sub>	hdi <sub>97%</sub>
frequency:is verb	0.22	0.04	0.15	0.28
is verb	-0.3	0.19	-0.64	0.06
cefr[C2]	-1.49	0.08	-1.63	-1.35
cefr[C1]	-0.91	0.06	-1.02	-0.81
cefr[B2]	-0.50	0.05	-0.60	-0.40
cefr[B1]	-0.25	0.05	-0.33	-0.15
cefr[A2]	-0.07	0.05	-0.16	0.02
scale(cefr-j)	0.12	0.02	0.09	0.16
scale(aoa)	0.10	0.02	0.06	0.13
frequency	-0.28	0.03	-0.34	-0.22
character length	-0.06	0.01	-0.07	-0.04
Intercept	-1.80	0.19	-2.17	-1.46

Table 4: HDI (3–97% interval) of model with interaction between being a verb and frequency (B+I). Table includes mean, standard deviation, and HDI boundaries of the posterior.

## 6 Deep Learning Models

We put forward baseline deep learning models trained for semantic error prediction using our dataset. The models are probes trained on embeddings from pre-trained transformer models (Vaswani et al., 2017).

### 6.1 Architecture

We use the English BERT and RoBERTa models as the basis of our architecture (Devlin et al., 2019; Liu et al., 2019).<sup>20</sup> In addition to these well-researched models, we also explore the more recent and larger LLAMA2-7B model (Touvron et al., 2023).

We use these models without fine-tuning to create embeddings of the word tokens in question. Due to subtokenisation, the base model might produce more than one embedding per word token, which we address with mean pooling of the subtoken embeddings. Research has suggested that the last layers of BERT-like transformer models are not best suited for lexical semantic tasks (Vulić et al., 2020). Therefore, we create our embeddings by mean pooling over layers 1–10 (inclusive), i.e. excluding the last two layers, for the BERT and RoBERTa models. For LLAMA2, the role of the different layers is not as well-established and we resorted to averaging the output of layers 1–30, i.e. ignoring the last two layers again.

Probes are fine-tuned on the word embeddings, which requires less computational resources than training the whole transformer. The probes consist of a hidden layer (size 100), an output layer, and a SoftMax pooling layer, described by the following

<sup>20</sup>We use the base-size models. All models are loaded using the huggingface transformers library (Wolf et al., 2020).

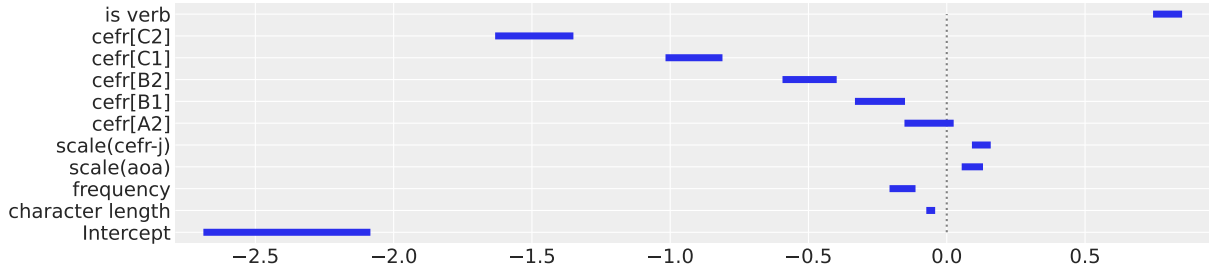


Figure 3: HDI credible intervals (3–97%) for coefficients of the base model (B).

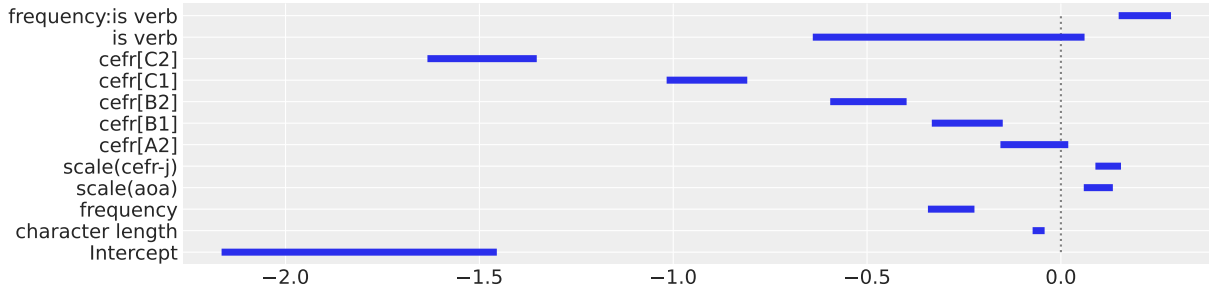


Figure 4: HDI (3–97% interval) for coefficients of the interaction model (B+I).

equations:

$$e = \text{embedding}$$

$$h = \text{ReLU}(W_{\text{hidden}}e + b_{\text{hidden}})$$

$$\text{scores} = \text{LogSoftMax}(W_{\text{output}}h + b_{\text{output}})$$

## 6.2 Training

The probes are trained by first performing hyperparameter-search using 5-fold cross-validation on the combined data from the train- and dev-split. The hyperparameter search randomly draws 20 hyperparameter settings from the space (see table 8 for details). The probe is then trained on the combined train- and dev-split using the hyperparameters reaching the highest  $F_1$ -score in cross-validation.<sup>21</sup> For training, we use the AdamW algorithm (Loshchilov and Hutter, 2018). The evaluation occurs on the eval-split.

The extreme label imbalance of the dataset can lead the probes to exhibit a bias towards assigning negative labels. To address this, we over-sample the positive labels during training, so that there is an equal number of positive and negative labels.

## 6.3 Results and Interpretation

Due to the imbalance of the labels, accuracy is not a meaningful metric for our dataset. Instead,

<sup>21</sup>The hyperparameter search space and the best hyperparameters for each probe are available in the online materials at [https://github.com/dstrohmaier/semantic\\_error\\_prediction/tree/main/probe\\_kwargs](https://github.com/dstrohmaier/semantic_error_prediction/tree/main/probe_kwargs).

we use the  $F_1$ -score and the area under the curve of the ROC (AUC). The AUC can be interpreted as the probability that a randomly chosen positive instance, i.e. a content word token that is a replacement, will have a higher score than a randomly sampled negative instance.

Table 5 provides the overall results, as well as the results for each CEFR level (including the “N” level for native speakers). We compare the deep learning models against a baseline that labels all tokens as corresponding to semantic errors (“all True”),<sup>22</sup> and the regression models discussed in section 5.<sup>23</sup>

With a threshold of 0.5, the logistic regression models fail to achieve an  $F_1$ -score of above 0%. The AUC score is more promising, consistently outperforming the 50%-threshold of the all True baseline. The transformer-embeddings based models outperform the regression baselines: with only one exception, there is at least one transformer model that outperforms the best regression model for every CEFR level. The exception is the AUC for the A1 level (B: 67.6%). In this case the information of the student CEFR level might be of sufficient importance to outweigh the performance

<sup>22</sup>Labelling all tokens negatively would lead to an  $F_1$  of 0.

<sup>23</sup>For the native test data, the CEFR label of the student is given as C2, since this is the closest available class. Otherwise the comparison to the regressions models favours the later, because they are not evaluated on missing data, e.g. when the age of acquisition of a word is not accessible.

	Overall		A1		A2		B1		B2		C1		C2		N	
	F <sub>1</sub>	AUC	F <sub>1</sub>	AUC	F <sub>1</sub>	AUC	F <sub>1</sub>	AUC	F <sub>1</sub>	AUC	F <sub>1</sub>	AUC	F <sub>1</sub>	AUC	F <sub>1</sub>	AUC
all True	4.0	50.0	6.4	50.0	6.8	50.0	6.0	50.0	4.7	50.0	3.6	50.0	2.3	50.0	1.1	50
B	0.0	68.7	0.0	<b>67.6</b>	0.0	60.1	0.0	59.8	0.0	62.6	0.0	55.6	0.0	69.4	0.0	54.7
B+I	0.0	69.0	0.0	66.7	0.0	60.4	0.0	60.3	0.0	62.4	0.0	58.4	0.0	68.7	0.0	54.0
BERT	9.6	67.8	10.4	64.3	13.1	<b>69.1</b>	15.2	66.8	9.4	66.5	7.3	<b>68.2</b>	8.5	73.2	2.7	<b>59</b>
RoBERTa	10.8	69.2	<b>12.4</b>	64.7	13.8	67.2	13.0	70.1	<b>14.1</b>	<b>72.8</b>	<b>10.5</b>	66.3	<b>10.8</b>	<b>78.1</b>	1.8	<b>59</b>
LLAMA2	<b>11.0</b>	<b>69.8</b>	11.9	64.6	<b>14.9</b>	68.1	<b>15.7</b>	<b>70.8</b>	12.4	68.8	6.5	67.4	3.0	72	<b>2.8</b>	58.2

Table 5: Scores in percentages. The baseline scores result from assigning True to all tokens or all content word tokens.

advantage of the transformer embeddings.

Looking across CEFR levels, no simple trend in performance holds. Both A1 (highest transformer AUC: 64.7%) and C1 (highest AUC: 68.2%) appear particularly challenging. One generalisation that can be made is that the numbers on the native data are the worst (F<sub>1</sub>: 2.8%, AUC: 59%). We assume that this is due to the absence of native essays in the training data. In effect, this result strongly suggests that the error patterns for native and L2 speakers differ considerably. After all, the C2 level, which is supposedly the closest to the native skill, has the highest performance! That being said, the native data are from a different source, the LOCNESS corpus (Granger, 1998), which might also explain the low performance.

That LLAMA2 has the highest overall F<sub>1</sub> (11%) and AUC (69.8%) suggests that the size of the language model is a factor. Generally, however, the differences are small and the highest AUC value is achieved by RoBERTa for the C2 level (78.1%).

In light of the dataset difficulty, it is not surprising that the F<sub>1</sub>-scores are low. The higher AUC are somewhat encouraging, especially for certain CEFR levels (e.g. C2 for RoBERTa reaching 78.1%). To support educational technologies, it will be important to better differentiate between complex and other words, i.e. to increase the AUC. That being said, the current scores can already be used as an input feature for downstream tasks, as we show in the next section.

## 7 Downstream Application

We show that the scores of one of our models support essay score prediction as a downstream task.

### 7.1 Setup

We use the ELLIPSE dataset (Crossley et al., 2023) for evaluation, which provides vocabulary scores for more than 6000 essays by L2 learners of English. We use the probability scores produced by

our LLAMA2-embeddings based model as it is the overall best performing model (see table 5) to predict these vocabulary scores using a Bayesian linear regression.

The vocabulary scores are on the essay-level, while our lexical complexity scores are on the token level, requiring us to perform pooling. We consider two forms of pooling: mean and max pooling.

In addition, we compare the regression using our model-derived lexical complexity scores with a simpler approach: For the simple approach, we use the proportion of times a word has been put forward as a correction. We use again mean and max pooling.

We also include other variables that can be used to assess vocabulary in our regression:

**Min. Frequency.** We use the same source of word frequencies as discussed in section 5.1. We apply min-pooling to the token frequencies, removing frequencies of 0.0 (default value).<sup>24</sup>

**CEFR-j.** We use the CEFR-j word list discussed in section 5.1, applying min-max-normalisation, so that each CEFR level corresponds to a 0.2 step, providing a range from 0–1 for comparison with our probe scores, which also range from 0 to 1. The CEFR-j scores for tokens are mean-pooled.

**Type-Token Ratio.** Following the literature on complexity (Bulté and Housen, 2012), we use the type-token ratio as a feature. The data is provided by the dataset, but we use the ratio rather than the percentage for comparability.

**Measure of Textual Lexical Diversity (MTLD).** The ELLIPSE dataset also provides MTLD data, a metric from lexical diversity derived from the type-token ratio (McCarthy, 2005), but accounting for text length. We rescale this data to a mean of 0 and standard deviation of 1.

<sup>24</sup>We also explored mean-pooling but found its coefficient to be indistinguishable from 0.

**Grade Level.** The ELLIPSE dataset includes students from grade 8 to 12. We use this information and min-max normalise the grade level to make it comparable with our probe scores.

We compare five regressions models:

1. **base:** Base model without any of our lexical semantic production complexity scores.
2. **max:** Model using the max-pooling of our lexical complexity scores in addition to base variables.
3. **mean:** Model using only the mean-pooling of our lexical complexity scores in addition to base variables.
4. **max+mean:** Model using both complexity scores.<sup>25</sup>
5. **proportion:** Model using the mean and max pooling error correction proportions instead, as described above.

## 7.2 Results and Discussion

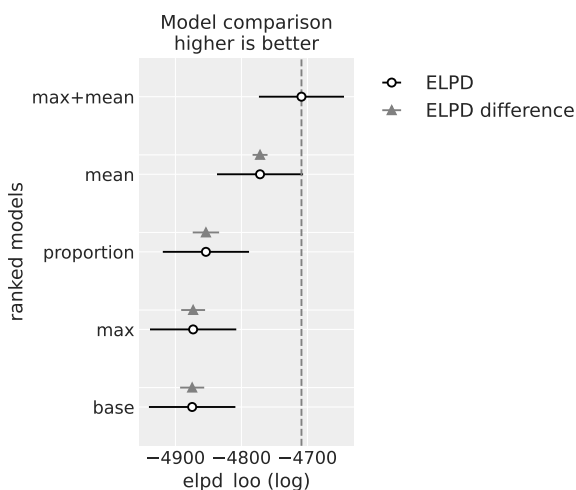


Figure 5:  $\text{elpd}_{100}$  scores for Bayesian linear regression models predicting vocabulary scores.

To compare our five models, we use the expected log pointwise predictive density, which is estimated using leave-one out cross-validation ( $\text{elpd}_{100}$ ). The  $\text{elpd}_{100}$  is a standard metric for comparing Bayesian models (see figure 5 and table 7) and can be written as (see Vehtari et al., 2017):

$$\text{elpd}_{100} = \sum_i^n \int p(y_i|\theta)p(\theta|y_{-i})d\theta$$

where  $y_{-i}$  are all datapoints except the  $i$ -th.

<sup>25</sup>It might appear more appropriate to use the median rather than the mean, as the latter also incorporates the max value. We found, however, that this made a negligible difference.

	mean	sd	hdi <sub>3%</sub>	hdi <sub>97%</sub>
intercept	3.10	0.07	2.98	3.24
vocabulary $\sigma$	0.50	0.00	0.49	0.51
grade level	-0.15	0.02	-0.18	-0.12
max scores	0.71	0.06	0.59	0.83
mean CEFR-j	3.76	0.22	3.37	4.19
mean scores	-4.46	0.24	-4.92	-4.01
min frequencies	0.05	0.01	0.04	0.07
scale(MTLD)	0.20	0.01	0.18	0.21
type-token ratio	-1.23	0.09	-1.40	-1.06

Table 6: Results of Max+Mean model for predicting the vocabulary scores of ELLIPSE essays.

We also provide the  $R^2$  metric in table 7 in the appendix, because it is more established within in NLP literature, although it neglects the probabilistic information provided by the Bayesian approach. It shows the same picture as the  $\text{elpd}_{100}$  for the five models.

The comparison suggests that adding both the mean- and the max-pooled scores contribute to the fit of the model. The max-pooling, however, contributes only substantially when combined with the mean-pooling. The max+mean model also outperforms the proportion model, showing that the neural models are helpful.

We provide the HDI for our best fitting model in table 6 and figure 7. Among the features, the mean pooled score of our model has the largest absolute coefficient.<sup>26</sup> The coefficient is, however, negative (HDI:  $[-4.92, -4.01]$ ), which might appear surprising at first glance. After all, a higher score should indicate more complex words, which in turn one might expect to indicate a higher proficiency. We believe that this puzzle can be explained by also taking into account the effect of the max-pooled scores.

The effect of the max-pooled scores is smaller, but with high probability positive (HDI:  $[0.59, 0.83]$ ), thus pointing in the expected direction. We interpret this suggestion as follows: the skilled learner produces few contexts that might easily lead to confusion, thus rendering the average word token easier to choose, but their most complex word is more challenging than that of a learner at a lower level.

The surprising negative coefficient is not just present for our scores, but also for type-token ratio<sup>27</sup> (HDI:  $[-1.40, -1.06]$ ) and grade levels of

<sup>26</sup>No direct comparison to frequencies or the scaled MTLD is possible due to the different scale.

<sup>27</sup>The MTLD, however, has the expected relationship, sug-

students (HDI:  $[-0.18, -0.12]$ ). In the case of the minimum frequencies, we find a somewhat surprising positive coefficient (HDI:  $[0.04, 0.07]$ ), suggesting that a higher vocabulary score is associated with avoiding very rare words. De Wilde (2023) has previously found for L2 writing that “more proficient learners use more frequent words” (p. 11), but also notes that the literature is divided on this. These inverted results suggest a non-linear relationship between L2 learner writing and features which the literature associates with lexical sophistication.

	elpd <sub>100</sub>	elpd <sub>diff</sub>	se	dse	R <sup>2</sup>
max+mean	-4709.1	0.0	64.4	0.0	0.26
mean	-4771.8	62.7	65.2	11.3	0.24
proportion	-4853.8	144.7	65.3	20.0	0.22
max	-4873.2	164.1	65.3	18.1	0.22
base	-4874.7	165.6	65.5	18.2	0.22
overall	-5218.0	508.9	59.6	55.0	0.28
phraseology	-5528.4	819.3	57.8	59.6	0.24

Table 7: elpd<sub>100</sub> metrics for downstream application task (predicting vocabulary scores. Besides the main elpd<sub>100</sub>-metric, the table provides the difference to the elpd<sub>100</sub> to the best model, as well as the standard error for these two values (se and dse respectively).

The ELLIPSE dataset also provides other types of scores for student essays against which a comparison is possible. From those we selected the overall score, as it is the most important one, and the phraseology score, as it is the one closest related from vocabulary. By performing a regression with the same features on these scores, we can see whether the features are specific to vocabulary, as intended. Indeed, we find this to be the case for elpd<sub>100</sub> (see results in table 7 and figure 8),<sup>28</sup> despite the well-established halo effect, which leads annotators to provide roughly similar scores (e.g. Engelhard, 1994).

Although further research into the connection between content word replacement errors and vocabulary scores is required, the initial results show that our complexity scores can improve the performance of downstream applications.

gesting that the negative coefficient of the type-token ratio might be due to the essay length.

<sup>28</sup>It is not the case for R<sup>2</sup> in the case of the overall score, but this comparison is not directly admissible, because the variance for the Vocabulary scores (0.36) differs from that of the Overall score (0.41). The comparison of the elpd<sub>100</sub> is only acceptable because the number of data points and the predicted variables share a scale.

## 8 Conclusion

We propose semantic error prediction as a task for investigating lexical semantic production complexity. Such an estimate of complexity is useful for many purposes in educational technology, including assessing output by learners and providing them with information for improving their writing skills.

Complex word identification systems, in contrast, are focused on difficulty in *comprehension* rather than *production*. Semantic error detection/correction system cannot be used this way, because they provide an estimate of how likely a word is to be wrong, not how difficult it was to produce the word in the first place. Semantic error prediction, thus, fills a gap in the CALL literature.

We propose and implement a method for creating semantic error prediction datasets from error correction datasets. Analysing the dataset with Bayesian logistic regressions, we found that verbs show a peculiar accumulation of semantic errors.

Furthermore, we train transformer-embedding based models for semantic error prediction. These models perform better than the baselines, although much room for improvement remains. Finally, we use the scores produced by the best of our models on the downstream task of predicting the vocabulary scores of student essays using a Bayesian linear regression. The results indicate that these lexical complexity scores improve the model.

## Limitations

The present proposal suffers primarily from three limitations:

First, factors other than lexical semantic complexity might lead to content word replacement errors, rendering the proposed error prediction task an imperfect proxy. Future research should investigate other measures for active vocabulary for comparison.

Second, the error correction dataset used for our investigation does not provide information about important properties influencing error patterns, such as the first language of the L2 learners. However, our method is applicable to other datasets providing such information.

Third, our investigation is limited to an English error correction dataset. Error patterns might differ between languages. In some languages, for example morphologically richer languages, content word



replacement errors might be harder to identify or have a weaker connection to lexical semantics.

## References

- Mohammed Qassem Al-Shormani and Yehia Ahmed Al-Sohbani. 2012. [Semantic errors committed by yemeni university learners: Classifications and sources](#). *International Journal of English Linguistics*, 2(66):p120.
- Desislava Aleksandrova and Vincent Pouliot. 2023. [Cefr-based contextual lexical complexity classifier in english and french](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, page 518–527, Toronto, Canada. Association for Computational Linguistics.
- David Alfter and Elena Volodina. 2018. [Towards single word lexical complexity prediction](#). In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, page 79–88, New Orleans, Louisiana. Association for Computational Linguistics.
- Aristotele. 1994. *Categories and De Interpretatione*, reprint edition. Clarendon Aristotle series. Clarendon Press, Oxford.
- Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. [The BEA-2019 Shared Task on Grammatical Error Correction](#). In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.
- Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. [Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.
- Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. [Grammatical Error Correction: A Survey of the State of the Art](#). *Computational Linguistics*, 49(3):643–701.
- Bram Bulté and Alex Housen. 2012. [Defining and operationalising L2 complexity](#), *Language Learning & Language Teaching*, page 21–46. John Benjamins Publishing Company.
- Tomás Capretto, Camen Piho, Ravin Kumar, Jacob Westfall, Tal Yarkoni, and Osvaldo A. Martin. 2022. [Bambi : A Simple Interface for Fitting Bayesian Linear Models in Python](#). *Journal of Statistical Software*, 103(15).
- Council of Europe CoE. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment; Companion Volume*. Council of Europe Publishing, Strasbourg.
- Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023. [The english language learner insight, proficiency and skills evaluation \(ellipse\) corpus](#). *International Journal of Learner Corpus Research*, 9(2):248–269.
- Anne Cutler. 1983. *Lexical complexity and sentence processing*, page 43–79. Wiley, Chichester, Sussex.
- Vanessa De Wilde. 2023. [Lexical characteristics of young l2 english learners’ narrative writing at the start of formal instruction](#). *Journal of Second Language Writing*, 59:100960.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). *arXiv:1810.04805 [cs]*.
- Julie L. Earles and Alan W. Kersten. 2017. [Why are verbs so hard to remember? effects of semantic context on memory for verbs and nouns](#). *Cognitive Science*, 41(S4):780–807.
- George Engelhard. 1994. [Examining rater errors in the assessment of written composition with a many-faceted rasch model](#). *Journal of Educational Measurement*, 31(2):93–112.
- May Fan. 2000. [How big is the gap and how to narrow it? an investigation into the active and passive vocabulary knowledge of l2 learners](#). *RELC Journal*, 31(2):105–119.
- Michael Flor, Steven Holtzman, Paul Deane, and Isaac Bejar. 2024. [Mapping of american english vocabulary by grade levels](#). *ITL - International Journal of Applied Linguistics*.
- Dedre Gentner and Ilene M. France. 1988. [The Verb Mutability Effect: Studies of the Combinatorial Semantics of Nouns and Verbs](#), page 343–382. Morgan Kaufmann.
- Sian Gooding and Ekaterina Kochmar. 2018. [CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting](#). In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, Louisiana. Association for Computational Linguistics.
- Sian Gooding, Ekaterina Kochmar, Seid Muhie Yimam, and Chris Biemann. 2021. [Word Complexity is in the Eye of the Beholder](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449, Online. Association for Computational Linguistics.

- Sylviane Granger. 1998. *The computer learner corpus: a versatile new source of data for SLA research*. Routledge.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength natural language processing in python*.
- Kipsamo E. Jeptarus and Patrick K. Ngene. 2016. Lexico-semantic errors of the learners of english: A survey of standard seven keiyo-speaking primary school pupils in keiyo district, kenya. *Journal of Education and Practice*, 7(13):42–54. ERIC Number: EJ1102824.
- Nan Jiang. 2000. *Lexical representation and development in a second language*. *Applied Linguistics*, 21(1):47–77.
- Mark D. Johnson. 2017. *Cognitive task complexity and l2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis*. *Journal of Second Language Writing*, 37:13–38.
- Alan W. Kersten and Julie L. Earles. 2004. *Semantic context influences memory for verbs more than memory for nouns*. *Memory & Cognition*, 32(2):198–211.
- Aziz Khalil. 1985. *Communicative error evaluation: Native speakers' evaluation and interpretation of written errors of arab efl learners*. *TESOL Quarterly*, 19(2):335–351.
- Minkyung Kim, Scott A. Crossley, and Kristopher Kyle. 2018. *Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality*. *The Modern Language Journal*, 102(1):120–141.
- Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. *Age-of-acquisition ratings for 30,000 English words*. *Behavior Research Methods*, 44(4):978–990.
- Batia Laufer. 1998. *The development of passive and active vocabulary in a second language: Same or different?* *Applied Linguistics*, 19(2):255–271.
- Batia Laufer and Paul Nation. 1995. *Vocabulary size and use: Lexical richness in l2 written production*. *Applied Linguistics*, 16(3):307–322.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. *arXiv:1907.11692 [cs]*.
- Ilya Loshchilov and Frank Hutter. 2018. *Decoupled Weight Decay Regularization*. In *International Conference on Learning Representations*.
- Philip M. McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.d., The University of Memphis, United States – Tennessee. 3199485.
- Masashi Negishi, Tomoko Takada, and Yukio Tono. 2013. *A progress report on the development of the cefr-j*. In *Exploring language frameworks: Proceedings of the ALTE kraków conference*, page 135–163. Citation Key: negishi2013progress.
- Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. *The CoNLL-2014 Shared Task on Grammatical Error Correction*. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.
- Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. *Lexical complexity prediction: An overview*. *ACM Computing Surveys*, 55(9):179:1–179:42.
- Musa Nushi, Roya Jafari, and Masoumeh Tayyebi. 2022. *Iranian advanced efl learners' perceptions of the gravity of their peers' written lexical errors: The case of intelligibility and acceptability*. *Iranian Journal of Foreign Language Teaching Innovations*, 1(1):41–56.
- Margareta Olsson. 1972. *Intelligibility: A Study of Errors and Their Importance*. ERIC Number: ED072681.
- Abril-Pla Oriol, Andreani Virgile, Carroll Colin, Dong Larry, Fannesbeck Christopher J., Kochurov Maxim, Kumar Ravin, Lao Jupeng, Luhmann Christian C., Martin Osvaldo A., Osthege Michael, Vieira Ricardo, Wiecki Thomas, and Zinkov Robert. 2023. *PyMC: A modern and comprehensive probabilistic programming framework in python*. *PeerJ Computer Science*, 9:e1516.
- Gustavo Paetzold and Lucia Specia. 2016. *SemEval 2016 Task 11: Complex Word Identification*. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.
- Alice Pintard and Thomas François. 2020. *Combining expert knowledge with frequency information to infer cefr levels for words*. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with READING Difficulties (READI)*, page 85–92, Marseille, France. European Language Resources Association.
- Muhammad Reza Qorib, Seung-Hoon Na, and Hwee Tou Ng. 2022. *Frustratingly Easy System Combination for Grammatical Error Correction*. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1964–1974, Seattle, United States. Association for Computational Linguistics.
- Muhammad Reza Qorib and Hwee Tou Ng. 2023. *System Combination via Quality Estimation for Grammatical Error Correction*. In *Proceedings of the 2023*

- Conference on Empirical Methods in Natural Language Processing*, pages 12746–12759, Singapore. Association for Computational Linguistics.
- Matthew Shardlow. 2013. The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, Sofia, Bulgaria. Association for Computational Linguistics.
- Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — A New Corpus for Lexical Complexity Prediction from Likert Scale Data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with READING Difficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.
- Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. **SemEval-2021 Task 1: Lexical Complexity Prediction**. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. **Llama 2: Open Foundation and Fine-Tuned Chat Models**.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. **Practical bayesian model evaluation using leave-one-out cross-validation and waic**. *Statistics and Computing*, 27(5):1413–1432.
- Elena Volodina, Christopher Bryant, Andrew Caines, Orphée De Clercq, Jennifer-Carmen Frey, Elizaveta Ershova, Alexandr Rosen, and Olga Vinogradova. 2023. MultiGED-2023 shared task at NLP4CALL: Multilingual Grammatical Error Detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 1–16, Tórshavn, Faroe Islands. LiU Electronic Press.
- Elena Volodina, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse, and Thomas François. 2016. **Swellex: Second language learners’ productive vocabulary**. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, page 76–84, Umeå, Sweden. LiU Electronic Press.
- Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. **Probing Pretrained Language Models for Lexical Semantics**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. **Transformers: State-of-the-Art Natural Language Processing**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017. CWIG3G2 - Complex Word Identification Task across Three Text Genres and Two User Groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 401–407, Taipei, Taiwan. Asian Federation of Natural Language Processing.

## A Further Results and Details

$$\mathbf{B:} \quad \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + c$$

$$\mathbf{B+I:} \quad \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + c + \beta_6 X_2 X_4$$

$X_1 = \# \text{ characters}$                        $X_4 = \text{word cefr-j level}$   
 $X_2 = \text{frequency}$                          $X_5 = \begin{cases} 1 & \text{if token is a verb} \\ 0 & \text{in other case} \end{cases}$   
 $X_3 = \text{age of acquisition}$      $c = (\beta_{A2} C_{A2} + \beta_{B1} C_{B1} + \dots) = \text{effect of student CEFR level}$

Figure 6: Equations describing the two Bayesian logistic regression models: Basic (B) and Basic with Interaction added (B+I).

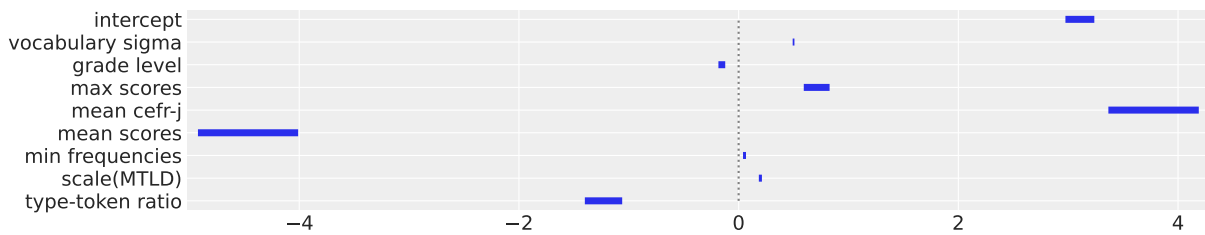


Figure 7: HDIs for Max+Mean model predicting the vocabulary scores. Max and mean scores refer to the pooled results of our neural model.

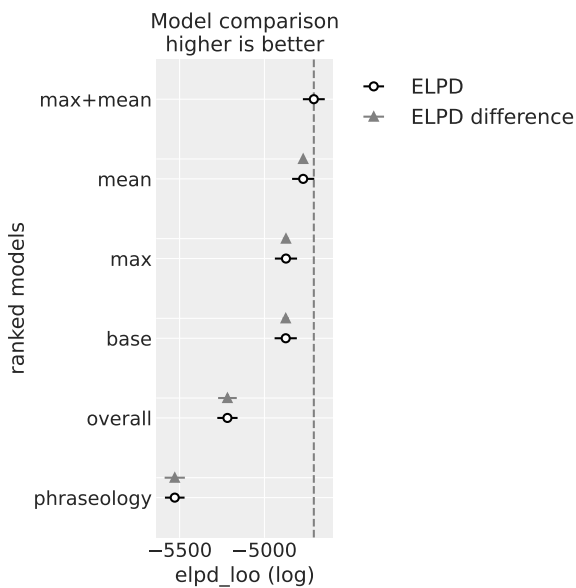


Figure 8:  $elpd_{loo}$  scores for Bayesian linear regression models predicting vocabulary scores (top 4 model) as well as Overall scores and Phraseology scores.

	space	BERT	RoBERTa	LLAMA2
midrule batch size	{640, 1280, 1920, 2560, 3200}	2560	2560	640
learning rate	$\{1 \cdot 10^{-2}, 5 \cdot 10^{-3}, 1 \cdot 10^{-3}, 5 \cdot 10^{-4}, 1 \cdot 10^{-4}, 5 \cdot 10^{-5}, 1 \cdot 10^{-5}\}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-5}$
epochs	{20, 30, 40, 50, 60}	50	50	40

Table 8: Hyperparameters search space and selected hyperparameters.

# GRAMEX: Generating Controlled Grammar Exercises from Various Sources

Guillaume Toussaint<sup>1</sup>

(1) CNRS, LORIA  
Vandœuvre-Lès-Nancy  
France

Yannick Parmentier<sup>2,3</sup>

(2) Université de Lorraine, LORIA  
(3) Université d'Orléans, LIFO  
Vandœuvre-Lès-Nancy, France

Claire Gardent<sup>1</sup>

(1) CNRS, LORIA  
Vandœuvre-Lès-Nancy  
France

firstname.lastname@loria.fr

## Abstract

This paper presents GRAMEX, an application designed to assist teachers in the creation of learning materials, namely grammar exercises. More precisely GRAMEX leverages state-of-the-art parsing techniques to morpho-syntactically annotate texts and turn these into grammar exercises while aligning these with official curricula. Allowing teachers to freely select excerpts of texts from which to generate specific grammar exercises aims to increase learners' engagement in educational activities. GRAMEX currently supports 4 types of exercises (Fill-in-the-Blanks, Mark-the-Words, Single and Multiple Choice questionnaires) and 3 output formats (JSON objects, printable workbooks, H5P interactive content). GRAMEX is under active development and has been experimentally used with teachers of L1-learners in elementary and middle French schools.

## 1 Introduction

Grammar learning is known to have a strong impact on language learning in general. Indeed, studies showed that a lack of self-confidence in one's own grammatical skills often leads to broader difficulties in language learning and writing (Ignacia-Dorronzoro and Klett, 2007; Castagné-Véziès, 2018). Further investigations also suggest that isolating grammar practice from other learning activities results in higher learning difficulties (Vincent, 2016). This, combined with the positive effects on learners' motivation observed by Peacock (1997), advocates for the use of authentic texts (possibly seen in various contexts) as a valuable resource for automatic generation of grammar exercises.

The GRAMEX project builds on this idea to provide (1) teachers with a digital environment which

can be used to generate grammar exercises from user-defined texts and learning goals and (2) learners with an online facility to train and monitor their progress. The generated exercises are annotated with fine-grained morphological and syntactic information along with readability scores (François and Fairon, 2012) and links to official curricula, allowing teachers to control exercise generation, ensuring the adequacy of the output material for target learners.

Along with this **control** on exercise generation, GRAMEX features include:

**robustness** : the use of efficient neural parsing techniques combined with error analysis on parse trees makes it possible to filter out sentences leading to ill-formed questions ;

**multilingualism** : two languages have been tested so far (French and English), yet GRAMEX relies on multilingual parsing engines covering 20+ languages ;

**extensibility** : thanks to its modular architecture, GRAMEX can easily be extended to other languages or new exercise types (see Section 3) ;

**interoperability** : GRAMEX comes with a REST Application Programming Interface (API) and 3 export formats (JSON objects, printable workbooks in docx format, interactive content in H5P format), allowing users to interact with GRAMEX in many ways, including within Learning Management Systems (LMS), external applications or in a classical paper-based setting.

The remaining of this paper is organized as follows. In Section 2, we present related work. In Section 3, we describe how teachers can use GRAMEX to generate grammar exercises, and how learners can complete these. In Section 4, we present GRAMEX's implementation. In Section 5,

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

we comment on GRAMEX status and discuss its current limitations and ongoing work. We finally conclude and present future work in Section 6.

## 2 Related work

There have been many approaches to automatic generation of grammar exercises over the last decades. The corresponding systems distinguish themselves according to their core functionalities.

A first distinction can be made between systems supporting **custom text input** and those relying on predefined resources (corpora, grammars and / or lexicons). The latter includes *ArikIturri* (Aldabe et al., 2006), *Grammar Exerciser* (Perez-Beltrachini et al., 2012) and *Lärka* (Volodina et al., 2014). Systems allowing users to enter free text, like GRAMEX, include *MIRTO* (Antoniadis et al., 2006), *Sakumon* (Hoshino and Nakagawa, 2007), *VIEW* (Meurers et al., 2010), *Language Muse* (Madnani et al., 2016), *Language Exercise App* (Perez and Cuadros, 2017), and *FLAIR* (Heck and Meurers, 2022b). These notably differ in the way input texts are pre-processed to extract candidate sentences. In the case of GRAMEX, sentence filtering is done by means of fine-grained morpho-syntactic annotations computed by state-of-the-art text parsers (namely, SpaCy<sup>1</sup> (Honnibal and Johnson, 2015) and Stanza<sup>2</sup> (Qi et al., 2020))<sup>3</sup> combined in tailored NLP pipelines (see Section 3), while other systems rely either on partial analyses involving specific part-of-speech tags or syntactic patterns (e.g. *MIRTO*, *Lärka*, *Sakumon*),<sup>4</sup> or on more abstract representations such as sentence vectorization (e.g. *Language Exercise App*).

A second distinction concerns their **degree of automation**. Most systems require human intervention (i.e., post-edition of questions, such as the selection of distractors in Multiple Choice questions) to create ready-to-use grammar exercises. GRAMEX is designed to limit such intervention as much as possible. Users are merely required to validate (and optionally reorder) output questions. This design choice is questionable, and may be revised in the light of experimental studies involving

<sup>1</sup><https://spacy.io>

<sup>2</sup><https://github.com/stanfordnlp/stanza>

<sup>3</sup>Note that these are not limited to syntactic analysis *sensu stricto*, they include many (neural and / or symbolic) modules for broader text analysis.

<sup>4</sup>Like GRAMEX, FLAIR uses state-of-the-art parsers, but only specific annotations are considered for exercise generation, following work of Pilán et al. (2016) on candidate sentence selection.

school teachers to be carried out in a near future.

A third distinction can be made on the **level of control** offered by these systems. Systems generally offer a limited control on the generation of exercises. Noticeable exceptions include *Language Exercise App*, where users can define target constructions, *MIRTO*, where users can also link questions to references providing learners with helpful information, *Language Muse*, which generates about 24 predefined activities at various levels (sentence, paragraph, discourse) and *FLAIR*, which comes with a highly configurable generation process, where users can for instance define additional parameters depending on the target grammatical phenomenon (Heck and Meurers, 2022b). In our case, a trade-off between configurability and usability is being sought. GRAMEX currently allows users to target precise predefined grammatical concepts extracted from official curricula (MENJS, 2018). A more fine-grained control is under development, allowing for instance to select target syntactic structures (see Section 5).

A fourth distinction concerns their **expressivity**, that is, the types of exercises they support. Most systems support Multiple Choice (MC) questionnaires since these can be automatically processed to evaluate learners' performances. The number and types of supported exercises vary from one system to another. GRAMEX currently supports 4 exercise types, namely Fill-in-the-Blanks (FiB), Mark-the-Words (MtW), Multiple Choice (MC) and Single Choice (SC). Other common exercise types, not yet supported by GRAMEX include Error Detection (ED), Memory (Mem), Shuffle (Sh) and Word Forms (WF). Table 1 summarizes the expressivity of the above-mentioned systems with respect to these types.

Finally, let us note that relatively few systems are able to **export** exercises to be integrated in external tools (i.e., Learning Management Systems) out-of-the-box.<sup>5</sup> Such systems include *VIEW*, which is a browser extension and as such can be integrated natively with web interfaces, *Language Exercise App* and GRAMEX, which can both export exercises in H5P format (interactive HTML5 content)<sup>6</sup> supported by many LMS.

<sup>5</sup>*ArikIturri* exports exercises in XML format, which is not directly usable e.g. in an LMS, but can be relatively easily converted to other formats for integration.

<sup>6</sup><https://h5p.org/>

System	SC	MC	ED	MtW	FiB	Mem	Sh	WF	Other
<i>MIRTO</i>				×	×				
<i>ArikIturri</i>		×	×		×			×	
<i>FAST</i>		×	×						
<i>Sakumon</i>		×							
<i>VIEW</i>		×		×	×				
<i>Grammar Exerciser</i>				×		×			
<i>Lärka</i>		×							
<i>Language Muse</i>	×	×							×
<i>Language Exercise App</i>		×		×	×		×		
<i>FLAIR</i>		×		×	×	×	×	×	
<i>GRAMEX</i>	×	×		×	×				

Table 1: Exercise types supported by exercise generation systems (these are in chronological order)

### 3 Workflow description

In a nutshell, GRAMEX is a web application allowing teachers to create exercises from custom texts depending on target grammatical phenomena and learner levels. These exercises can be shared with other users or exported for reuse in other applications (e.g. LMS). Teachers can furthermore create collections of activities (so-called lessons) which follow given learning paths. In the following subsections, we go through the various steps involved in exercise generation.

#### 3.1 Selecting and annotating input data

In order to generate exercises, users need to first select an adequate input text.<sup>7</sup> They can select from the following sources : Wikipedia articles, web pages (identified by their URL), local files<sup>8</sup> and free (e.g. copy-pasted) texts.

From this source, the text is extracted (i.e., formatting information is removed) and fed to a custom yet classical NLP pipeline for text annotation. This pipeline builds on state-of-the-art parsers to perform various tasks sequentially: sentence segmentation, tokenization, part-of-speech tagging, morphological analysis and syntactic dependency parsing. As a result, each sentence from the input text is annotated with morpho-syntactic information in CoNLL format (Buchholz and Marsi, 2006) and stored in GRAMEX’s database.

Additionally, we also compute and store, for each annotated sentence, its readability scores

<sup>7</sup>In our approach, the adequacy between a text and a target grammatical phenomenon is *by design* left to the teacher.

<sup>8</sup>For now, only text and pdf files are allowed, docx files will be supported soon.

(e.g. Flesch–Kincaid (Kincaid et al., 1975)).<sup>9</sup> These scores may be used by teachers to order exercises depending on their readability or to adapt activities to pupils with special educational needs.<sup>10</sup>

The sentences which have been annotated (e.g., whose length is above a given threshold and which contain at least one of the target grammatical phenomena) can be inspected as illustrated in Figure 1. Sentences which should not be used in exercises (e.g. due to an inadequate vocabulary) can be manually filtered out by teachers at this step.

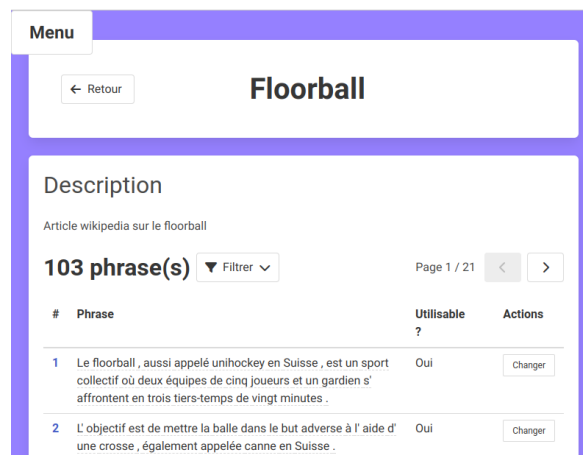


Figure 1: Teacher’s interface to inspect annotated texts

<sup>9</sup>Other readability assessment Machine Learning-based techniques have been implemented following work by Hernandez et al. (2022), see (Ngo and Parmentier, 2023) ; their precision on representative data is yet to be evaluated prior to integrating these into GRAMEX.

<sup>10</sup>Currently GRAMEX does not use these readability scores for exercise generation, they are only displayed to teachers in order to help them to select which sentences to use in exercises. An automatic ordering of questions based among others on these metrics, will be explored in a near future.



### 3.2 Filtering annotated data

Once the selection and annotation phase described above is done, user-validated annotated sentences are stored in GRAMEX’s database, together with their readability scores. Neural text analysis modules by their statistical nature may produce erroneous annotations (e.g., wrong morphological features). This is especially true since our NLP pipeline applies *pre-trained* dependency parsers to potentially out-of-domain data.

In order to detect annotations which are likely to be erroneous before exercise generation, a statistical filtering is applied. In brief, we compiled a corpus  $C_{err}$  of annotation errors by comparing our pipeline’s annotations with a gold-standard (i.e., manually annotated) dataset made of 23,750 sentences coming from the French section of the Universal Dependency corpus (Nivre, 2016). From this corpus  $C_{err}$ , we experimented with various machine (deep and non-deep) learning algorithms in order to predict whether a given annotated sentence should be flagged as invalid. The best results were obtained by using gradient boosting (Friedman, 2002) reaching an F-score of 0.63.<sup>11</sup>

Note that, whatever the result of this filtering step is, the annotated sentence is kept in the database so that users can manually inspect or edit it should they want to.

### 3.3 Aligning annotated data with target grammatical phenomena

In order to control exercise generation with respect to target grammatical concepts,<sup>12</sup> we define an alignment between these and morpho-syntactic annotations generated by our NLP pipeline. This candidate selection is based on curricula-based predefined filters.

To facilitate the maintenance and extension of GRAMEX, these alignment filters are defined in configuration files and use a custom description language inspired by the Grew corpus query language (Guillaume, 2021) to specify which morpho-syntactic annotations contribute to a given target grammar concept. As an illustration, let us consider the following specification:

```
[upos=VERB&Mood=Ind&Tense=Fut]
```

Here, we specify a combination of annotations which are characteristics of sentences having a

<sup>11</sup>Filtering is work in progress, especially since all annotation errors are not of equal importance in our context.

<sup>12</sup>Recall that these concepts come from official curricula.

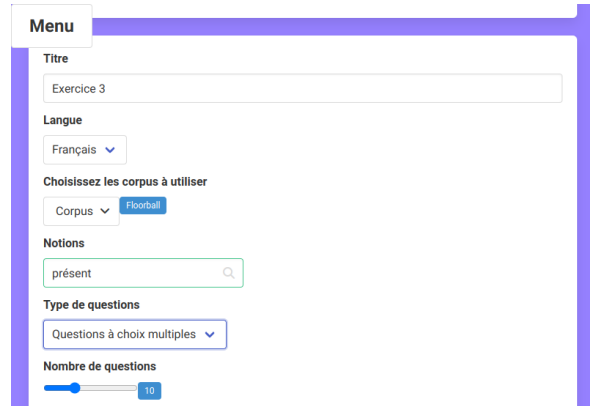


Figure 2: Teacher’s interface to create exercises

verb in future tense. It reads as follows: the sentence must contain a token whose part-of-speech tag is VERB, and whose morpho-syntactic features include Mood, Ind and Tense, Fut as key,value pairs.

Concretely, once a text is fully annotated by our NLP pipeline, these alignment filters are used to check the presence of any target grammatical concept in annotated sentences and, in case of success to keep their locations in the sentence (and store them together with the morpho-syntactic annotations in GRAMEX’s database).

It is worth noting that, although most of the grammatical phenomena listed in official curricula are correctly flagged, some (such as simple past in French) are consistently not. We suspect this is due to the under-representation of these phenomena in parsers’ training data. In order to circumvent this issue, we use a rule-based approach (e.g. a verb conjugation algorithm) to overwrite the annotations given by the parser. In case of ambiguity (same morpheme for several tenses), we keep all possible annotations in the database.

### 3.4 Generating exercises

In order to create an exercise, users have to choose (i) a text (within the corpus of texts they have previously asked GRAMEX to annotate), (ii) a target grammatical concept to work on, (iii) a type of exercise (among the 4 types currently supported by GRAMEX, namely FiB, MtW, SC and MC), and (iv) a number of questions, see Figure 2. Exercises are also given a title and optional keywords to facilitate their indexing and reuse.

From this configuration, GRAMEX retrieves in the selected annotated text, the expected number of sentences exhibiting the target grammatical

concept. A transformation rule is then applied to the corresponding sentences to turn them into the required exercise type. Note that, if the selected text contains more occurrences of the target grammatical concept than the required number of questions, a random selection is done. This is subject to modification in the future (see Section 5).

The user is then presented with the generated exercise, and has the option to replace questions and / or re-order them. Figure 3 shows an example of a FiB exercise on past perfect in French generated by GRAMEX.

### 3.5 Exporting exercises

Export refers to the possibility for users to download exercises in a given format. This is useful for creating backups, post-editing exercises or else sharing exercises with other teachers (who will import them). Supported export formats include JSON for programmatic uses, word documents for paper-based activities, and H5P components for use in dedicated (on-line or desktop) environments equipped with an H5P player (e.g., Lumi<sup>13</sup>).

Figure 4 gives an example of FiB and MC exercises exported in H5P format. Note that if needed, H5P components can be modified using the free H5P editor.<sup>14</sup>

### 3.6 Sharing exercises

Sharing refers to the possibility for teachers to give access to their exercises to other users. Sharing can either be public (that is, to all registered users) or else restricted to specific users only. Public exercises can be retrieved using a text-based search on their title, keywords and content.

Exercises shared with specific users (learners or groups of learners) can be accessed on invitation or else by using auto-generated access codes.

### 3.7 Taking exercises

Learners can access exercises from their dashboard directly if they have been invited by their teacher, or else by using their access code. In both cases, questions can be answered in a dedicated interface (see Figure 5). Once the exercise is completed, students are presented with a summary of their successes and failures (see Figure 6). All attempts can also be monitored by the teacher.

<sup>13</sup><https://lumi.education/en/>

<sup>14</sup><https://h5p.org/installation>

## 4 Implementation

GRAMEX relies on a client-server architecture, with a front-end in JavaScript / VueJS<sup>15</sup> and a back-end in Python. These components interact with a MySQL database following a classical Model-View-Controller design pattern (Krasner and Pope, 1988) as illustrated in Figure 7.

GRAMEX's database basically contains information about users (teachers and learners) and learning materials. These pieces of information are organized as follows. Teachers can manage learners' accounts, corpora (collections of annotated texts) and learning activities. Activities can either be a so-called *lessons* gathering textbooks and exercises, or *tests* (standalone exercises). Both lessons and tests can be shared with specific learners depending on their profile and / or teachers' pedagogical choices.

Note that GRAMEX's exercise generation module is used in a similar way when creating lessons or tests. The only difference lies in whether they are used in the context of formative or summative assessment (Sadler, 1998) by teachers (unlike exercises belonging to lessons, exercises from tests can be taken only once).

GRAMEX comes with a web user interface built with the Bulma CSS framework<sup>16</sup>. Users can use GRAMEX through responsive web pages designed for computers and tablets.

The back-end hosts GRAMEX's NLP pipeline and database. It also offers a REST API developed in Flask<sup>17</sup> allowing programs (including the front-end) to interact with the hosted components via the controller module. The back-end handles all data manipulations, from text annotation to exercise generation and export.

Note that the back-end also includes a typescript module responsible for generating H5P components, which are served by the API for download and reuse in other applications.

## 5 Current status

GRAMEX is under active development. Design choices are subject to modification depending on feedback from teachers. In the following subsections, we briefly report on a first 2-week experiment, which highlighted some limitations calling for further development.

<sup>15</sup><https://vuejs.org/>

<sup>16</sup><https://bulma.io/>

<sup>17</sup><https://flask.palletsprojects.com/>





















Numéro	Phrase	Réponse	Lisibilité	Actions
1	Il y avait une famille qui les _____ (avoir) _____ (voir) et qui s'est dit : «	avait ; vus	108	   
2	De son côté, le chat était très content : il _____ (avoir) _____ (trouver) une chatte.	avait ; trouvé	105	   
3	Et ça y est, elle l' _____ (avoir) _____ (choisir) !	avait ; choisi	104	   
4	Mimi lui dit qu'il _____ (être) _____ (partir) à la recherche de son frère et ils rentrèrent à la maison.	était ; parti	93	   
5	Pourtant, chaque fois, il _____ (avoir) _____ (essayer) de se retenir très fort.	avait ; essayé	92	   

Figure 3: Fill-in-the-Blanks exercise generated targeting the past perfect tense

Conjugué les verbes au présent de l'indicatif

On y \_\_\_\_\_ (évoluer) à 4 joueurs, dont un gardien.

○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

---

Les buts sont distants de 18 m (la moitié d'un grand terrain ou la taille d'un terrain de volley) mais **conservent** la même taille que sur le grand terrain.

Les mots en gras sont :

au passé simple de l'indicatif

au présent de l'indicatif

au plus-que-parfait

à l'indicatif

○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

Figure 4: Exercises imported on an external H5P player

## 5.1 Report on experimental uses

GRAMEX has been used by a pool of 4 school teachers whose pupils' age range from 9 to 15 years old. They focused on the development of exercises from raw data (that is, they did not create lessons). In a few cases (where computers were available in classrooms), generated exercises were presented to pupils. In the end of the experiment,

Menu

**Questions "Test 1 - Exercice 1"**

---

Question n°1/5

Dans la caisse de bois  un troisième homme qui avait fini de se battre – un homme que le Wild avait vaincu, qu'il avait harcelé jusqu'à ce que son corps ait cessé pour toujours de se mouvoir.

Conjugué les verbes à l'imparfait de l'indicatif

Figure 5: Learner's interface for taking questions

a questionnaire was sent to teachers to get their feedback on GRAMEX's usability (how easy / convenient it is to use GRAMEX?) and performance (how pertinent are the generated exercises?).

On the usability side, some pupils had difficulty logging in with auto-generated passwords. Teachers recommended the use of QR-codes to provide learners with a connection link. Teachers had troubles understanding the logic behind exercises and lessons. On the performance side, teachers encountered issues with complex web pages (extracted texts were noisy), and were wishing one could feed PDF files to the application. Finally teachers indicated they would need more control on sentence selection. For instance, they would like to be able to control the presence of various syntactic constructions in selected questions.

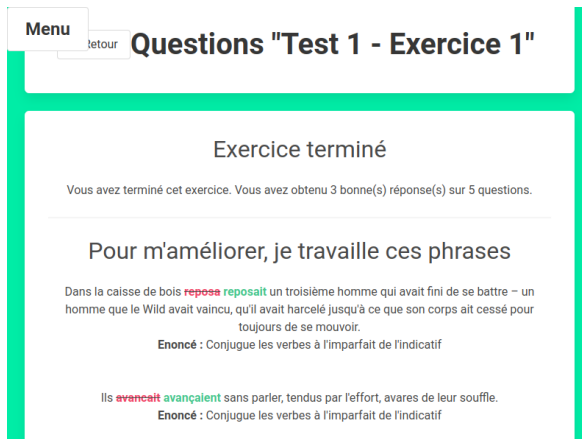


Figure 6: Learner’s interface on exercise completion

## 5.2 Limitations and current work

As mentioned above, GRAMEX has been designed to facilitate exercise generation by providing users with a semi-automatic process requiring only a lightweight configuration. Candidate grammatical concepts (and corresponding morpho-syntactic annotations) are predefined and can be used out-of-the-box. It turns out that this configuration is not sufficient as teachers cannot precisely control the structure of generated questions. GRAMEX’s workflow is thus being extended to give teachers the possibility to define syntactic constraints on the selected sentences. These constraints are written in the same description language as curricula-related filters (see Section 3.3).

Another main limitation of GRAMEX lies in its use of pre-trained neural modules for text analysis. As mentioned above, these modules are applied on unknown texts (potentially out-of-domain). Even though a statistical filter is applied, teachers cannot be guaranteed that the provided annotations (and thus exercises) are correct. We are currently working on the development of another annotation error detection module. Two paths are being considered : using ensemble techniques which would basically compare between annotations computed by distinct parsers following work by [Surdeanu and Manning \(2010\)](#), and using a rule-based approach where predicted dependency rules would be compared with a dependency grammar extracted from manually annotated data following work by [Rehbein and Ruppenhofer \(2018\)](#).

Another limitation of GRAMEX corresponds to the limited types of exercises it supports. This combined with the fact that FiB does not support answers which would deviate from original texts

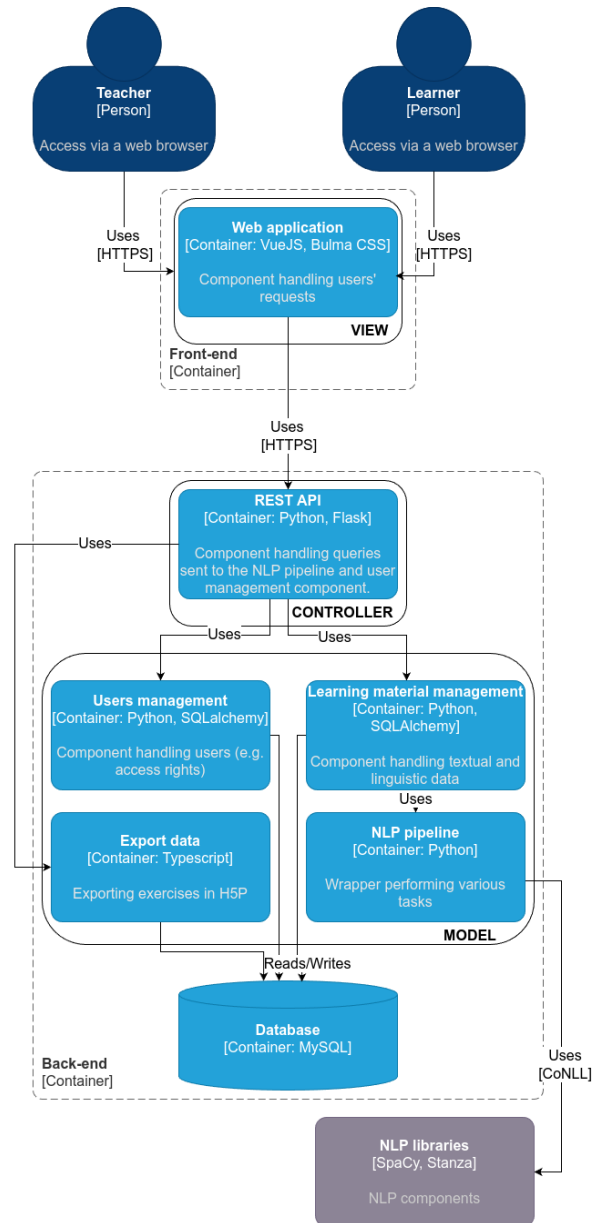


Figure 7: GRAMEX architecture

while being correct, makes it crucial to extend GRAMEX with new exercise types.

## 6 Conclusion and perspectives

We presented GRAMEX, an environment for CALL using state-of-the-art parsers to generate grammar exercises in line with official curricula. GRAMEX aims to help teachers to create adequate learning materials with minimal efforts. GRAMEX is work in progress and benefits from cooperations with field teachers. Future improvements include the configuration of exercises by means of an expressive search engine following [Heck and Meurers \(2022a\)](#), and the extension to new languages.

## References

- Itziar Aldabe, Maddalen Lopez de Lacalle, Montse Maritxalar, Edurne Martinez, and Larraitz Uria. 2006. Arikiturri: An automatic question generator based on corpora and nlp techniques. In *Intelligent Tutoring Systems*, pages 584–594, Berlin, Heidelberg, Springer Berlin Heidelberg.
- Georges Antoniadis, Sandra Echinard, Olivier Kraif, Thomas Lebarbé, and Claude Ponton. 2006. [Modélisation de l’intégration de ressources TAL pour l’apprentissage des langues : la plateforme MIRTO](#). *ALSIC - Apprentissage des Langues et Systèmes d’Information et de Communication*, 08(1):65–79.
- Sabine Buchholz and Erwin Marsi. 2006. [CoNLL-X shared task on multilingual dependency parsing](#). In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City. Association for Computational Linguistics.
- Clotilde Castagné-Véziès. 2018. [La Grammaire à l’épreuve de la langue et de la métalangue](#). *Cognition, représentation, langue (Corela)*, 16(1).
- Thomas François and Cédric Fairon. 2012. [An “AI readability” formula for French as a foreign language](#). In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, Jeju Island, Korea. Association for Computational Linguistics.
- Jerome H. Friedman. 2002. [Stochastic gradient boosting](#). *Computational Statistics & Data Analysis*, 38(4):367–378. Nonlinear Methods and Data Mining.
- Bruno Guillaume. 2021. [Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion](#). In *EACL 2021 - 16th conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Kiev/Online, Ukraine.
- Tanja Heck and Detmar Meurers. 2022a. [Generating and authoring high-variability exercises from authentic texts](#). In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 61–71, Louvain-la-Neuve, Belgium. LiU Electronic Press.
- Tanja Heck and Detmar Meurers. 2022b. [Parametrizable exercise generation from authentic texts: Effectively targeting the language means on the curriculum](#). In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 154–166, Seattle, Washington. Association for Computational Linguistics.
- Nicolas Hernandez, Nabil Oulbaz, and Tristan Faine. 2022. [Open corpora and toolkit for assessing text readability in French](#). In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with READING Difficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 54–61, Marseille, France. European Language Resources Association.
- Matthew Honnibal and Mark Johnson. 2015. [An improved non-monotonic transition system for dependency parsing](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.
- Ayako Hoshino and Hiroshi Nakagawa. 2007. A cloze test authoring system and its automation. In *Proceedings of the 6th International Conference on Advances in Web Based Learning, ICWL’07*, page 252–263, Berlin, Heidelberg. Springer-Verlag.
- María Ignacia-Dorronzoro and Estela Klett. 2007. [Le rôle de la grammaire dans l’enseignement de la lecture en langue étrangère](#). *Éla. Études de linguistique appliquée*, 148:499–511.
- J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.
- Glenn E. Krasner and Stephen T. Pope. 1988. A cookbook for using the model-view controller user interface paradigm in smalltalk-80. *Journal of Object Oriented Programming*, 1(3):26–49.
- Nitin Madnani, Jill Burstein, John Sabatini, Kietha Biggers, and Slava Andreyev. 2016. [Language muse: Automated linguistic activity generation for English language learners](#). In *Proceedings of ACL-2016 System Demonstrations*, pages 79–84, Berlin, Germany. Association for Computational Linguistics.
- Ministère de l’Éducation Nationale de la Jeunesse et des Sports MENJS. 2018. [Programmes d’enseignement. cycle des apprentissages fondamentaux \(cycle 2\), cycle de consolidation \(cycle 3\) et cycle des approfondissements \(cycle 4\) : modification \(publication du bulletin officiel no mene1820169a\)](#).
- Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. [Enhancing authentic web pages for language learners](#). In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18, Los Angeles, California. Association for Computational Linguistics.
- Duy Van Ngo and Yannick Parmentier. 2023. [Towards Sentence-level Text Readability Assessment for French](#). In *Second Workshop on*

- Text Simplification, Accessibility and Readability (TSAR@RANLP2023)*, Varna, Bulgaria.
- Joakim Nivre. 2016. [Universal Dependencies: A cross-linguistic perspective on grammar and lexicon](#). In *Proceedings of the Workshop on Grammar and Lexicon: interactions and interfaces (GramLex)*, pages 38–40, Osaka, Japan. The COLING 2016 Organizing Committee.
- Matthew Peacock. 1997. [The effect of authentic materials on the motivation of EFL learners](#). *ELT Journal*, 51(2):144–156.
- Naiara Perez and Montse Cuadros. 2017. [Multilingual CALL framework for automatic language exercise generation from free text](#). In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 49–52, Valencia, Spain. Association for Computational Linguistics.
- Laura Perez-Beltrachini, Claire Gardent, and German Kruszewski. 2012. [Generating grammar exercises](#). In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 147–156, Montréal, Canada. Association for Computational Linguistics.
- Ildikó Pilán, Elena Volodina, and Lars Borin. 2016. [Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation](#). *Traitement Automatique des Langues*, 57(3):67–91.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- Ines Rehbein and Josef Ruppenhofer. 2018. [Sprucing up the trees – error detection in treebanks](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 107–118, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- D. Royce Sadler. 1998. [Formative assessment: revisiting the territory](#). *Assessment in Education: Principles, Policy & Practice*, 5(1):77–84.
- Mihai Surdeanu and Christopher D. Manning. 2010. [Ensemble models for dependency parsing: Cheap and good?](#) In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 649–652, Los Angeles, California. Association for Computational Linguistics.
- François Vincent. 2016. [Articulation entre la grammaire, l’écriture et la lecture : résultats d’une recension d’écrits](#).
- Elena Volodina, Ildikó Pilán, Lars Borin, and Therese Lindström Tiedemann. 2014. [A flexible language learning platform based on language resources and web services](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 3973–3978, Reykjavik, Iceland. European Language Resources Association (ELRA).

# LLM chatbots as a language practice tool: a user study

Gladys Tyen, Andrew Caines, Paula Buttery

ALTA Institute, Dept. of Computer Science & Technology

University of Cambridge

{gladys.tyen, andrew.caines, paula.buttery}@cl.cam.ac.uk

## Abstract

Second language learners often experience language anxiety when speaking with others in their target language. As the generative capabilities of Large Language Models (LLMs) continue to improve, we investigate the possibility of using an LLM as a conversation practice tool. We conduct a user study with 160 English language learners, where an LLM chatbot is used to simulate real-world conversations. We present our findings on 1) how an interactive session with a chatbot might impact performance in real-world conversations; 2) whether the learning experience differs for learners of different proficiency levels; 3) how changes in difficulty affects the learner's experience; and 4) how online, synchronous conversation provided by an LLM compares with a purely receptive experience. Additionally, we propose a simple yet effective way to detect linguistic complexity on-the-fly: clicking on words to reveal dictionary definitions. We demonstrate that clicks correlate well with linguistic complexity and indicate which words learners find difficult to understand.

## 1 Introduction

Rapid advancements in natural language processing technology, brought on by large language models (LLMs), have opened up new directions and methods for learning and education. In particular, language learners have been making use of LLMs' language generation abilities to support their learning experience (e.g. [PrettyPolly, 2023](#); [Microsoft, 2023](#)).

In this paper, we investigate the possibility of using an LLM for conversational practice in language learning. Many existing approaches restrict the LLM in some way (e.g. [Duolingo Team, 2023](#); [Zhang and Huang, 2024](#)), requiring manual crafting of prompts or syllabuses. Restrictions are

common for pre-LLM chatbots in language learning ([Bibauw et al., 2019](#)), as they are rule-based and can often fail to parse user input correctly. However, as LLM technology advances, these restrictions may no longer be needed.

In our study, we test the limits of LLM capabilities by using an LLM directly without any restrictions on topic, context, or grammatical form. We conduct a user study with 160 English learners, who are asked to interact with an online chatbot. In our implementation, our chatbot is designed to simulate a typical conversationalist so that the learner can practise chatting in English.

We seek to answer the following research questions:

- RQ1.** Does chatting with an online chatbot have any educational impact on real-life interaction?
- RQ2.** How does the language learning experience change for learners at different proficiency levels?
- RQ3.** Does adjustment of difficulty level affect the learner's experience, either positively or negatively?
- RQ4.** How does a conversational setting (combining comprehension and production) compare to a comprehension-only setting?

Overall, our results suggest that chatbots for conversational practice have positive educational impact, though further investigation is required in some areas. We find that this setup is more suited to learners at lower proficiency levels; that it provides more enjoyment over plain reading; and that personalised difficulty adaptation prevents dialogues from becoming too easy. Detailed findings can be found in Section 4.

Additionally, we propose a simple but effective way to identify linguistic complexity during

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

a chatbot conversation: clicking to reveal dictionary definitions. This function can be seamlessly integrated into any web interface, and our results demonstrate a clear correlation between clicking and what the learner finds difficult.

## 2 Background

Before the advent of transformers and LLMs, chatbots for computer-assisted language learning (CALL) were typically rule-based and were only used for constrained scenarios. Bibauw et al. (2019) present a pre-LLM survey of dialogue systems for language learning, and observe that most systems have implicit or explicit constraints, on either the content of the user response or the grammatical form. Ones that allow free dialogue are typically rule based and prone to producing ungrammatical or nonsensical messages (e.g. Coniam, 2014; Jia, 2009)

However, as most chatbots worked within these constraints, it was also easier to introduce adjustments to the chatbot for language learning purposes. One of the most common adjustments is the adaptation of difficulty level based on the user’s linguistic proficiency or previous performance, for example as implemented by Hassani et al. (2016); Lu et al. (2006); Ní Chiaráin and Ní Chasaide (2016); Su et al. (2015); Vlugter et al. (2009).

With the introduction of neural dialogue systems and later LLMs, the performance of chatbots improved greatly (Papangelis et al., 2021; Adwardana et al., 2020; Roller et al., 2021). This technology made it possible to build chatbots for CALL with little to no constraints, while generating grammatical sentences. For example, Tyen et al. (2022) propose a chatbot setup where the difficulty of generated text can be adjusted to user’s proficiency level; Lee et al. (2023) propose a system (with some restriction on context) that produces feedback for students; Zhang and Huang (2024) investigate how vocabulary acquisition is affected by 4 types of chatbots for 4 contexts, all connected to an LLM backend. Additionally, the release of ChatGPT (OpenAI, 2023) prompted some language learners to use the service to help them learn (Microsoft, 2023), even though ChatGPT is not specifically designed for language learning.

Despite advances in technology and commercial chatbots for language learning, there is limited research on the effect of using unconstrained

LLM chatbots to learn a second language. Previous studies use chatbots that are limited to predetermined contexts (Lee et al., 2023; Zhang and Huang, 2024), or that are rule-based (Coniam, 2014; Jia, 2009), with the feedback that the chatbot is difficult to understand or responds with ungrammatical or nonsensical messages.

In our paper, we use an open-domain LLM chatbot, with no restrictions on context, topic, or grammatical form. Our chatbot is designed to simulate a typical conversationalist, so that learners may practise conversing in their target language. To our knowledge, this work is the first to perform user evaluations on open-domain LLM chatbots for language learning.

## 3 Study setup

We recruit 160 participants via Prolific<sup>1</sup> for our user study. All participants are screened to ensure that their first language is not English. They are then directed to our website, where they navigate through 4 sections:

1. The first section consists of basic profiling questions to ascertain the participant’s linguistic background, such as their first language (L1). The most common L1s were Polish, Portuguese, and Italian (full list in the appendix).
2. The second section is a proficiency test consisting of 25 multiple choice questions to estimate their proficiency level. The questions and answers are taken from the Cambridge English Test Your English application<sup>2</sup>. Scores from the test are mapped to the Common European Framework of Reference (CEFR) (Council of Europe, 2020), a 6-point scale representing proficiency, allowing easy comparison with existing work.
3. The third section is the main interaction with the chatbot. This involves chatting directly with the chatbot, or reading messages from chatbots; variations are described below.
4. The final section consists of closing questions asking the participant about their experience, including 2 attention questions to eliminate low-effort responses. We enclose the full

<sup>1</sup><https://www.prolific.com/>

<sup>2</sup><https://www.cambridgeenglish.org/test-your-english/general-english/>



list in the appendix, but highlight individual questions in our Findings section.

Additional details of the user study setup can be found in the appendix.

Each participant is randomly assigned different experimental conditions in a  $2 \times 2 \times 2$  design:

- **Chatting VS reading**

To understand the difference between receptive reading and interactive conversation, we assign half of our participants to the chatting condition, and the remaining half to the reading condition. In the chatting condition, each participant is asked to converse with a chatbot. They send messages to the chatbot directly and can actively steer the conversation topic. In the reading condition, the participant cannot send messages, and instead navigates through a conversation between two identical chatbots. Everything else, such as the user interface, remains the same.

- **Adaptive difficulty VS non-adaptive difficulty**

One common feature in language learning chatbots is the capability of adapting chatbot messages based on the user’s proficiency level. However, it is unclear to us how this may affect the learning experience, so we apply the adaptation for half of the participants, while the other half receive messages generated with standard top- $k$  sampling ( $k = 40$ ) (Fan et al., 2018). For the adaptation, we follow Tyen et al. (2022) and use a reranking method with sub-token penalties and filtering, as described in their paper<sup>3</sup>. See the appendix for further details on the re-ranking model and implementation of penalties.

- **Dictionary lookup VS no dictionary lookup**

In the dictionary lookup condition, participants are able to click on words to look up their definitions. This function is only available for words in messages that are sent from the chatbot. All messages are tokenised by the RASP parser (Briscoe et al., 2006).

Full details can be found in the appendix.

<sup>3</sup>Implementation found at <https://github.com/WHGTYen/ControllableComplexityChatbot>.

For all three pairs of conditions, participants are split evenly into two groups, where one group is assigned one condition and the other group is assigned the other condition: for example, there are 80 participants in the chatting condition and 80 participants in the reading condition as well. The splitting is done in a way that ensures equal coverage across all combinations of conditions: e.g. there are 20 participants who are chatting *and* have adaptive difficulty *and* dictionary lookup; 20 participants who are reading *and* have adaptive difficulty *and* dictionary lookup; and so on.

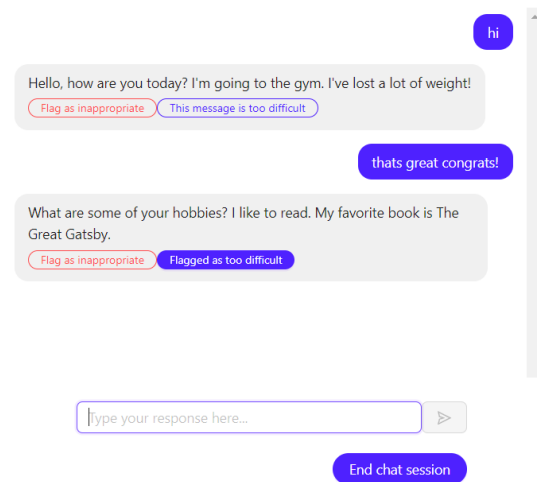


Figure 1: Chat interface presented to participants in the chatting condition. Messages in blue bubbles are sent from the user, while messages in grey bubbles are sent from the chatbot. In this example, the most recent message is flagged by the user as being too difficult.

### 3.1 Chatbot

We use BlenderBot (2.7B parameters) (Roller et al., 2021) as the base LLM. BlenderBot was chosen because the model is not instruction-tuned, and has been fine-tuned on the Blended Skill Talk dataset (Smith et al., 2020), which combines various conversational skills. This allows us to simulate a real conversationalist rather than a virtual assistant. Additionally, BlenderBot was previously used by Tyen et al. (2022) for difficulty adjustment. We use the same setup<sup>3</sup> to enable a clear comparison: for participants in the adaptive condition, we use a decoding method proposed by Tyen et al. (2022) (method 5), which allows us to adjust the difficulty level of generated messages.

In terms of chatbot quality, Roller et al. (2021) report extensive evaluation results on BlenderBot, including self-chat human evaluation and interac-

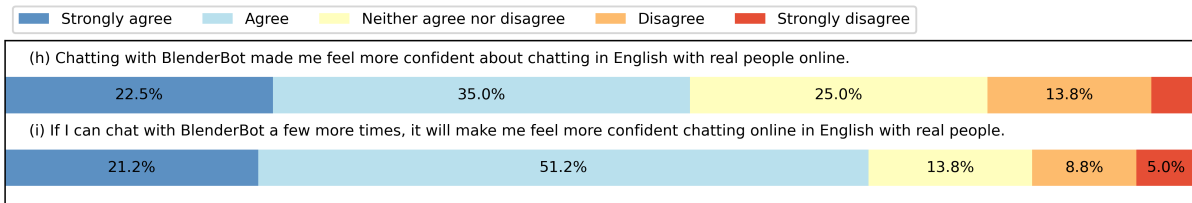


Figure 2: Responses to confidence-related Likert questions from participants in the chatting condition.

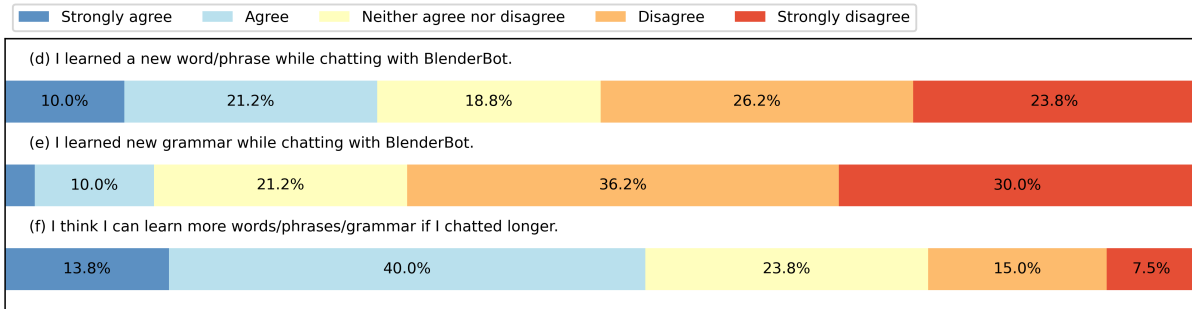


Figure 3: Responses to learning-related Likert questions from participants in the chatting condition.

tive human evaluation. Their results show that generative BlenderBot (2.7B) performs better than Meena Adiwardana et al. (2020) and narrowly loses to human participants in terms of engagingness (49% versus 51%). Tyen et al. (2022) report self-chat evaluation results of the adapted decoding method based on the Sensibleness and Specificity Average Adiwardana et al. (2020) and grammaticality. Method 5 from their paper was found to be statistically equivalent to the non-adapted version in terms of sensibleness, specificity, as well as grammaticality.

Additionally, to disentangle effects of prompt crafting or manual changes to the learning experience, and to minimise effects on chatbot quality, our current chatbot setup does not use any prompts, predetermined responses, or linguistic syllabuses (though they may be added in future work). All user input goes directly to the LLM, and all generated messages are sent directly to the user.

Figure 1 shows a screenshot of the interface used to interact with the chatbot. Participants are asked to spend at least 15 minutes on this section, after which the “End chat session” button would appear. Participants can also choose to spend more time with the chatbot if they wished.

## 4 Findings

### 4.1 RQ1: Impact on real-life interaction

#### Increased self-confidence in real-life interaction

Two of our feedback questions (h) and (i), shown in Figure 2, focus on the learner’s sense of self-confidence when it comes to real-life settings. We rely on self-reports as confidence is inherently about perception of the self, and arguably can only be measured via self-reports (Paulhus et al., 2007).

The results show that more than half of the participants in the chatting condition agree that they felt more confident about chatting with real people, even after 1 session of conversing with the chatbot. This number increases further to 72% in question (i), where we ask participants for *predicted* self-confidence levels, if given more opportunities to converse with the chatbot.

#### Limited learning may increase in the long term

Questions (d), (e), and (f), shown in Figure 3, focus on the learning of new words, phrases, or grammatical constructions. While some participants report learning after just one session, most disagree with the statements, particularly regarding grammatical constructions. This suggests that a single chatbot session is unlikely to provide benefits for language learning.

On the other hand, participants are more optimistic when asked to *predict* learning, if given fur-

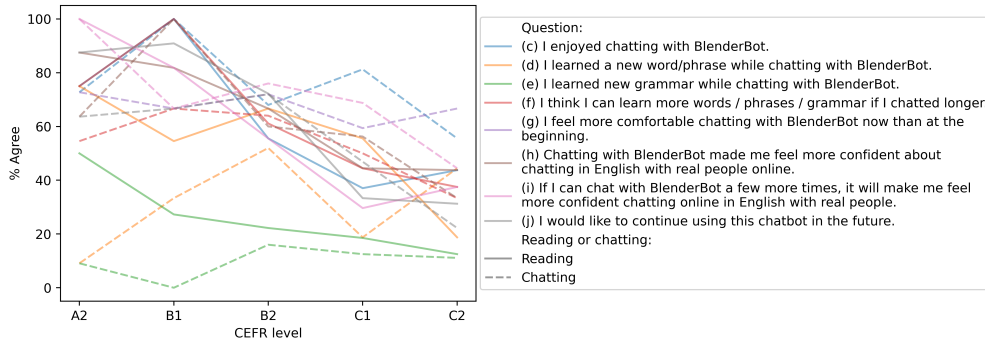


Figure 4: Proportion of *Agree* or *Strongly agree* responses to each Likert question, sorted by CEFR level.

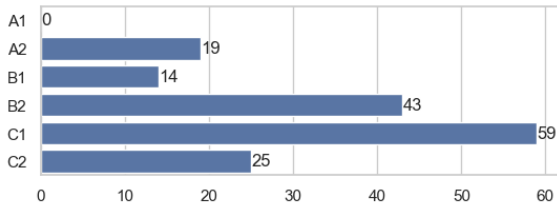


Figure 5: Distribution of CEFR levels across all participants. CEFR levels are ordered from least to most proficient.

ther opportunity to converse with the chatbot. This is in line with our previous finding about confidence, where participants also predict more positive outcomes if given more time with the chatbot. As our user study only consists of one session and is not designed to test longitudinal effects, we are unable to verify whether there are any actual long term benefits. However, it is noteworthy that users themselves have a positive opinion on long-term chatbot usage, suggesting that their experience had a motivational effect.

#### 4.2 RQ2: Variation in proficiency levels

All participants are asked to complete a series of multiple choice questions, which are used to gauge their proficiency level. The distribution of CEFR levels is shown in Figure 5. None of our participants are found to be at A1 (most beginner) level: this is likely due to the initial recruitment and navigation through the consent form, which requires a minimal level of proficiency to understand.

Proportion of *Agree* or *Strongly agree* responses sorted by approximate CEFR level are visualised in Figure 4. Note that *Agree* and *Strongly agree* represent positive outcomes in our Likert questions, while *Disagree* and *Strongly disagree* represent negative outcomes.

We then compute Spearman’s rank correlation

coefficient ( $\rho$ ) between test scores and answers to our Likert questions.

Our results show that **less proficient learners are more likely to report and predict positive outcomes**. We find that participants’ scores in the proficiency test significantly negatively correlate with:

- enjoyment (question (c),  $\rho = -0.25, p < 0.002$ )
- perceived learning of grammatical constructions (question (e),  $\rho = -0.33, p < 0.00003$ )
- predicted learning in the long term (question (f),  $\rho = -0.27, p < 0.0005$ )
- predicted self-confidence levels in the long term (question (i),  $\rho = -0.36, p \ll 0.00001$ )
- interest in continued usage (question (j),  $\rho = -0.37, p \ll 0.00001$ )

Questions (f), (i), and (j) all pertain to participants’ predictions, suggesting that lower proficiency participants find greater potential for future benefits than high-proficiency participants. This is a reasonable outcome as more beginner language learners would require more practice than more experienced learners. For question (e), we hypothesise that the difference between high- and low-proficiency learners is because grammar is often taught at earlier stages of learning. High-proficiency learners are more likely to struggle with advanced concepts such as use of humour and slang, linguistic style, etc.

Note that the above correlation scores are computed for all participants (in the reading and chatting conditions). Figure 4 shows that the effect is stronger for the reading condition than the chatting condition, where learners at a higher proficiency level give more positive responses than in the reading condition, particularly for questions (c) on enjoyment and (i) on predicted confidence.

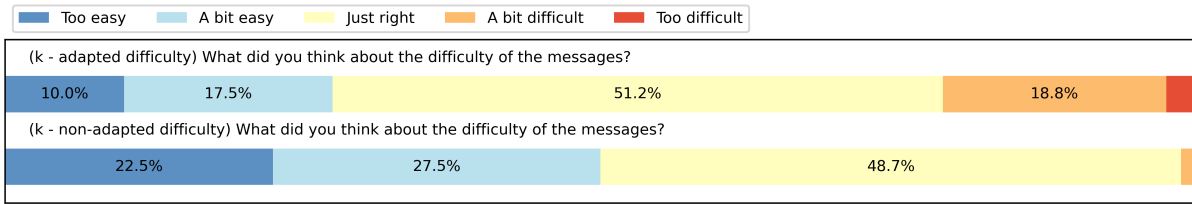


Figure 6: Responses to question (k) on perceived difficulty from participants in the adapted and non-adapted difficulty conditions.

Overall, our results suggest that learners at a lower proficiency level are more likely to benefit from interactions with an LLM chatbot, but correlations are not strong and many high-proficiency learners also report positive outcomes.

### 4.3 RQ3: Difficulty adaptation

For half of the participants, the chatbots are adjusted to their CEFR level (Tyen et al., 2022) based on their scores in the pre-test. For the remaining half, the chatbots use standard top- $k$  sampling (Fan et al., 2018). At the end of the study, participants are asked about the difficulty level of the messages in question (k), where the potential responses are: *Too easy*, *A bit easy*, *Just right*, *A bit difficult*, and *Too difficult*.

Firstly, our results show that there is a significant difference in perceived difficulty between those in the adapted condition and those in the non-adapted condition ( $p < 0.00009$ ). When comparing specific responses, we find that participants in the adapted version are **significantly more likely to respond with *A bit difficult*** ( $p < 0.0001$ ), while the number of responses for *Too difficult* remain the same, and there are non-significant reductions in the number of *Too easy* and *Easy* responses. Figure 6 contains a visualisation of the responses.

The fact that the non-adapted version of the chatbot is *Too easy* for many participants is in line with the finding in Tyen et al. (2022) that BlenderBot with no adaptations generates messages at B1 level. If the default difficulty level is B1, many participants at B2 level or above would consider the messages to be too easy. Therefore, difficulty adjustment methods are required.

Our results indicate that difficulty adjustment via decoding (Tyen et al., 2022) is effective at introducing language aspects which are more difficult, but are not so difficult that the learner is unable to comprehend it. According to Krashen’s Input Hypothesis of second language acquisition

(Krashen, 1992), successful second language acquisition occurs when the learner is exposed to input that contains ‘ $i + 1$ ’, referring to “an aspect of language that the acquirer has not yet acquired but that he or she is ready to acquire”. This suggests that the ideal perceived difficulty level is between *Just right* and *A bit difficult*. Following this hypothesis, we surmise that exposure to text with adjusted difficulty levels is likely more beneficial for second language learning than to text that is not adjusted. However, to fully test this theory, a longitudinal study is required to measure learning progress.

### 4.4 RQ4: Conversational interaction versus receptive reading

In both the chatting condition and reading condition, messages from the chatbot(s) are generated on-the-fly using the same decoding strategy. Despite using the same setup, we observe distinct linguistic differences between the content generated in the chatting and reading conditions, likely due to influence from the user. For example, messages generated in the reading condition are shorter on average ( $p < 0.0002$ ); messages in the chatting condition are more likely to contain questions ( $p < 0.00001$ ).

Overall, the Jaccard similarity between chatbot-generated messages in the chatting and reading conditions is relatively high at 0.35. For comparison, the Jaccard similarity between all chatbot-generated messages and messages in the Blended Skill Talk dataset (Smith et al., 2020) (which BlenderBot was fine-tuned on) is 0.26; and the Jaccard similarity between chatbot-generated messages and user-written messages in the user study is 0.12.

We additionally explore the impact of reading versus chatting via survey responses. Surprisingly, our results show only one main difference between learners in the chatting and reading conditions: **chatters enjoy the experience more than read-**

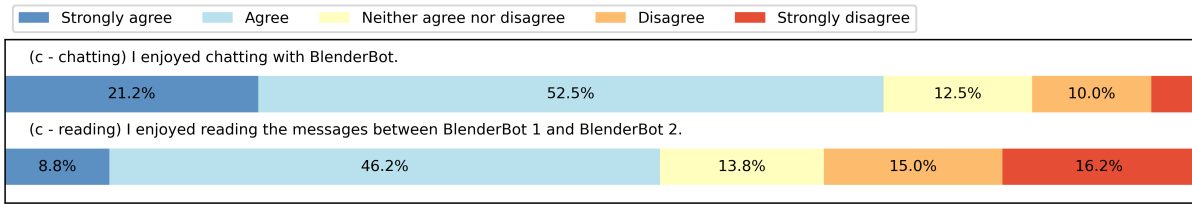


Figure 7: Responses to Likert question on enjoyment from participants in the chatting and reading conditions.

**ers do.** Figure 7 below shows a comparison of their responses to question (c), which asks whether participants enjoyed the chatbot session. Chatters are significantly more likely to give a more positive response ( $p < 0.001$ ).

Among all survey question responses, other than enjoyment, we find no other significant differences between the chatting and reading conditions, whether with adaptive or non-adaptive difficulty, or with or without dictionary lookup. This is a surprising result given the differences in text content, and the fact that second language production is inherently differently from second language comprehension (Laufer, 1998; Gernsbacher and Kaschak, 2003).

There are some suggestive, but non-significant differences: for example, users in the chatting condition are slightly more likely to predict boosts in confidence levels, while users in the reading condition are slightly more likely to report learning new words. However, further study with a larger group of users is required to understand if these effects are linked to interaction (or lack thereof).

## 5 Clicking for dictionary lookup as an indicator of complexity

In our user study, we implement a clicking mechanism where learners can click on words to reveal their dictionary definition. This function is simple to implement and integrates seamlessly with the existing user interface, yet can provide valuable information about the user’s learning experience.

We find that clicks are a strong indicator of when a learner finds a word difficult. We report in Table 1 three statistics that are often correlated with lexical complexity (Shardlow et al., 2021), and compare them for words that are clicked on versus words that are *not* clicked on. We find that words that are clicked on are more complex, as they are significantly longer ( $p << 0.0001$ ), less frequent ( $p << 0.0001$ ), and have a smaller number of definitions ( $p < 0.0002$ ).

Statistic	Clicked	Unclicked
Avg. character length	<b>8.07</b>	3.80
Avg. Zipf frequency	<b>5.69</b>	6.82
Avg. num. of definitions	<b>2.59</b>	5.26

Table 1: Statistics correlated with lexical complexity for words that are clicked on, versus words that are not clicked on. Zipf frequency refers to the base-10 logarithm of frequency per 1 billion words; the number of definitions refers to the number of synsets on WordNet (Miller, 1994). Bold font denotes the statistic that indicates higher complexity. All 3 statistics are shown to be significantly different between clicked and unclicked words ( $p << 0.0001$  for length and frequency;  $p < 0.0002$  for number of definitions).

Furthermore, clicks are also associated with the reported difficulty level of the overall message. During our study, participants are able to flag messages that they consider to be too difficult (see Figure 1). We find that messages that are flagged as difficult are 5 times more likely to have words that are clicked on (11.3%), compared to messages that are not flagged (2.2%). This demonstrates that learners are clicking on words that *they* consider complex, rather than e.g. out of curiosity, or due to random, unintentional clicking.

Despite strong evidence that clicks are indicative of lexical complexity, we observe that only 33 out of 80 participants in the clicking condition make use of this feature. For the 33 participants, 4751 messages are sent from the chatbot, but only 377 clicks are recorded in total. Possible reasons for the low click-rate include: 1) Participants rarely encounter any words that they find sufficiently difficult; 2) Participants are engaged in conversation and prefer to continue rather than pausing to read definitions; 3) Participants find the dictionary definitions unhelpful; or 4) Participants forget they have access to this function. Note that all participants in the clicking condition are informed of this mechanism before their chatbot session.

Due to the low click-rate, our data is insufficient to draw conclusions about potential benefits

or drawbacks of clicking. Additionally, we find no significant differences in survey response questions between the groups with and without this dictionary lookup function. This is also the case when looking at groups with or without adaptive difficulty, or in the chatting or reading conditions. Further work is required to understand clicking behaviour and its impact on the learning experience.

## 6 Limitations and future work

**Scope of user study** Our user study involves a small sample of 160 participants, whose first languages are mostly European languages, and whose CEFR proficiency levels are skewed towards the higher end. Additionally, due to the small number of participants, we are unable to properly measure interaction effects despite the  $2 \times 2 \times 2$  design. Further work is required to ascertain if our findings hold at a larger scale and with a different population, and to clarify how LLM chatbots facilitate language learning.

**Measured performance** Some of our observations rely on participants' self reports rather than measured linguistic performance. Based on previous research, our results show promise and are likely associated with improved performance, but our study does not measure this directly. In future work, we can measure linguistic improvement over the course of multiple chatbot sessions by comparing performance before and after the fact.

**LLM capability** For our user study, we use a small (2.7B parameters) model for the ease of deployment and inference speed. It is possible to improve the capability of the chatbot by replacing it with larger models such as LLaMA (Touvron et al., 2023) and BLOOM (BigScience Workshop et al., 2022). We expect that results related to enjoyment are likely to improve with a larger model, and the conversational experience would be more realistic.

**Personalisation using clicking data** Our current study does not make use of the clicking data to adjust the generated messages, but future work on computer-assisted language learning can make use of clicks to adapt content on-the-fly to the user.

## 7 Conclusion

In this paper, we report our findings from our user study, where we recruit 160 second lan-

guage speakers to interact with LLM-based chatbots. Our results show that using an LLM chatbot as a language practice tool can improve self-confidence, and provides a more enjoyable learning experience compared to purely receptive reading tasks. Although learning outcomes are not apparent after one session, many participants predict more positive effects in the long term, if given further opportunity to interact with the chatbot. This is especially true for learners at a lower proficiency level.

In terms of implementation, we introduce clicking as a way to reveal dictionary definitions during the user study. We find that this method effectively detects words which the learner finds complex, on-the-fly. For the chatbot, we implement a decoding method that adjusts the difficulty of generated messages (Tyen et al., 2022). Our results show that this method generates text that is more often considered *A bit difficult*, which is likely to facilitate learning (Krashen, 1992).

Overall, our findings demonstrate that LLM chatbots as a language practice tool can bring benefits to different aspects of language learning. We leave it to future work to measure long-term learning outcomes of chatbot interaction.

## Acknowledgements

This paper reports on research supported by Cambridge University Press & Assessment. We thank the anonymous reviewers for their comments.

## References

- Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. [Towards a human-like open-domain chatbot](#).
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Serge Bibauw, Thomas François, and Piet Desmet. 2019. Discussing with a computer to practice a foreign language: research synthesis and conceptual framework of dialogue-based CALL. *Computer Assisted Language Learning*, 32(8):827–877.
- BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel

- Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Ted Briscoe, John Carroll, and Rebecca Watson. 2006. [The second release of the RASP system](#). In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 77–80, Sydney, Australia. Association for Computational Linguistics.
- David Coniam. 2014. The linguistic accuracy of chatbots: usability from an esl perspective. *Text & Talk*, 34(5):545–567.
- Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching Assessment*, 3rd edition. StrasBourg.
- Duolingo Team. 2023. [Introducing duolingo max, a learning experience powered by gpt-4](#). Accessed on January 31, 2024.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Morton Ann Gernsbacher and Michael P. Kaschak. 2003. [Neuroimaging studies of language production and comprehension](#). *Annual Review of Psychology*, 54(1):91–114.
- Kaveh Hassani, Ali Nahvi, and Ali Ahmadi. 2016. Design and implementation of an intelligent virtual environment for improving speaking and listening skills. *Interactive Learning Environments*, 24(1):252–271.
- Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. [Ai alignment: A comprehensive survey](#). *arXiv preprint arXiv:2310.19852*.
- Jiyoun Jia. 2009. An ai framework to teach english as a foreign language: Csiec. *Ai Magazine*, 30(2):59–59.
- Stephen Krashen. 1992. The input hypothesis: An update. *Linguistics and language pedagogy: The state of the art*, pages 409–431.
- Batia Laufer. 1998. [The Development of Passive and Active Vocabulary in a Second Language: Same or Different?](#) *Applied Linguistics*, 19(2):255–271.
- Seungjun Lee, Yoonna Jang, Chanjun Park, Jungseob Lee, Jaehyung Seo, Hyeonseok Moon, Sugyeong Eo, Seunghoon Lee, Bernardo Yahya, and Heuseok Lim. 2023. [PEEP-talk: A situational dialogue-based chatbot for English education](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 190–207, Toronto, Canada. Association for Computational Linguistics.
- Chun-Hung Lu, Guey-Fa Chiou, Min-Yuh Day, Chong-Shyong Ong, and Wen-Lian Hsu. 2006. Using instant messaging to provide an intelligent learning environment. In *Intelligent Tutoring Systems*, pages 575–583, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Microsoft. 2023. [How ChatGPT can be used to help with foreign language learning](#). Accessed on February 1, 2024.
- George A. Miller. 1994. [WordNet: A lexical database for English](#). In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.
- Neasa Ní Chiaráin and Ailbhe Ní Chasaide. 2016. [The digichaint interactive game as a virtual learning environment for irish](#). In *CALL communities and culture – short papers from EUROCALL 2016*, pages 330–336. Research-publishing.net.
- OpenAI. 2023. [ChatGPT](#). Accessed on February 1, 2024.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Alexandros Papangelis, Paweł Budzianowski, Bing Liu, Elnaz Nouri, Abhinav Rastogi, and Yun-Nung Chen, editors. 2021. *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*. Association for Computational Linguistics, Online.
- Delroy L Paulhus, Simine Vazire, et al. 2007. The self-report method. *Handbook of research methods in personality psychology*, 1(2007):224–239.
- PrettyPolly. 2023. [Prettypolly - learn a language by practicing speaking with ai](#). Accessed on January 31, 2024.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. [Recipes for building an open-domain chatbot](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.
- Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. [SemEval-2021 task 1: Lexical complexity prediction](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. [Can you put it all together: Evaluating conversational agents’ ability to blend skills](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Pei-Hao Su, Chuan-Hsun Wu, and Lin-Shan Lee. 2015. A recursive dialogue game for personalized computer-aided pronunciation training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):127–141.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Gladys Tyen, Mark Brenchley, Andrew Caines, and Paula Buttery. 2022. [Towards an open-domain chatbot for language practice](#). In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 234–249, Seattle, Washington. Association for Computational Linguistics.

P. Vlugter, A. Knott, J. McDonald, and C. Hall. 2009. [Dialogue-based CALL: a case study on teaching pronouns](#). *Computer Assisted Language Learning*, 22(2):115–131.

Jiashuo Wang, Haozhao Wang, Shichao Sun, and Wenjie Li. 2023. Aligning language models with human preferences via a bayesian approach. *arXiv preprint arXiv:2310.05782*.

Zhihui Zhang and Xiaomeng Huang. 2024. The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon*, 10(3).

## A Study setup details

**Screening** Our participants are recruited from Prolific and filtered using the built-in screening process. Participants must have a non-English language for their first language, primary language, and earliest language in life. As this does not guarantee that each participants’ first language is *not* English (one can have multiple first languages), we also ask for their first languages later in the study. Additionally, we filter out participants living in countries where English speakers are in the majority (e.g. US, UK, Australia, etc.).

All participants’ first languages can be found in Table 2.

First language	Number of participants
Polish	60
Portuguese	32
Italian	17
Greek	11
Spanish	11
Hungarian	8
German	7
Russian	3
Czech	3
Slovene	2
Afrikaans	2
Latvian	1
French	1
Arabic	1
Romanian, Moldovan	1
Urdu	1
Dutch	1
Turkish	1
Tagalog	1
Ukrainian	1

Table 2: All first languages among our participants. Note that each participant can specify more than one first language.

**Payment** Before the study begins, participants are told that they will be paid a minimum of £7 for roughly half an hour of their time, including at least 15 minutes of chatbot interaction. Pay will increase with every additional 15 minutes spent with the chatbot(s), up to a maximum of £13. All entries are manually verified before payment to remove low-effort or invalid entries.

**Consent form** Participants are redirected to our website for the study, where they are presented with a consent form detailing how their data will be used. The consent form was written with second language speakers in mind, to ensure that beginner learners can also understand it. Participants can also contact the authors via email or the messaging system on Prolific regarding any concerns about the study. To proceed to the next section, participants must consent to their data being used for research purposes. However, they can withdraw their consent at any point, up to 6 months after the study. They may also exit the task any time they wished.

**Profiling questions** There are two questions in this section:

1. *What is/are your first language(s)?*



Participants can select one or more languages out of a list of ISO-639 languages.

## 2. *How long have you been learning English?*

Participants enter a number followed by a choice of “years” or “months”.

**Proficiency questions** 25 multiple choice questions were used to estimate the proficiency level of users. Questions are taken from the Cambridge English Test Your English application (General English)<sup>4</sup>. Participants are asked to select one of 3 or 4 options for each question. Scores are then converted to CEFR levels, as done on the website. This CEFR level is used as input to the difficulty adaptation mechanism (Tyen et al., 2022).

**Chatbot interaction** At the beginning of this section, participants are informed that:

1. They should not reveal any personal information, even if asked.
2. The chatbots are not real people, despite what the messages may say, but messages will be read by researchers afterwards.
3. There is a risk that the chatbots may generate inappropriate messages. Participants can flag messages as inappropriate by clicking on the ‘Flag as inappropriate’ button. Clicking on the button again un-flags the message.
4. Information or opinions in the generated messages should not be taken for fact.
5. If participants are finding the messages difficult, they can flag messages as too difficult by clicking on the ‘Flag as too difficult’ button. Clicking on the button again un-flags the message.
6. There will be attention questions in the next section, so participants should read messages carefully.
7. (For those in the dictionary lookup condition) Participants can click on words to look them up in the dictionary.

After acknowledging the above, participants may begin the chatbot interaction. In both reading and chatting conditions, messages are generated on-the-fly, using an NVIDIA Tesla V100 GPU.

<sup>4</sup><https://www.cambridgeenglish.org/tes-t-your-english/general-english/>

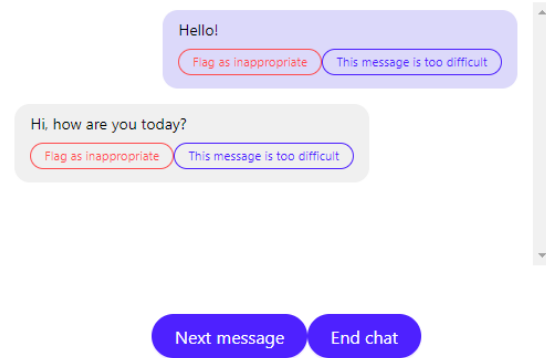


Figure 8: Interface presented to participants in the reading condition. Messages on both sides are chatbot-generated using the same parameters.

**Reading condition** In the reading condition, the user reads a conversation between two identical chatbots with the same settings. The user interface can be found in Figure 8. Unlike the UI for the chatting condition (in Figure 1), the user presses a button to reveal the next message, instead of typing in a text input field. Note that to maintain fair comparison, all messages in either the reading or chatting conditions are generated in real time.

**Adaptive condition** In the adaptive condition, all chatbot messages are generated using a weighted reranking decoding method (Tyen et al., 2022). This method consists of 3 components:

1. Sub-token penalties to adjust probabilities of tokens during generation
2. A reranker model to assign adjusted scores to each generated candidate message
3. A filter to remove generated candidates that contain ungrammatical words

For the reranker model, we use weights directly from [https://github.com/WHGTyen/ControllableComplexityChatbot/tree/master/complexity\\_model](https://github.com/WHGTyen/ControllableComplexityChatbot/tree/master/complexity_model) without performing any additional fine-tuning. The final score of each generated candidate is calculated as the average rank between ranked probability scores and ranked complexity scores, weighting both equally:

$$\frac{r(P(C)) + r(|L_{\text{user}} - LC|)}{2} \quad (1)$$

$C$  is the candidate message;  $r$  is a ranking function returning a rank out of 20 candidates;  $L_{\text{user}}$  is the

CEFR level of the user, and  $L_C$  is the predicted CEFR level of the candidate message.

For the vocabulary filter, we use a list of English words from <https://github.com/dwyl/english-words>, but ignore capitalized words (indicating proper nouns). For the sub-token penalties, the probability of each token  $t$  is given by:

$$P(t) = \begin{cases} P(t) \cdot \varphi(L_t - L_{\text{user}}) & \text{if } L_t > L_{\text{user}} \\ P(t) & \text{otherwise} \end{cases} \quad (2)$$

where  $L_t$  refers to the CEFR level of token  $t$  and  $L_{\text{user}}$  refers to the user’s CEFR level, determined by proficiency test scores at the beginning of the user study. The level is determined before any text is generated, does not change throughout the conversation, and is implemented in the same way regardless of reading/chatting or lookup conditions. For the function  $\varphi$  representing the normal distribution, we follow parameters used in the original paper,  $\mu = 0$  and  $\sigma = 2$ .

**Inappropriate language** Participants have the ability to flag messages as being inappropriate. Of the 21,283 messages sent by a chatbot, 359 (1.69%) were flagged as such. A small sample reveals that about half of these messages were flagged due to being nonsensical, or logically or pragmatically unsuitable for the context, rather than offensive – this may be due to some participants misinterpreting the word “inappropriate”. The remaining half generally touch on politically sensitive topics, use politically incorrect terms, or are offensive or insulting in some way.

The existence of these messages is concerning for chatbot usage in educational settings, especially for younger learners. Recent work on AI alignment has produced considerable improvements over the past few years (see Ji et al. (2023) for a comprehensive survey), but it is still possible to elicit inappropriate messages, especially when under specially crafted attacks (Shayegani et al., 2023). In its current form, we believe that LLM chatbots are best suited for an adult audience who are aware and informed of the nature of language models. However, current technology on LLM safety is improving rapidly, and new methods for mitigating toxicity are being developed constantly (e.g. Ouyang et al. (2022); Bai et al. (2022); Wang et al. (2023)), so it may soon be possible to deploy chatbots that are safe for younger audiences.

**Feedback questions** Table 3 shows the full list of questions asked after each chatbot session. Questions vary slightly depending on whether the participant is assigned the chatting or reading condition.

Questions (a) and (b) are attention questions used to eliminate low-effort entries where the participant failed to engage with the task. Among our submissions, only 4 are removed for this reason.

Chatting condition	Reading condition
<b>Attention questions</b>	
(a) Were there messages from BlenderBot that did not make sense? If so, can you give some examples?	Were there messages from BlenderBot 1 or 2 that did not make sense? If so, can you give some examples?
(b) Tell us one fact about BlenderBot that you learned from this conversation.	Tell us one fact about either BlenderBot 1 or 2 that you learned from this conversation.
<b>Likert questions</b>	
(c) I enjoyed chatting with BlenderBot.	I enjoyed reading the messages between BlenderBot 1 and BlenderBot 2.
(d) I learned a new word/phrase while chatting with BlenderBot.	I learned a new word/phrase while reading these messages.
(e) I learned new grammar while chatting with BlenderBot.	I learned new grammar while reading these messages.
(f) I think I can learn more words / phrases / grammar if I chatted longer.	I think I can learn more words / phrases / grammar if I read more of these messages.
(g) I feel more comfortable chatting with BlenderBot now than at the beginning.	N/A
(h) Chatting with BlenderBot made me feel more confident about chatting in English with real people online.	Reading these messages made me feel more confident about chatting in English with real people online.
(i) If I can chat with BlenderBot a few more times, it will make me feel more confident chatting online in English with real people.	If I can read more of these messages, it will make me feel more confident chatting online in English with real people.
(j) I would like to continue using this chatbot in the future.	I would like to continue reading similar messages in the future.
<b>Feedback questions</b>	
(k) What did you think about the difficulty of the messages? Options: Too easy / A bit easy / Just right / A bit difficult / Too difficult	
(l) Do you have any other thoughts, comments, or feedback for us? (free text response)	

Table 3: Questions answered by each participant after their chatbot session.

# Sailing through multiword expression identification with Wiktionary and Linguse: A case study of language learning

Till Überrück-Fries and Agata Savary

Université Paris-Saclay

CNRS, LISN

till@ueberfries.de

agata.savary@lisn.upsaclay.fr

Agnieszka Dryjańska

University of Warsaw

Institute of Romance Studies

a.dryjanska@uw.edu.pl

## Abstract

Multiword expressions (MWEs), due to their idiomatic nature, pose particular challenges in comprehension tasks and vocabulary acquisition for language learners. Current NLP tools fall short of comprehensively aiding language learners when encountering MWEs. While proficient in identifying MWEs seen during training, current systems are constrained by limited training data. To address the specific needs of language learners, this research integrates expansive MWE lexicons and NLP methodologies as championed by Savary et al. (2019a). Outcomes encompass a specialized MWE corpus from Wiktionary, the enhancement of Linguse, a reading application for language learners, with MWE annotations, and empirical validation with French language students. The culmination is an MWE identifier optimally designed for language learner requirements.

## 1 Introduction

Second language acquisition is a complex process that involves developing and refining a range of competences. One such competence—lexical competence—includes the knowledge of and ability to use a certain category of lexical items, known in the field of Natural Language Processing (NLP) as multiword expressions (MWEs). Examples of such items are *all of a sudden* ‘suddenly’, *a hot dog* ‘a sausage sandwich’, *larger than life* ‘attracting attention’, *to carry out* ‘to perform’ or *to do one’s best*. In language teaching, this category is often referred to as “fixed expressions,” which consist of multiple words learned as cohesive units (Council of Europe, 2001, p. 110). Despite the differing terminologies across computational and educational spheres, the essence of these lexical items remains consistent: they pose distinct idiomatic challenges

that resist straightforward grammatical or semantic interpretation.

The concept of MWEs, defined by Baldwin and Kim (2010), encompasses lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic, and/or statistical idiomaticity. It is precisely this idiomaticity that makes MWEs a notable stumbling block for language learners and a significant computational challenge in NLP.

Given these complexities, there is a compelling need for computer-assisted language learning solutions that address the acquisition of such lexical items. We address this need by focusing on the integration of MWE identification techniques into Linguse, a reading application designed for language learners. The aim is to bridge the gap between the pedagogical requirements of second language learners and the capabilities of state-of-the-art NLP systems.

## 2 Related work

Challenges encountered when processing MWEs include ambiguity, idiomaticity, flexibility, and lexical proliferation (Sag et al., 2002). Two main tasks in this context are: MWE discovery and MWE identification. Discovery aims to find new MWEs in text corpora, while identification deals with annotating known MWEs in running text (Constant et al., 2017). Our focus is on MWE identification, as it allows MWEs to be cross-referenced with lexical resources, which is crucial in language learning.

Traditional approaches to MWE processing included treating them as ‘words with spaces’ (Smadja, 1993; Evert, 2005) but one category has proven particularly resistant to this treatment: verbal multiword expressions (VMWEs). They exhibit non-adjacency of components (*spend a lot of time*), syntactic and word order variability (*time spent*), and syntactic ambiguity (*turn on the heating* vs. *turn on the floor*) (Savary et al., 2017).

---

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

To address these challenges, the PARSEME network created standardized corpora with VMWE annotations in 26 languages and organized a series of multilingual shared tasks on automatic identification of VMWEs (Savary et al., 2017; Ramisch et al., 2018, 2020). The focus in evaluation gradually moved from generic performance measures to those focusing on previously unseen VMWEs (Ramisch et al., 2020), which proved critically hard to identify. Savary et al. (2019a) claim that this difficulty stems from the inherent nature of VMWEs' idiosyncrasy, which resists generalisation over unseen data.

However, much progress can still be achieved in identification of seen VMWEs, by addressing their morpho-syntactic flexibility, as shown by Pasquer et al. (2020b) with the *Seen2020* system, underpinned by rule-based candidate extraction and filtering techniques. In edition 1.1 of the shared task it yielded a macro-average  $F_1$  score of 0.83, surpassing four other systems in the identification of seen VMWEs. In edition 1.2 it was rebranded as *Seen2Seen* for the closed track (in which only the annotated corpora provided by the shared task organizers are used) and as *Seen2Unseen* with some modifications for the open track (in which other external resources can also be used) (Pasquer et al., 2020a). It shows limited performance in the unseen MWE-based category (4<sup>th</sup>/7,  $F_1$ : 13.7) but remains competitive in the global (seen+unseen) MWE-based category (1<sup>st</sup>/2 in the closed track, 2<sup>nd</sup>/7 overall,  $F_1$ : 63.0). It was only outperformed by one of the neural models (Taslimipour et al., 2020), employing a fine-tuned, multilingual BERT (Devlin et al., 2019) for joint parsing and identification.

These outcomes suggest that rule-based systems like *Seen2020* can be highly competitive for seen MWE identification, even when juxtaposed with more sophisticated models. Their principal limitation is the relatively low diversity of MWEs seen during training. This problem might be tackled by using fully unsupervised methods, e.g. inspired by metaphor detection, in which contextual and static word embeddings are used to represent the idiomatic and literal meaning of a potential MWE, respectively (Zeng and Bhat, 2021).

Another solution is to leverage existing MWE lexicons, which possibly contain many MWEs not seen in manually annotated corpora. Namely, the lexicon entries known to be MWEs, can be automat-

ically identified in large corpora with a relatively high reliability. This is due to the fact that, although VMWEs are potentially ambiguous (*take the cake* can be understood idiomatically or literally), they seldom appear in their literal or accidental forms in corpora (Savary et al., 2019b). Thus, the sentences containing lexicon entries can be used as an augmented training corpus, as shown by Kanclerz and Piasecki (2022) for English and by Hadj Mohamed et al. (2024) for Arabic. Sentences illustrating the usage of an MWE can also be found in the lexicon itself, as is the case for Wiktionary (Muzny and Zettlemoyer, 2013), and leveraged for MWE identification (Tedeschi et al., 2022). Importantly for our work, such methods enable linking the identified MWEs with human-readable definitions, useful for language learners. They also facilitate the control over the precise list of identifiable MWEs. This might pave the way towards adapting MWE identification to the learners' proficiency level.

### 3 Didactic framework

As the purpose of this study is the integration of automatic MWEs identification and annotation in teaching French as a foreign language, an exclusive focus on technical solutions may fall short of meeting the diverse needs of language learners. To remedy this shortcoming, it is essential to integrate a didactic framework, strongly inspired by linguistic approaches.

In contemporary linguistics not only single word forms but also MWEs are considered an essential component of language, particularly of its lexical subsystem (Mejri, 1999; Sułkowska, 2013; Tutin, 2018). A profoundly modified understanding of the concept of *meaning* in linguistics, strongly impacted by cognitive science and the renewal of semantics, revealed that not only single words but also some syntactically complex items should be perceived as fully fledged *units of meaning*. Consequently, in Foreign Language Teaching (FLT), MWEs, referred to as fixed expressions, are introduced within the framework of communicative language competences, notably lexical and semantic ones (Council of Europe, 2001, pp. 108–109). However, these expressions represent a major issue in both fields owing to syntactic constraints, including degree of combinatorial fixity and discontinuity, and semantic features such as non-compositionality vs. opacity and their gradation (Cavalla, 2016; Tutin, 2018). This complexity gives rise to a par-

ticularly broad category of phenomena encompassing sentential formulae, phrasal idioms, and fixed frames (Council of Europe, 2001, pp. 110–111) that a language learner must internalize as whole *units of meaning* to effectively communicate, which leads to difficulties in both receptive and productive activities (Cavalla, 2009; Cavalla and Labre, 2019). Moreover, the didactic approaches present in French student’s books (e.g. the Edito series) hardly help learners to cope with these difficulties as not sufficient attention is paid to the multi-stage procedure of teaching new lexical or grammatical items (Puren, 2016), especially at the conceptualization and training levels (Dryjańska, 2024).

Two main lexical approaches in FLT can be distinguished: incidental (Fr. incident) and explicit (Fr. explicite). The former subordinates lexical learning to the objectives of reading or writing activities whereas the latter implies a structured lexical progress based on lexical exercises and the appropriation of metalexical concepts (Grossmann, 2011). The incidental lexical approach has much in common with the concept of *synthetic reading* (Fr. lecture synthétique), a linear reading process that aids the introduction of new language structures and simultaneously encourages a focus on the text as a whole, satisfying learners’ curiosity, enriching their experience and helping them to develop their personality (Cornea (2010). Grossmann (2012), when exploring the role of lexical competence in the reading process from a cognitive perspective, observes that it is based on the reader’s ability to match encountered lexical units with representations, such as mental images, and to integrate them into their evolving mental model.

In our project we combine the above lexical and reading approaches. The automatic identification and annotation of MWEs developed within its framework is supposed to foster the process of the acquisition of new fixed expressions while reading independently, which additionally contributes to the development of some general competences such as the ability to learn (Council of Europe, 2001, p. 101, 106). However, it should be noted that the didactic framework seems to impose some specific constraints on MWE identification and annotation regarding evaluation in terms of the metrics like *precision*, *recall* and *F1 score* (cf. Section 7). Although there is an obvious tendency to increase the recall of the process, if it is followed by a diminution of the precision, on account of a

higher number of erroneously identified fixed expressions, the quality of such a tool will be poorly assessed according to teaching objectives. Low precision risks injecting noise and confusion into the learning environment. While these metrics offer insights into the efficacy of MWE identification systems, a genuinely holistic assessment can only be achieved when integrated within broader learning tools and measured against the actual benefits conferred upon the learner. Therefore, we introduce Linguse (cf. Section 8), a tool dedicated to language learning through reading, which encompasses MWE identification as one of its original features.

## 4 Assumption and hypothesis

The ambition of our work is to connect the domains of NLP and language learning by supporting learning activities with MWE identification. A secondary aim is to receive downstream feedback from end users, and connect them in this way to ongoing research on MWEs. In doing so, we seek to reconcile the practical needs of language learning with the theoretical work in NLP.

Inspired by the two preceding sections, we make the following assumption:

Assumption: A large MWE coverage is desirable when automatically annotating text for language learners. This ensures its utility to learners in various stages of language mastery and equips them with the linguistic flexibility they need in real-world scenarios.

This assumption motivated our preference for a large MWE lexicon offering example sentences even for rare expressions, as discussed in the following sections. Grounded in the assumption, our research posits the following hypothesis:

Hypothesis: A rule-based system, trained on example sentences from a lexicon, can successfully extend MWE coverage while maintaining satisfactory performance metrics.

The following sections describe the practical approach taken to corroborate this hypothesis.

## 5 Data

This section describes the MWE material employed in this project, outlining the various sources of MWE data and the construction of a lexicon-driven MWE corpus.

## 5.1 Sources of MWE data

Alongside the theoretical work on aligning notions of MWEs, various data sources on French MWEs were reviewed from both the NLP and the language learning domains. The goal is to identify sources suitable for direct evaluation and those that shed light on MWEs relevant to language learners. Unfortunately, traditional language learning resources like textbooks often lack explicit MWE data, making them less adequate for systematic identification of MWEs relevant to education. Additionally, copyright constraints prevent their exploitation.

Despite this, four supplementary data sources were identified, two from the realm of language learning—FLELex and PolylexFLE—and two from the field of NLP—PARSEME and Deep-Sequoia.

**FLELex:** A graded lexicon for learners of Français Langue Etrangère (FLE) (François et al., 2014). It offers normalized word frequencies by CEFR competence level and includes MWEs<sup>1</sup>.

**PolylexFLE:** Tailored to MWEs in French and aiming to facilitate second language acquisition (Todirascu et al., 2024). It contains 4,525 MWEs and their CEFR competence levels and focuses on verbal MWEs<sup>2</sup>.

**PARSEME 1.2:** An NLP corpus for French, mainly annotated for VMWEs (Ramisch et al., 2020). It comprises 20,961 manually annotated sentences<sup>3</sup>.

**Deep-Sequoia:** Providing multi-layer annotations on French sentences (Candito et al., 2017). Its 3,099 sentences overlap with the French PARSEME corpus but extend MWE annotations beyond VMWEs<sup>4</sup>.

All datasets exhibit a relatively low count of unique MWEs, as summarized in Table 1. For standardization of the counts, MWEs sharing the same multiset of lemmas were considered duplicates. MWE headwords in FLELex and PolylexFLE were

<sup>1</sup>The dataset comes along in two versions and only the CRF-tagged version contains MWE data. The levels are A1, A2, B1, B2, C1 and C2 according to the Common European Framework of Reference for Languages. See <https://ceantal.uclouvain.be/cefrlex/flelex/>.

<sup>2</sup>During the execution of our project, the data was not yet publicly available but a sample of 136 MWEs was graciously provided to our consideration. The full dataset can now be accessed at <https://github.com/amaliatodirascu/PolylexFLE>.

<sup>3</sup><https://gitlab.com/parseme/sharedtas-data/-/tree/master/1.2/FR>

<sup>4</sup><https://deep-sequoia.inria.fr/>

Table 1: Unique MWE Counts Across Datasets

DATASET	# MWEs (UNIQUE)
FLELEX	1,979
POLYLEXFLE	4,525
PARSEME 1.2	1,800
DEEP-SEQUOIA	2,109

automatically tokenized and lemmatized, while PARSEME 1.2 and Deep-Sequoia employed original lemmas.

## 5.2 Extracting structured data from Wiktionary

To address the scarcity of unique MWEs in existing datasets, we created a lexicon-based training corpus. The choice of lexicon required careful consideration, and the **Wiktionary Project**<sup>5</sup> emerged as an ideal candidate. It offers an open, community-driven platform under a Creative Commons Share-Alike license, ensuring both accessibility and adaptability for research applications.

Beyond these merits, Wiktionary provides data for multiple languages, facilitating the future scalability of our methodology. It also supplies example sentences and supplementary linguistic information, both crucial for building an MWE-rich training corpus and providing language learners with additional information about annotated MWEs. Especially the latter makes Wiktionary a great data source for applications targeting language learners (e.g. Simonnet et al., 2024).

Wiktionary is primarily an unstructured wiki maintained by thousands of volunteers with varying degrees of technical skills. Therefore, its source code is expressed in easily formatable and human-readable wikicode, a light-weight markup language leveraging templates and modules in the Lua programming language for formatting content. This setup necessitates the extraction of structured data from Wiktionary to accomplish automated downstream tasks.

Among the several existing extraction projects, DBnary (Sérasset, 2015) and Wiktextextract (Ylonen, 2022) are the most robust and advanced candidates. After comparison, Wiktextextract emerged as the superior option owing to its ability to flexibly expand Lua Templates, thereby achieving a higher extraction quality. A particular concern was that DBnary—due to lacking the same flexibility—

<sup>5</sup><https://wiktionary.org/>

exhibited undesired artifacts in example sentences, which would compromise the integrity of our training corpus.

While the Wiktextextract project publishes a fully extracted dataset of the English Wiktionary which also includes a large number of French headwords<sup>6</sup>, this dataset unfortunately provides only limited coverage of French example sentences—a crucial feature for our project. We, therefore, had to adapt the Wiktextextract script to parse the French Wiktionary directly. The adapted version was able to extract headwords, part-of-speech tags, and word senses. For each word sense, a gloss and example sentences (if present) were extracted as well as potential subsenses, whereby additional tags, categories and meta data such as source and authors were placed in separate fields resulting in clean text for glosses and example texts<sup>7</sup>.

### 5.3 Corpus creation

To support our hypothesis, it is essential to demonstrate that an MWE identification system can be trained using example sentences from a lexicon. These sentences may undergo automatic preprocessing but should require minimal manual intervention. This requirement necessitates that the MWE identification system be capable of learning solely from positive examples, as lexical example sentences provide only positive instances for each MWE.

However, to validate and refine the system, negative examples are needed to measure precision. Consequently, constructing a development and test set involves some degree of manual annotation to identify occurrences and non-occurrences of MWEs. Ultimately, to confirm that the system meets the performance goal of being useful to language learners, a fully annotated test set is required. This test set should ideally be drawn from a corpus representative of general French, rather than from a distribution of lexical example sentences.

In the following sections, we detail the process of creating the training and test sets used to evaluate WikTSeen.

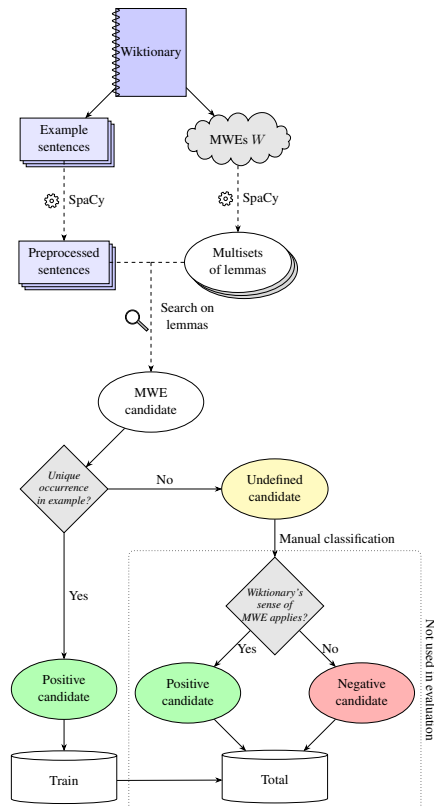


Figure 1: Building a train set from Wiktionary example sentences

#### 5.3.1 Preprocessing the Wiktionary corpus

Following the extraction of structured data from a Wiktionary dump dated 07.04.2023, several steps of data processing were undertaken to construct a coherent training corpus. Figure 1 schematically illustrates this process.

Our initial concern was to identify MWEs among the extracted lexical entries, or more formally, the set  $W$  of MWE types present in Wiktionary. We used whitespace characters within the headwords as discriminating markers for MWEs. Next, the Wiktionary-specific part-of-speech (POS) tags were mapped to Universal POS tags to facilitate universality and integration with existing NLP tools. Furthermore, we flattened the ‘senses’ and ‘subsenses’ fields into a consolidated list of glosses and example sentences for each lemma-POS pair.

Applying this heuristic, we identified 119,561 MWEs, of which 31,794 were plural forms, i.e. having only one gloss containing the string ‘pluriel d’, disregarding capitalization. These plural forms are not useful for two reasons: (i) the corresponding Wiktionary entries contain no definitions other than a reference to the single form entry (we need a definition to explain the meaning of MWEs to the

<sup>6</sup>See <https://kaikki.org/index.html>.

<sup>7</sup>The adapted script is available in the pull request to the main Wiktextextract project: <https://github.com/tatuylonen/wiktextextract/pull/223>. The Wiktextextract project has since expanded and now parses the French Wiktionary edition out of the box.



user), (ii) the occurrences of these forms can still be spotted in text by our MWE identification method, which is based on lemmas of the MWE components. After excluding these plural forms, we were left with 87,767 MWEs in  $W$ , each characterized by a unique lemma-POS combination.

In order to identify the necessary components of each MWE, the lemma of each MWE—represented by the entry’s headword—was automatically tokenized and lemmatized. In this manner, we derived for each MWE type a multiset of single word lemmas whose joint occurrence we consider a necessary condition for the occurrence of the MWE as a whole (e.g. *la crème de la crème* (lit. ‘the cream of the cream’) ‘the best part’ yields {*crème, crème, de, le, le*}).

This process occasionally led to minor inaccuracies, such as converting the headword *a priori* to *\*avoir\_priori* ‘have\_priori’ caused by the added complexity of lemmatizing fragmented text. In adopting this approach, we deferred responsibility for the delicate question of determining the canonical form and necessary components of an MWE to the Wiktionary authors—a pragmatic choice which will have to be justified by the outcomes<sup>8</sup>.

Finally, example texts underwent a SpaCy (Honribal and Montani, 2017) processing pipeline consisting of tokenization, POS-tagging, lemmatization, and dependency parsing. These texts were then partitioned into individual sentences based on parsing outcomes. While these newly delineated sentence boundaries largely matched the original example sentences, they were occasionally more liberal. This strategy was intentional: shorter sentences reduce the complexity of searching for MWE candidates and also mirror the preprocessing steps that our MWE identifier will eventually employ on unprocessed real-world text.

### 5.3.2 Training set

The initial extraction process transforms rich-text formatted Wiktionary entries into plain text, eliminating the specific formatting that often but not always (and not always correctly) marks MWE occurrences in example sentences. Consequently, we needed to re-identify the spans of MWE occurrences within these examples.

To address this, we ran a systematic search for

<sup>8</sup>For instance, while *commencer à* ‘start to (do something)’ is a MWE entry in Wiktionary, it is considered a single verb with a selected preposition (i.e. a word combination relevant to valency rather than to idiomaticity) in Sequoia.

MWEs as defined by their multisets of lemmas across all preprocessed sentences, not limiting the search to just the single MWE a sentence was an example of. To manage the computational complexity of the search, we assumed that MWEs tagged with the POS labels ‘ADJ’ (e.g., *bon à rien* (lit. ‘good for nothing’) ‘unable to succeed’), ‘ADV’ (e.g., *de temps en temps* ‘from time to time’), ‘ADP’ (e.g., *au lieu de* ‘instead of’), ‘CONJ’ (e.g., *à mesure que* ‘as’), ‘INTJ’ (e.g., *à la bonne heure* ‘splendid!’), ‘NOUN’ (e.g., *lune de miel* ‘honeymoon’), and ‘PROPN’ (e.g., *Académie française* ‘the French Academy’) must manifest as continuous lemma sequences in the text. For all other POS tags, we allowed any discontinuities as long as a complete multiset of lemmas was present in an individual sentence. For very prevalent multisets of lemmas, we stopped the search after having found more than 1,000 occurrences.

The search yielded a comprehensive list of MWE candidates. An MWE candidate was automatically included in the training set when it was the sole candidate in a sentence which was known to be an example of that MWE. All other candidates were kept as undefined candidates for potential manual classification. Figure 1 describes this automatic derivation of our training set.

The figure also illustrates a partial manual annotation process of undefined candidates. The outcomes of this effort are included in our total corpus but were used neither in training nor evaluation.

### 5.3.3 Test set

The training set (as well as the manually annotated parts of the total corpus) is composed exclusively of lexical example sentences which, *prima facie*, have no claim to being representative of modern French. To evaluate the real-life performance of our system, as experienced by language learners, a general corpus annotated with MWEs is required. As discussed in Section 5.1, few such corpora exist, with Deep-Sequoia being notable for its inclusion of MWE annotations beyond just verbal MWEs.

The main difficulty in evaluating our system on Deep-Sequoia is the potential discrepancy between its notion of MWEs (the set  $S$  of MWE types) and that of Wiktionary (the set  $W$  of MWE types). Some MWEs are annotated in Deep-Sequoia and included in Wiktionary ( $W \cap S$ ), such as *à peu près* ‘approximately’. Others are included in Wiktionary but not annotated in Deep-Sequoia ( $W \setminus S$ ), such as *en aval de* ‘downstream of’ (only *en aval*

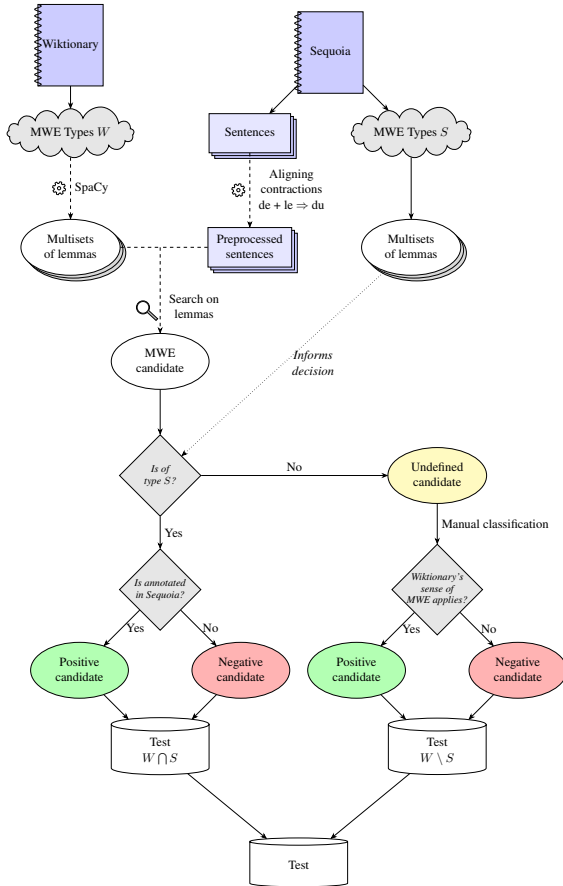


Figure 2: Creating a test set based on Deep-Sequoia

would be annotated in  $S$  according to the Deep-Sequoia annotation guidelines).

Given our choice, as guided by the hypothesis, to follow Wiktionary’s notion of an MWE, we need to evaluate our system’s performance on the entirety of  $W$ . Claiming to annotate all MWEs in  $W$  (extending coverage) while in practice evaluating on only the labels provided in Deep-Sequoia ( $W \cap S$ ), would reduce any claim about satisfactory performance to just the limited subset of  $W \cap S$ . The fact alone that MWEs in  $W \cap S$  have undergone a formal check of their MWE-hood,<sup>9</sup> while those in  $W \setminus S$  are based on the looser standards of the Wiktionary community, necessitates close attention to the latter group.

We, therefore, decided to create our test set based on the Deep-Sequoia corpus, employing a two-pronged approach. For the MWEs from  $W \cap S$ , we reused the annotations from Deep-Sequoia. For the MWEs from  $W \setminus S$ , we added manual anno-

<sup>9</sup>The MWE annotation guidelines used for Sequoia have the form of decision diagrams driven by formal linguistic tests. They are available at [https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/-/wikis/Guide-annotation-PARSEME\\_FR-chapeau](https://gitlab.lis-lab.fr/PARSEME-FR/PARSEME-FR-public/-/wikis/Guide-annotation-PARSEME_FR-chapeau).

tations. Figure 2 describes the creation process of our test set.

Similar to our approach for the training set, we searched for MWE candidates in the Deep-Sequoia corpus using multisets of lemmas corresponding to the MWEs of type  $W$ . We retained the provided corpus annotations (tokenization, POS tags, dependency parsing) but adjusted the contraction of *du* ‘of.the.MASC.SING’ to align with our automatic preprocessing pipeline<sup>10</sup>.

The resulting MWE candidates were pre-selected for either automatic or manual annotation by comparing their multisets of lemmas to those corresponding to MWE types  $S$ . In particular, if *any* occurrence of a given multiset of lemmas was annotated in Deep-Sequoia as an MWE, then a specific occurrence of that multiset of lemmas was automatically classified as either a positive or negative candidate based on whether or not it had an MWE label. This heuristic assumes that Deep-Sequoia is consistent, meaning that, if an MWE was annotated once, all its occurrences are annotated. All other MWE candidates were then manually classified, applying the decision rule: label it as a positive candidate if one of the senses of the MWE entry is present; otherwise, as a negative candidate.

The combined test set comprises MWEs from both  $W \cap S$  and  $W \setminus S$ , covering the entirety of  $W$ <sup>11</sup>.

This dichotomy introduces some label consistency issues. For instance, Deep-Sequoia does not distinguish between MWEs with the same lemma but different POS labels, whereas Wiktionary does (e.g., *à court terme* ‘in the short term’ has an entry as an ADJ and as an ADV). Consequently, an occurrence of *à court terme* ‘in the short term’ might be labelled as both ADJ and ADV if automatically annotated, whereas manual classification would disambiguate the part-of-speech. We expect these inconsistencies to be minimal and consider them an acceptable trade-off for reducing manual annotation efforts.

<sup>10</sup>While Deep-Sequoia tokenizes *du* ‘of.the’ to *de le* ‘of the’, SpaCy keeps it as a single token. Aligning the lemmatization protocols is crucial since our identification system searches for MWE candidates based on the multisets of lemmas seen during training.

<sup>11</sup>Since we are only concerned with identifying and evaluating seen MWEs, in practice, the test set only covers the subset  $\bar{W}$  of  $W$ , which corresponds to the MWEs our system sees during training, i.e. the MWEs included in the training set.

Table 2: Corpus statistics

NO. OF	TOTAL	TRAIN	TEST (SEQUOIA)
MWES	87,767	28,459	1,318
SENT.	126,558	48,020	3,099
TOKENS	2,555,207	1,194,824	68,615
POS. C.	57,053	49,165	1,972
NEG. C.	31,052	0	10,494
UND. C.	1,102,488	233,170	0

## 5.4 Corpus statistics

As a result of the steps described above, distinct outcomes of this paper are a unique MWE corpus, comprising Wiktionary’s example sentences, and the Deep-Sequoia corpus, annotated with MWEs from the Wiktionary lexicon.

Table 2 presents comprehensive statistics for the total corpus, its subset used for training, and the Deep-Sequoia test set, fully annotated with trainable MWEs from Wiktionary<sup>12</sup>. For each set, the table reports the number of MWEs with unique lemma-POS pairings (that have at least one candidate occurrence), sentences, tokens, and the number of found MWE candidates, classified as positive (a true MWE), negative (a literal or incidental occurrence of the constituent lemmas of an MWE in a sentence), or undefined (awaiting manual classification).

One significant achievement is the scale of unique MWEs included in the corpus, which exceeds that of previous data sources by an order of magnitude (compare Section 5.1).

## 6 MWE identification with WiktSeen

The corpus described in the previous section became a cornerstone of *WiktSeen*, a rule-based MWE identification system, closely modeled after the *Seen2020* system developed by Pasquer et al. (2020b). We opted to base *WiktSeen* on this particular model due to its strong performance in identifying seen MWEs during the PARSEME shared task edition 1.1. The rule-based nature of *Seen2020* offers several advantages that align with our research goals. Firstly, it allows for relatively straightforward implementation and customization. Secondly, it is able to learn from positive examples alone, eliminating the need for labeling negative examples (or of making sure to catch all positive examples

<sup>12</sup>It is worth noting that the count of undefined candidates is a conservative estimate. The search for new candidates for MWEs with high-frequency lemma multisets was halted after identifying the first 1000 candidates.

in the dataset). Thirdly, its rule-based architecture enables reasoned analysis and debugging of the system’s performance. This last point is especially important in our setup since it allows us to distinguish errors introduced by the system from errors introduced during the task and dataset design.

These attributes make *WiktSeen* instrumental for testing our hypothesis: that a rule-based system, trained on lexically-rich example sentences, can extend MWE coverage without compromising performance metrics. The previous achievements of *Seen2020* in the PARSEME shared task bolster our confidence in this hypothesis, allowing us to focus more on corpus design and informing the further course of research through user experiments.

A notable enhancement in our implementation is the integration of *WiktSeen* as a custom SpaCy pipeline component. This plug-and-play compatibility enables seamless integration with other natural language processing tasks, facilitating easy deployment in downstream applications<sup>13</sup>.

In the subsequent sections, we will outline the key features of *WiktSeen*, emphasizing where it diverges from the original *Seen2020* system. For a more comprehensive understanding of the underlying architecture, we direct the reader to the original work by Pasquer et al.

### 6.1 Candidate extraction

*WiktSeen* employs a two-stage process for MWE identification, with the first stage dedicated to candidate extraction. During the training phase, the system registers multisets of lemmas corresponding to the necessary components of an MWE for each observed POS and MWE lemma combination. In the prediction stage, *WiktSeen* searches each sentence for matches to these registered multisets of lemmas, effectively identifying initial candidate occurrences of MWEs.

To enhance search efficiency, *WiktSeen* allows for configuration of POS-specific continuous candidate matching. By default, continuous matching is applied to MWEs with the POS tags: ‘ADJ’, ‘ADV’, ‘ADP’, ‘CONJ’, ‘INTJ’, ‘NOUN’, and ‘PROPN’. Candidates that pass this initial extraction are then forwarded to the subsequent stage for further filtering<sup>14</sup>.

<sup>13</sup>The pipeline component is available at <https://github.com/empiriker/mwe-detector>.

<sup>14</sup>It’s worth noting that the candidate extraction stage follows the same logic as our search for annotation candidates during corpus creation. This necessarily impacts the interpre-

## 6.2 Trainable Rule-Based Filters

The second stage in *WiktSeen*'s MWE identification pipeline focuses on enhancing precision through filtering. The system utilizes a combination of seven filters, F1 to F7, that take the observed morphosyntactic properties of MWE components into account.

One key distinction between *WiktSeen* and the original *Seen2020* is in how these filters are trained. While the latter learns filter settings for each MWE class based on PARSEME VMWE tags, *WiktSeen* learns individual filter settings for each specific MWE, except for the global filters F5 and F6. The 7 filters are defined as follows:

### F1: Components should be disambiguated

This filter only accepts candidates with multisets of POS tags that were observed during training (e.g. *point*/VERB *out*/ADV) but not *point*/NOUN *out*/ADV).

### F2: Components should appear in specific orders (Ignoring discontinuities)

This filter only accepts candidates whose POS tags appear in the same order as observed in the training data, disregarding any discontinuities (e.g. *point*/VERB *out*/ADV but not *out point*).

### F3: Components should appear in specific orders (Considering discontinuities)

Similar to F2, but it takes into account all POS tags from the first to the last candidate token, considering discontinuities (e.g. *point* that/PRON *out* but not *point that*/SCONJ *it*/PRON *is*/VERB *out*).

### F4: Components should not be too far

This filter only accepts candidates whose largest discontinuity is no greater than the largest observed discontinuity.

### F5: Closer components are preferred

This global filter selects the candidate with the smallest discontinuity among all matches for a given multiset of lemma within a sentence.

### F6: Components should be syntactically connected

Another global filter that passes candidates where the tokens form a (weakly) connected dependency subgraph or/and are in a grandparent/grandchildren relation.

### F7: Nominal components should have seen inflection

If a candidate match contains exactly one noun, this filter expects the noun to appear with a previously observed inflection (*turn tables* but not *turn table*). If there are zero or

tation of our results which we discuss in the next section.

more than one noun, the candidate automatically passes this filter.

The original *Seen2020* system featured an eighth feature concerned with nested VMWEs. Due to the practical absence of nested MWEs in the Wiktionary-based MWE training corpus, this filter is set permanently to true in *WiktSeen*.

## 6.3 Tuning active filters

In the original *Seen2020* paper, an 8-bit parameter was tuned on the development set to determine which filters should be active during prediction. This 8-bit parameter was trained per language present in the data set and then applied globally for all classes of VMWEs.

Following this lead, we ran all combinations of a 7-bit parameter on a small development set and kept the the best performing filter combination, determined by the  $F_1$ -score, before evaluating on the test set.

In the future, a separate active filter parameter could be trained for each different POS class of MWEs (verbal, nominal...). However, initial experiments have shown that this technique requires quite a large development set. Otherwise, filter tuning would quickly overfit the few MWEs of each POS class present in the development set. We, therefore, opted to only tune a single set of global filters.

## 7 Results

The evaluation of the *WiktSeen* system faces several initial difficulties: a) lack of negative examples in the created French MWE corpus, b) small overlap in MWE-hood with existing corpora, and c) an atypical distribution of MWEs in the training set. These issues were largely addressed through manual creation of a Sequoia-based test set (see Section 5.3.3).

However, the methodology used for corpus creation has its own consequences for interpreting the results. Specifically, the same candidate generation method was used to search for annotation candidates as is used by *WiktSeen* in the candidate extraction stage. This implies that our evaluation method can only reasonably evaluate the second stage, i.e., the filtering stage. Consequently, the baseline recall of our model (without any filtering) would be 100%.

We deem this acceptable in the context of language learning, where it may not be necessary to

match a formally precise span of an MWE. Responsibility for defining the constituent parts of an MWE is delegated to Wiktionary. Furthermore, filtering is considered the harder part compared to candidate extraction, and it is the aspect we aim to evaluate more strictly.

## 7.1 Evaluation procedure

With this in mind, a three-step evaluation procedure was adopted.

**Model training** The model is trained on the training set, which comprises the bulk of the available data without manual classification.

**Filter tuning** We use a (random sentence-based) 20% split of our Deep-Sequoia corpus as a development set. The trained model’s second stage predicts filter values on this set, allowing us to calculate binary classification metrics. We then identify optimal filter settings based on the  $F_1$ -score, balancing precision and recall. This approach ensures that filter tuning occurs on a sample distribution matching the final test set.

**Final evaluation** The model, trained only on the original training set, is evaluated on the remaining 80% of the Deep-Sequoia corpus using the optimal filter settings determined in step 2. This evaluation provides an estimate of the model’s performance on a natural distribution of MWE occurrences, serving as an empirical check on its utility.

Through this evaluation process, we aim to assess WiktSeen’s capabilities in a way that aligns with the project’s objectives and underlying assumptions.

## 7.2 Filter tuning

Figure 3 displays the  $F_1$ -scores across different filter settings. Notably, the highest-performing combinations involve the activation of filters F2, F5 and F6. These filters respectively require the POS tags of MWE components to match the order observed during training (F2), prefer closer components among candidates of the same MWE (F5), and enforce syntactic connectedness (F6).

Apart from the optimal filter set, the figure contains many hints on how to improve filters in a future iteration. Just to give an example, F7 (nominal components should have seen inflection) seems to extraordinarily benefit precision albeit at a huge price in recall. A conclusion might be that only some MWE classes profit from F7, or that the training set was not diverse enough in terms of MWEs

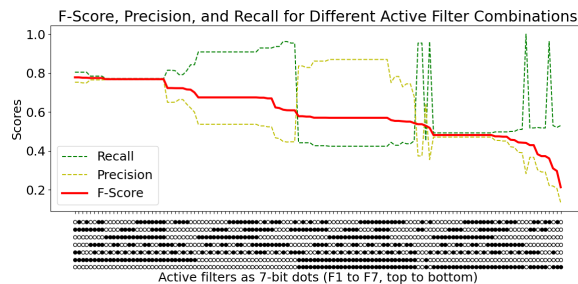


Figure 3: Performance for different filters on dev

whose nominal components are not fixed.

## 7.3 Results on Deep-Sequoia

We report the results on the Deep-Sequoia corpus with optimal filter settings for the entire test set (80% split) and its partitions by annotation process and POS. Table 3 presents the global metrics and metrics for subsets corresponding to the MWE types  $W \cap S$  and  $W \setminus S$  (see Figure 2). Table 4 provides metrics per POS class. For better interpretability, both tables include the number of MWE candidates (positive/positive+negative candidates) and the number of unique MWE candidates (with at least one positive candidate/with any positive or negative candidate) per respective subset.

On the full test set, *WiktSeen* achieves an  $F_1$ -score of 0.776. However, a significant disparity emerges when comparing the results on  $W \cap S$  and  $W \setminus S$ . For MWEs adhering to the formal definition of MWE-hood in Deep-Sequoia, the identification task appears nearly solved with an  $F_1$ -score of 0.929. However, for MWEs introduced only in Wiktionary, the  $F_1$ -score drops to 0.535.

This disparity can be partly attributed to the composition of each data slice in terms of unique MWEs with positive versus any candidate occurrence. In  $W \cap S$ , the ratio of true candidates to all candidate matches is  $\frac{1100}{1554} \approx \frac{7}{10}$ , compared to  $\frac{624}{8716} \approx \frac{7}{100}$  in  $W \setminus S$ . This suggests that expressions considered MWEs by Deep-Sequoia exhibit multisets of lemmas that are more likely to be true candidates, whereas Wiktionary introduces many MWEs lacking this property, making identification much harder in  $W \setminus S$ . In a sense, this is the opposite relationship of what Savary et al. (2019c) have found for verbal (!) MWEs: i.e., that any morphological and syntactical candidate structure that exhibits features of a VMWE is much more likely to be a true occurrence of the MWE than a literal reading. Apparently,  $W \setminus S$  introduces many, mostly non-verbal MWEs that exhibit the opposite

Table 3: Performance on test with top filters

	TEST	TEST <sub> w<sub>ns</sub></sub>	TEST <sub> w<sub>ls</sub></sub>
<b>F1</b>	<b>0.776</b>	0.929	0.535
<b>PRECISION</b>	<b>0.751</b>	0.939	0.484
<b>RECALL</b>	<b>0.804</b>	0.92	0.598
<b># OCCS</b>	1,734/10,270	1,110/1,554	624/8,716
<b># MWES</b>	709/1,258	427/432	282/826

Table 4: Performance on test with top filters by POS

POS	F <sub>1</sub>	PREC.	REC.	# OCC.	# MWE
ADJ	0.664	0.531	0.886	88/232	41/78
ADP	0.618	0.787	0.509	283/714	64/89
ADV	0.771	0.660	0.927	327/1,112	162/266
CONJ	0.718	0.832	0.632	133/199	36/48
INTJ	0.714	0.556	1.000	5/20	3/13
NOUN	0.913	0.907	0.919	467/523	237/261
PRON	0.732	0.872	0.631	65/2,024	6/31
PROPN	0.722	0.700	0.745	47/76	17/25
VERB	0.777	0.708	0.862	318/5,078	142/428
X	1.000	1.000	1.000	1/292	1/19

relationship between occurrences of their multisets of lemmas and true idiomatic occurrences.

Examining results by POS class, *WiktSeen*'s performance remains relatively stable across different groups. It performs best on nominal MWEs, averaging on verbal MWEs, and worse on adjective and adpositional MWEs. These results indicate that *WiktSeen* generalizes well across MWE classes but also highlight areas for improvement. The low precision for adjective MWEs is partly due to the difficulty in distinguishing them from adverbial MWEs, which often share the same lemma multisets. The poor recall for adpositional MWEs may result from F6's check for syntactic connectedness disproportionately affecting this MWE class. These observations suggest directions for error analysis and future enhancements.

Overall, the global F<sub>1</sub>-score of 0.776 is encouraging. We hypothesize that human language learners, even without expert knowledge of their target language, can tolerate some noise in MWE identification without compromising its usefulness. While its performance leaves room for improvement, *WiktSeen* can likely already provide real-world value. We tested this hypothesis through application in Linguse and subsequent user experiments, as discussed in the following sections.

## 8 Linguse

Linguse is a reading application for language learners that predates this research<sup>15</sup>. As a web application, it allows learners to upload texts in various formats and provides an interface optimized for reading comprehension and vocabulary acquisition. This is achieved by identifying all lexical items in a text, facilitating context-aware retrieval of glosses and translations, and cross-referencing them with lexical items previously read by the user. Originally, Linguse's identification of lexical items was limited to single words. In our research, we collaborated with Linguse to enhance its reading interface by integrating MWE identification, enabling us to test how language learners interacted with and appreciated the identification of MWEs in their reading material.

While e-books and reading devices are widely used by foreign language learners, research on their educational use, particularly on the impact of their dictionary functionality (typically involving single-word identification and annotation) on the development of reading and lexical skills, is scarce in foreign language teaching literature (Davidson and Carliner, 2014; Rettberg, 2020). MWE identification is rarely implemented in reading devices,<sup>16</sup> highlighting the significance of our efforts to develop this functionality in Linguse and test its effectiveness through user experiments. This gap in existing tools motivated our development and assessment of MWE identification within Linguse, aiming to enhance language learning outcomes.

## 9 User experiments

The primary aim of the didactic part of this study is to collect feedback from end-users (French language learners), offering valuable insights into their specific needs and practical considerations. This, in turn, is expected to inform the scientific community, refining the scope of scientific tasks in alignment with real-world applications and shaping the trajectory of future research. These experiments were undertaken in partnership with the Institute of Romance Studies at Warsaw University. A class of 12 students studying the French language at the B1 level participated in the study. The experiments,

<sup>15</sup> Accessible via <https://linguse.com>.

<sup>16</sup> Kindle provides definitions for some manually selected phrases in English. See also <https://github.com/BoTiG/ebook-reader-dict/blob/master/docs/fr/README.md>

conducted from mid-May to mid-June 2023, were guided by three primary objectives:

1. to assess the impact of MWE identification on language learning,
2. to evaluate Wiktionary's utility as a guideline and knowledge base for MWE annotation,
3. to understand the practical needs and expectations of B1-learners with respect to MWE identification.

The user experiments were designed with a focus on gathering qualitative data, but quantitative part was also necessary. Participants were given a set of three tasks to be performed in their own time: a prequiz to assess their prior knowledge of MWEs, a reading task based on a series of French texts (Fournier, 2011) within the Linguse application, internally annotated with MWEs (throughout this period, they were supposed to take notes on any aspects they found confusing, useful, or interesting), a postquiz to assess any improvement or changes in their understanding of MWEs.

The pre- and the postquizzes, providing data for the quantitative evaluation, were based on the *Vocabulary Knowledge Scale* (Paribakht and Wesche, 1996) that requires the participants to evaluate their understanding of an MWE on a 5-level scale. While the first two levels take the learner's self-evaluation at face value, for the subsequent categories, participants were requested to provide evidence of their knowledge, such as synonyms, translations, or example sentences. This combination of self-reporting and evidence-based scoring allows us to gauge not just the breadth but also the depth of participants' MWE knowledge. The qualitative evaluation consolidated insights from both a semi-structured group feedback session and semi-structured individual interviews.

The user experiments reveal several key findings that contribute to both theoretical and practical discourse on MWE identification in language learning. Regarding the quantitative evaluation, the most salient outcome pertains to the difference of prequiz and postquiz results. The score obtained in the postquiz, representing the knowledge of 10 MWEs randomly chosen from the text read by the students during the second task of the experiment, was 2.575 and it increased by 0.775 compared to the score in the prequiz regarding the same MWEs. This result may suggest a positive influence of MWE-annotated texts on lexical competency; however, the robustness of these findings is limited by the

low participant count, and therefore, further studies are needed for more conclusive evidence.

As far as the qualitative feedback is concerned, overall, three themes closely related to our objective to evaluate the efficiency of the MWE identification and annotation for a reading tool emerged from the feedback, whose conclusions are very briefly presented in the following:

**General Experience:** Users generally expressed a positive to very positive sentiment towards the tool, affirming its utility in aiding their reading in a foreign language, as it can be confirmed by this statement: "Normally I want to look up all unknown words; here it was easy to focus on the text".

**Reading Assistance:** The tool's multi-faceted reading assistance, which includes word and MWE definitions, but also the availability of alternative help, like translations, useful when definitions were insufficient, was praised by the students. We noted the following opinion: "I liked that there were often multiple definitions for a word. Though sometimes definitions were missing or not sufficient. Then the translation feature helped me".

**Annotation Quality:** Some students noted inadequacies regarding annotation, but they were forgiving of minor annotation errors, suggesting that perfect accuracy is not required for the reading tool to be beneficial. It can be illustrated by the following statement: "When reading it was most important to understand the bigger picture, small annotation errors didn't matter".

To sum up, the overarching need for MWE identification tools in language acquisition was validated by user experience. It should also be emphasized that the utility of providing comprehensive lexical information emerged as crucial, reinforcing the strengths of our lexicon-based approach, which, by design, links to lexical data sources. Furthermore, our innovative didactic approach to grounding MWE identification in a community-driven lexicon faced no objections from participants, who are frequent users of resources like Wikipedia or Wiktionary. This suggests the practicality of the reading tool developed in our study and indicates a negligible impact of any inaccuracies on its overall usefulness.

## 10 Conclusions

This research project, situated at the intersection of NLP and language learning, aimed to enhance

learning activities through MWE identification while providing valuable insights from end users to MWE research.

Our findings support the hypothesis that a rule-based system, trained solely on positive MWE examples from a lexicon, can significantly expand MWE coverage while maintaining satisfactory performance metrics. The MWE coverage of our system is an order of magnitude larger compared to other sources. User experiments confirmed that language learners highly value broad MWE coverage, which is essential for assisting learners at various levels of expertise. Although the performance metrics of our rule-based system, *WiktSeen*, are not outstanding, they are deemed satisfactory because they do not detract from its utility for language learners. On the contrary, user experiments indicate that second language learners can handle noisy assistance as long as a multitude of resources are provided in context.

## 11 Implications and Future Work

The outcomes of this project offer promising avenues for future research and development. Specifically, the user-oriented components of the project, such as the MWE-annotated reading interface, have demonstrated practical benefits for language learning.

The immediate next step could be to provide a larger development set by expanding Wiktionary-based MWE annotations to the PARSEME corpus. This would allow for a more nuanced evaluation of the system's performance and potentially lead to class-specific filter optimizations. Other aspects of diversity, such as assessing the variety and disparity of MWE types (Lion-Bouton et al., 2022) both in the dataset and in system predictions, might prove beneficial for the lexical competence of language learners.

Further enhancements to the system itself should also be explored. New filters could be devised to target prevalent error sources. While it is tempting to explore advanced machine-learning algorithms such as transformers for MWE identification, we consider a gradual approach. Preliminary results and user feedback indicate that significant real-world benefits can still be obtained using the existing rule-based system, thus questioning the immediate need for adding complexity through a vector-/transformer-based approach.

We would also like to explore how well our

method translates to other languages in order to provide assistance to learners of target languages other than French, too. Wiktextextract has recently started to extract and make available data from the Chinese, German, Japanese, Polish, Russian, and Spanish editions of Wiktionary which considerably improves the availability of MWE lexica with example sentences. Finally, the fact that *WiktSeen* is based on *Seen2020* which was tested and evaluated on 14 languages (one of which was French) with good overall results, gives us reason to optimism that, using our approach, similar results are possible for more languages.

## Acknowledgments

This work was funded by an internship grant from the Graduate School in Computer Science of the Paris-Saclay University, as well as by the French Agence Nationale pour la Recherche, through the SELEXINI project (ANR-21-CE23-0033-01).

## References

- Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkha and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2 edition, pages 267–292. CRC Press, Boca Raton, USA.
- Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Yannick Parmentier, Caroline Pasquer, and Jean-Yves Antoine. 2017. [Annotation d'expressions polylexicales verbales en français](#). In *24e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, volume 2 of *Actes de TALN, volume 2 : articles courts*, pages 1–9, Orléans, France.
- Cristelle Cavalla. 2009. La phraséologie en classe de FLE. *Les Langues Modernes*, (1).
- Cristelle Cavalla. 2016. [Comment analyser sémantiquement les expressions figées ?](#) *Revue de Sémiotique et Pragmatique*, (39).
- Cristelle Cavalla and Virginie Labre. 2019. L'enseignement en FLE de la phraséologie du lexique des affects. In Iva Novakova and Agnès Tutin, editors, *Le Lexique des émotions*, pages 297–316. UGA Éditions.
- Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. [Multiword Expression Processing: A Survey](#). *Computational Linguistics*, 43(4):837–892.
- Cristiana Cornea. 2010. [Le rôle de la lecture dans l'apprentissage et l'utilisation du FLE](#). In *Le français*



- de demain : enjeux éducatifs et professionnels*, Sofia. Colloque international. 2010-10-28/2010-10-30.
- Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Press Syndicate of the University of Cambridge, Cambridge, U.K.
- Ann-Louise Davidson and Saul Carliner. 2014. *e-Books for Educational Uses*, pages 713–722. Springer New York, New York, NY.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Agnieszka Dryjańska. 2024. **Elementy językowego obrazu świata w nauczaniu języka francuskiego w kontekście filologicznym**. *Neofilolog*, (62/2):409–426.
- Stefan Evert. 2005. *The statistics of word cooccurrences : word pairs and collocations*. Doctoral Thesis, University of Stuttgart. Accepted: 2005-09-01. Alternative Title: Zur statistischen Analyse von Wortkombinationen: Wortpaare und Kollokationen.
- Jean-Louis Fournier. 2011. *Où on va, papa?* 3. éd. Librairie générale française, Paris.
- Thomas François, Nùria Gala, Patrick Watrin, and Cédric Fairon. 2014. **FLELex: a graded lexical resource for French foreign learners**. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3766–3773, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Francis Grossmann. 2011. **Didactique du lexique : état des lieux et nouvelles orientations**. *Pratiques*, (61):149–150.
- Francis Grossmann. 2012. **Le rôle de la compétence lexicale dans le processus de lecture et l'interprétation des textes**. *Forumlecture – Littératie dans la recherche et la pratique*, 2012(1). Section: Artikel.
- Najet Hadj Mohamed, Agata Savary, Cherifa Ben Khelil, Jean-Yves Antoine, Iskandar Keskes, and Lamia Hadrach-Belguith. 2024. **Lexicons gain the upper hand in Arabic MWE identification**. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 88–97, Torino, Italia. ELRA and ICCL.
- Matthew Honnibal and Ines Montani. 2017. **spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing**. To appear.
- Kamil Kanclerz and Maciej Piasecki. 2022. **Deep Neural Representations for Multiword Expressions Detection**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 444–453, Dublin, Ireland. Association for Computational Linguistics.
- Adam Lion-Bouton, Yagmur Ozturk, Agata Savary, and Jean-Yves Antoine. 2022. **Evaluating diversity of multiword expressions in annotated text**. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3285–3295, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Salah Mejri. 1999. **Unite lexicale et polylexicalité**. *Linx*, (39).
- Grace Muzny and Luke Zettlemoyer. 2013. **Automatic idiom identification in Wiktionary**. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421, Seattle, Washington, USA. Association for Computational Linguistics.
- T. Sima Paribakht and Marjorie Wesche. 1996. **Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition**. In James Coady and Thomas Huckin, editors, *Second Language Vocabulary Acquisition: A Rationale for Pedagogy*, Cambridge Applied Linguistics, pages 174–200. Cambridge University Press, Cambridge.
- Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020a. **Seen2Unseen at PARSEME Shared Task 2020: All Roads do not Lead to Unseen Verb-Noun VMWEs**. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 124–129, online. Association for Computational Linguistics.
- Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020b. **Verbal Multiword Expression Identification: Do We Need a Sledgehammer to Crack a Nut?** In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Christian Puren. 2016. **La procédure standard d'exercitation en langue**. site de didactique des langues-cultures. <https://www.christianpuren.com/>.
- Carlos Ramisch, S. Cordeiro, Agata Savary, V. Vincze, Verginica Barbu Mititelu, Archana Bhatia, Maja Buljan, Marie Candito, P. Gantar, Voula Giouli, Tunga Güngör, A. Hawwari, U. Inurrieta, J. Kovalevskaite, Simon Krek, Timm Lichte, Chaya Liebskind, J. Monti, Carla Parra Escartín, Behrang Q. Zadeh, Renata Ramisch, Nathan Schneider, I. Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. **Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions**. In *Proceedings of the Joint Workshop on Linguistic*

- Annotation, Multiword Expressions and Constructions*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archana Bhatia, Uxo Iñurrieta, Voula Giouli, Tunga Gungor, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.
- Scott Rettberg. 2020. [Teaching electronic literature using electronic literature](#). *Matlit Revista do Programa de Doutorado em Materialidades da Literatura*, 8:23–44.
- Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. [Multiword Expressions: A Pain in the Neck for NLP](#). In Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, and Alexander Gelbukh, editors, *Computational Linguistics and Intelligent Text Processing*, volume 2276, pages 1–15. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.
- Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019a. [Without lexicons, multiword expression identification will never fly: A position statement](#). In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.
- Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxo Iñurrieta, and Voula Giouli. 2019b. [Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir](#). *The Prague Bulletin of Mathematical Linguistics*, 112(1):5–54.
- Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxo Iñurrieta, and Voula Giouli. 2019c. [Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir](#). *The Prague Bulletin of Mathematical Linguistics*, 112:5–54.
- Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang Qasemizadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. [The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions](#). In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.
- Enzo Simonnet, Mathieu Loiseau, Émilie Magnat, and Élise Lavoué. 2024. [Spread the Word! BaLex, A Gamified Lexical Database for Collaborative Vocabulary Learning](#). In *Proceedings of the 16th International Conference on Computer Supported Education*, pages 388–395, Angers, France. SCITEPRESS - Science and Technology Publications.
- Frank Smadja. 1993. Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19(1).
- Monika Sułkowska. 2013. *De la phraséologie à la phraséodidactique*. Wydawnictwo Uniwersytetu Śląskiego, Katowice.
- Gilles Sérasset. 2015. [DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF](#). *Semantic Web*, 6(4):355–361.
- Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. [MTLB-STRUCT @parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models](#). In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.
- Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. [ID10M: Idiom identification in 10 languages](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.
- Amalia Todirascu, Thomas François, and Marion Cargill. 2024. [PolylexFLE: A MWE database for French L2 language learners](#). *ITL - International Journal of Applied Linguistics*.
- Agnès Tutin. 2018. [Les expressions polylexicales transdisciplinaires dans les articles de recherche en sciences humaines : retour d’expérience \(Chapitre 4\)](#). In M.P. & Tutin A. Jacques, editor, *Lexique transversal et formules discursives des sciences humaines*, pages 91–112. ISTE.
- Tatu Ylonen. 2022. [Wiktextextract: Wiktionary as Machine-Readable Structured Data](#). In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1317–1325, Marseille.
- Ziheng Zeng and Suma Bhat. 2021. [Idiomatic expression identification using semantic compatibility](#). *Transactions of the Association for Computational Linguistics*, 9:1546–1562.



## Linköping Electronic Conference Proceedings Nr. 211

eISSN 1650-3740 (Online)

ISSN 1650-3686 (Print)

ISBN 978-91-8075-774-4 (Print)

2024