CLARIN

Selected papers from the **CLARIN Annual Conference 2019**

Leipzig, Germany

Selected Papers from the

CLARIN Annual Conference 2019

Leipzig, 30 September - 2 October 2019

edited by Kiril Simov, Maria Eskevich

CLARIN

Common Language Resources and Technology Infrastructure

Front Cover Illustration:
Picture Composition by CLARIN ERIC

Linköping Electronic Conference Proceedings 172
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2020
ISBN 978-91-7929-807-4

Introduction

Franciska de Jong
Executive Director CLARIN ERIC
Universiteit Utrecht, The Netherlands
f.m.g.dejong@uu.nl

Kiril Simov
IICT, Bulgarian Academy of Sciences
Sofia, Bulgaria
Programme Committee Chair
kivs@bultreebank.org

This volume presents the highlights of the 8^{th} CLARIN Annual Conference 2019 held in Leipzig, Germnany, on 30^{th} September —2^{nd} October 2019.

CLARIN ERIC[1] is the European Research Infrastructure for Language Resources and Technology aimed at supporting researchers from the Social Sciences and Humanities (SSH) and beyond in their use of language data and technologies. CLARIN works towards lowering barriers in doing research by giving access to language resources distributed across the countries involved in the infrastructure and by offering advanced, user-friendly and effective applications that enable the analysis of textual data, speech recordings, as well as multimodal material in a wide diversity of research tasks.

Since the establishment of the ERIC in 2012, CLARIN has grown ins size considerably. Currently there are 21 member countries, 3 observers, and more than 100 associated research institutions who are all encouraged and supported to be represented at the annual conference which is meant to be a central event for CLARIN community and which is one of the crucial instrument for CLARIN to function as a knowledge hub. At the conference, consortia from all participating countries and the various communities of use meet, in order to exchange ideas, experiences and best practices in using the CLARIN infrastructure. The conference covers a wide range of topics, including the design, construction and operation of the CLARIN infrastructure, the data, tools and services that are or could be on offer, its actual use by researchers, its relation to other infrastructures and projects, and the CLARIN Knowledge Infrastructure. The aim is to attract researchers from all the various SSH fields that work with language materials, i.e. the people who are the raison d'être for CLARIN.

For the 8^{th} edition of the CLARIN Annual Conference the special topic was "Humanities and Social Science research enabled by language resources and technology". Early in 2019 a call[2] was issued for which 56 abstracts were submitted.

All submissions were reviewed anonymously by three reviewers (PC members and reviewers invited by PC members). Out of the 56 submitted abstracts 44 submissions were accepted for presentation at the conference (acceptance rate 0.79). The three topics that attracted the most of proposals were (a) language resources and tools, (b) design and construction of the CLARIN infrastructure, and (c) Interoperability and technical issues. This year 14 papers with a link to the special topic were accepted. Three of them had the special topic as the main focus, and thus formed the core of the plenary session devoted to it, while the others were presented either in other oral sessions or as posters. The accepted contributions were published in the online Proceedings of the Conference[3]. Several papers reported collaborative efforts by researchers from different institutes, either from one country or from institutional nodes in multiple countries. Some of the papers describe work that was carried out together with researchers from countries outside of Europe, such as South Africa, Japan, and Russia.

Following the well received novelty introduced at the 2018 edition of the CLARIN Annual Conference, a student poster session was organised with 13 presentations by PhD students who were selected by the national coordinator of their country. The abstracts of the student presentations were published in the online CLARIN 2019 Book of Abstracts[4].

[1] http://clarin.eu

[2] https://www.clarin.eu/content/call-abstracts-clarin-annual-conference-2019

[3] https://office.clarin.eu/v/CE-2019-1512_CLARIN2019_ConferenceProceedings.pdf

[4] https://www.clarin.eu/clarin-annual-conference-2019-abstracts

Franciska de Jong and Kiril Simov 2020. Introduction. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 i–iii.

The conference hosted two invited talks:

- Professor Scott Rettberg (University of Bergen, Norway) introduced the audience to the various aspects of Electronic Literature: Documenting and Archiving Multimodal Computational Writing. In his presentation he addressed efforts to disseminate, document, and archive the field of electronic literature. After providing some examples of genres of electronic literature, and a number of projects was presented such as the Electronic Literature Collections, the ELMCIP Electronic Literature Knowledge Base, and the Electronic Literature Archive that seek to preserve a corpus of work and criticism for the future.

- Professor Elke Teich (University of the Saarland, Saarbrücken, Germany) gave a talk dedicated to corpus-driven investigation of language use, variation and change. Taking the perspective of a "humanist-as-scientist", in her talk she reflected on the requirements for empirically investigating language use, variation and change with special regard to computational resources, models and tools.

In April 2019 CLARIN has launched the CLARIN Ambassadors Programme[5] with the aim to raise awareness about CLARIN in disciplines and communities that are not yet fully familiar with the potential benefits of using the CLARIN infrastructure. One of the three appointed Ambassadors, Toine Pieters (Utrecht University, The Netherlands) gave an invited talk at the conference that was entitled "Towards a Universe of Local Time Machines - building an open eco-system for applied heritage fuelled by common language resources and existing infrastructure".

In addition, on the event page[6] CLARIN published a rich set of materials related to the conference:

- The complete conference programme and most of the slides presented: https://www.clarin.eu/content/programme-clarin-annual-conference-2019

- Recordings of most talks. (Note that the two invited lectures and several other video materials are available on a dedicated channel of VideoLectures: http://videolectures.net/clarinannualconference2019_leipzig/)

After the conference, the authors of the accepted papers and student submissions were invited to submit full versions of their papers to be considered for the post-conference proceedings volume. The papers were anonymously reviewed, each by three PC members. We received 18 (including 1 student paper) full length submissions, out of which 16 were accepted for this volume. All the main topics addressed at the conference are covered in this volume.

We would like to thank all PC members and reviewers for their efforts in evaluating and re-evaluating the submissions, Maria Eskevich from CLARIN Office for her indispensable support in the process of preparing these proceedings, and at Linköping University Electronic Press, who has ensured that the digital publication of this volume came about smoothly.

Members of the Programme Committee for the CLARIN Annual Conference 2019:

- Lars Borin, Språkbanken, University of Gothenburg, Sweden

- António Branco, Universidade de Lisboa, Portugal

- Griet Depoorter, Institute for the Dutch Language, The Netherlands/Flanders

- Koenraad De Smedt, University of Bergen, Norway

- Roald Eiselen, South African Centre for Digital Language Resources, South Africa

- Tomaž Erjavec, Jožef Stefan Institute, Slovenia

[5] https://www.clarin.eu/news/meet-clarin-ambassadors

[6] https://www.clarin.eu/event/2019/clarin-annual-conference-2019-leipzig-germany

- Eva Hajičová, Charles University Prague, Czech Republic
- Erhard Hinrichs, University of Tübingen, Germany
- Nicolas Larrousse, Huma-Num, France
- Krister Lindén, University of Helsinki, Finland
- Monica Monachini, Institute of Computational Linguistics “A. Zampolli”, Italy
- Karlheinz Mörth, Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences, Austria
- Costanza Navarretta, University of Copenhagen, Denmark
- Jan Odijk, Utrecht University, The Netherlands
- Maciej Piasecki, Wrocław University of Science and Technology, Poland
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center, Greece
- Eirikur Rögnvaldsson, University of Iceland, Iceland
- Kiril Simov, IICT, Bulgarian Academy of Sciences, Bulgaria (Chair)
- Inguna Skadiņa, University of Latvia, Latvia
- Marko Tadič , University of Zagreb, Croatia
- Jurgita Vaičenonienė, Vytautas Magnus University, Lithuania
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary
- Kadri Vider, University of Tartu, Estonia
- Martin Wynne, University of Oxford, United Kingdom

Additional reviewers of this volume:

- Laska Laskova, Sofia University “St. Kliment Ochridski”, Bulgaria

Contents

Studying Disability Related Terms with Swe-Clarin Resources

Lars Ahrenberg[1], **Henrik Danielsson**[2], **Staffan Bengtsson**[3], **Hampus Arvå**[1],
Lotta Holme[2], **Arne Jönsson**[1]
[1] Department of Computer and Information Science, Linköping University, Sweden
lars.ahrenberg|hampus.arva|arne.jonsson@liu.se
[2] The Swedish Institute for Disability Research, Linköping University, Sweden
henrik.danielsson|lotta.holme@liu.se
[3] The Swedish Institute for Disability Research, Jönköping University, Sweden
staffan.bengtsson@ju.se

Abstract

In Swedish, as in other languages, the words used to refer to disabilities and people with disabilities are manifold. Recommendations as to which terms to use have been changed several times over the last hundred years. In this exploratory paper we have used textual resources provided by Swe-Clarin to study such changes quantitatively. We demonstrate that old and new recommendations co-exist for long periods of time, and that usage sometimes converges.

1 Introduction

Digitisation (with Optical Character Recognition, OCR) of textual material previously available only in print has enabled large-scale quantitative studies of the recorded past. Coupled with methodological developments in natural language processing (NLP) and other fields, researchers in the humanities and social sciences can study the past in new and powerful ways. Well known examples can be found in the study of literature (Moretti, 2005; Jockers, 2013), in cultural history (Michel et al., 2011) and language change (Tahmasebi et al., 2018).

It is noteworthy that much of the research so far has been conducted on English data. As the quantity of historical Swedish texts that are digitised is increasing, linguistic change in Swedish, whether ”natural” or prompted by technological innovations or by the recommendations of public authorities, can begin to be studied by digital methods.

In this study, we are concerned with lexical changes in the domain of disability. This domain is of special interest in a Swedish setting as the understanding of what disability is, and what it means, has been the subject of much debate, causing new recommendations to be issued from time to time as regards appropriate terminology (see Section 2).

There are some qualitative studies in the area of disability research in Sweden, (for example (Gustavsson Holmström, 2005; Gardeström, 2006; Marie-Louise Larsson-Severinsson and Tideman, 2009; Holme, 2000; Lindberg and Grönvik, 2011; Oliver, 2013; SOU 2019:23, 2019; Söder, 1982; Söder, 2013), that describe and analyse the changes in disability concepts. A problem with these studies is that they often are based on each other and on a selection of texts that both present and argue for renewing the concepts of disability in politics and society. Some of these studies therefore show similar results focusing the very same narrative.

In this article we argue that quantitative studies can contribute to enhance and enrich knowledge about changes of disability concepts. A more varied picture of the concepts of disability may emerge when new possibilities arise for analysing larger amounts of data. In this paper we give some examples to support this view, using Swe-Clarin resources and word space models. To the best of our knowledge, this is the first quantitative study of Swedish disability terms.

The study is a collaboration between computational linguists on the one hand and historians and disability researchers on the other. It is ongoing; in this paper, we report some early results.

Lars Ahrenberg, Henrik Danielsson, Hampus Arvå, Staffan Bengtsson, Lotta Holme and Arne Jönsson 2020. Studying Disability Related Terms with Swe-Clarin Resources. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 1–11.

2 The concept of disability (in Sweden)

Several models and perspectives have been discussed and proposed in relation to disability. The traditional way to approach this field has been labelled the *medical* or *individual* model. This is foremost a term that has been introduced by its opponents, as a contrast, and can hardly be said to have many advocates. The medical model tends to reduce the phenomenon to body functions and bodily deficits. Thus, disability occurs on an individual level, since it is the restrictions caused by physical or mental deviations or flaws that explain why someone experiences problems in everyday life. The inherent logic of the medical model is to a large extent guided by ideas of bodily normality and therefore much of the attention is directed to compensation.

This way of approaching and understanding disability has been challenged by the environmental turn first materialised in the so-called *social model*, a model that emerged within British disability activism in the 1970s. As opposed to the medical model, disability is rather viewed as the outcome of social, structural and institutional barriers. What turns an impairment into a disability depends on how the society is constructed. If society creates barriers in forms of both physical inaccessibility and degrading attitudes leading to various actions of discrimination the answer is not about compensational measures but to change how society works. According to the social model, disability is about combating these social barriers.

One objection to the social model has been its alleged neglect of impairments and the body as well as the experience of the individual. As a part of this criticism competing models have been developed. In Scandinavia, the *relative*, or *relational model* has gained ground. According to this approach, the question of what becomes a disability is not given but is shaped as a result of the interaction between the individual and the surrounding environment. A person with a certain impairment can be disabled in one specific context or situation but not in another. It depends on how the environment is constructed and what type of support is available. While the social model's claim of universal barriers, injustice and discrimination is difficult to maintain, the relational model is close to it by emphasising that disability must be understood in relation to the environment.

A great breakthrough for the disability movement in Sweden came in the 1970s when the Disability Federation Central Committee introduced a joint disability programme, called A Society for All (Centralkommitté, 1972) As early as the 1960s, the concept of disability in official documents and legal texts included some social model elements, and in the programme A Society for All it was claimed that society and the environment should be designed according to the needs of all citizens. It was not enough to bring the individual to society; society must also be made accessible.

Given this history we set out to study the usage of disability related terms over time with the following questions in mind:

- How fast are new understandings of disability, and new recommendations as to term usage, in particular, followed in official reports and media?
- What can quantitative studies based on language technology contribute to our understanding of disability related terms and concepts?

3 Methods and results

We decided to use the Official Reports of the Swedish Government (henceforth: SOU[1]) as our primary source for the study. The SOU's are ordered by the government and are primarily used by them, the parliament and organisations that may have an interest in the issues presented in an SOU. They include a number of reports where disability is the main focus, as well as reports with other main themes where disability may be a side theme or just touched upon in some sections.

[1] An acronym for Statens Offentliga Utredningar.

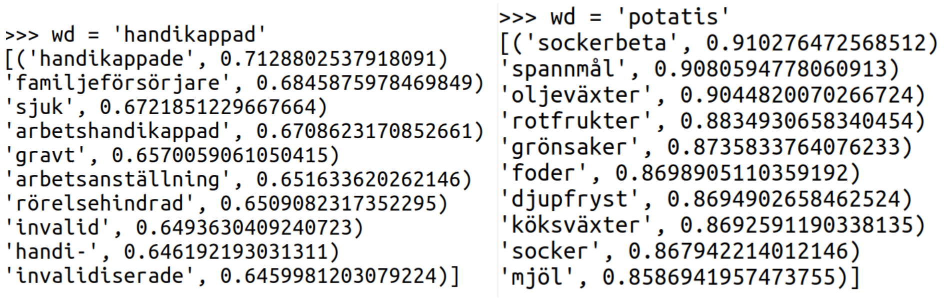
Four types of studies have been performed:

- Frequency studies, where the main methodological issue has been handling inflection in noisy text (Section 3.3)
- Studies of co-occurring terms, using the Swe-Clarin Korp web service (Section 3.4)
- Word space modelling of relevant terms (Sections 3.1 and 3.5)
- A study of temporal analogies based on transformations of word spaces (Section 3.6)

3.1 Data and preprocessing

The SOU reports covering the years from 1922 to 2016 were downloaded from the Swedish Language Bank[2], the resource hub of Swe-Clarin[3]. For the study on co-occurring terms, we have included newspaper texts, also from the Swedish Language Bank.

A main issue for the methods we wanted to use was the quality of the older texts that have been scanned using optical character recognition (OCR). There are frequent misreadings of certain letter combinations, but more seriously, hyphenated words are recognised as two separate units and the OCR does not at all handle the common two-column format, thus, often completely misrepresenting the original text. It was our hope, however, that the methods would be robust enough to allow general trends in the data to be captured even in the presence of noise. To check this we looked at the ten nearest neighbours in vector spaces created as described below in Section 3.5. This was done both for disability related words and for common words, and for each decade covered by the corpus. It turned out that the absolute majority of neighbours were either morphological variants, misrepresented variants, semantically related words such as near synonyms or co-hyponyms, or syntactic attributes. As shown in Figure 1, the word *handikappad* for the period 1970-79 had neighbours such as *handikappade*, a morphological variant, *handi-*, a hyphenated part, *gravt*, 'seriously', a syntactic attribute, and a number of semantically related words. For the word *potatis*, 'potato(es)', all neighbours are related to food.



```
>>> wd = 'handikappad'
[('handikappade', 0.7128802537918091)
'familjeförsörjare', 0.6845875978469849)
'sjuk', 0.672185122967664)
'arbetshandikappad', 0.6708623170852661)
'gravt', 0.6570059061050415)
'arbetsanställning', 0.651633620262146)
'rörelsehindrad', 0.6509082317352295)
'invalid', 0.6493630409240723)
'handi-', 0.646192193031311)
'invalidiserade', 0.6459981203079224)]
```

```
>>> wd = 'potatis'
[('sockerbeta', 0.910276472568512)
'spannmål', 0.9080594778060913)
'oljeväxter', 0.9044820070266724)
'rotfrukter', 0.8834930658340454)
'grönsaker', 0.8735833764076233)
'foder', 0.8698905110359192)
'djupfryst', 0.8694902658462524)
'köksväxter', 0.8692591190338135)
'socker', 0.867942214012146)
'mjöl', 0.8586941957473755)]
```

Figure 1: Ten nearest neighbours for the words *handikappad* and *potatis* for the sub-corpus of SOU:s covering the period 1970-79.

From these initial trials we concluded that the corpus, in spite of its noisiness, had sufficient quality to yield interesting results. However, we can not rely on absolute numbers, but the relative changes and differences that can be observed are sizeable enough to be trusted.

Another issue is that we discovered that there were some missing SOU-reports in the available files. This should be rectified in future releases of them.

For the studies on frequencies and word embeddings the texts were lowercased and stop words were removed. They were then grouped into decades. It was necessary to use this coarse granularity as reports covering topics related to disability are unevenly distributed over years.

[2] `https://spraakbanken.gu.se/swe/resurs/rd-sou#tabs=information` and `https://spraakbanken.gu.se/swe/resurs/sou#tabs=information`.

[3] The most recent version at the time of writing being published in July, 2017.

General terms: *handikap*, *handicap*, funktionsneds*, funktionshind*, funktionsvari*, funktionsnivå*, funktionsrätt*, *funkis*, *funkofobi*, funktionbegräns*, *delaktighet*
Contested / Group-specific terms: rörelsehind*, rullstolsbu*, vanför*, lam, lyt*, krympling*, invalid*, fallandesjuk*, svagsint*, andesvag*, vansinnig*, ofullkom*, sinneströg*, vanskapt*, imbecill*, kretin*, debil*, fåne, fånig, utvecklingsstör*, sinnesslö, obildbar*, idiot, efterbliven, döv, döf*, hörselskadad*, dövstum*, dumbe , blind, synskad*, dövblind, fallandesot, epilepsi
Vague terms: *abnorm*, halt, ofärdig, missbildad*, onormal*, avvikande, galen*, särskol*
partiellt arbetsför*, personer med, för alla, särskilda behov, särskilt stöd

Table 1: A list of Swedish terms and general words referring to disabilities. The asterisk indicates positions where the term may be extended.

3.2 Disability-related terms in Swedish

The term *handikapp* ('handicap') was introduced in the 1950s as an umbrella term for the many different terms that denoted special types of disability. *Handikappad* ('handicapped') was something a person was, but with the introduction of an environment related view, other words such as *funktionsnedsättning* ('functional impairment'), *funktionshinder* ('disability'), and *funktionsvariation* ('functional variation') were recommended. More recently these words, too, have been put into question, and a shift of attention to enabling measures has been proposed signalled by terms such as *delaktighet* ('participation'). These changes are not only replacements of forms but of (desirable) denotations and connotations.

3.3 Frequency changes

An initial list of some 60 words and phrases referring to disabilities and/or disabled persons over the last 100 years was manually produced by the disability researchers. See Table 1.

The list included scientific terms from disability studies, the recommended and contested terms of disability politics, and terms and general words for disabilities affecting specific capabilities, such as sight, hearing, and movement. From the initial experiments of finding nearest neighbours reported above, we found a few additional terms including adjectives such as *psykisk*, 'mental', which are common in multi-word disability terms. It also turned out that some of the words were too infrequent to be included in the study.

The words are either nouns or adjectives, which means that they occur in Swedish text in a variety of inflectional forms, up to eight for nouns, up to ten for adjectives. They also form derivatives and compounds. We have assumed that sharing of a common stem implies sharing a meaning[4]. This is a simplification, but does not prevent the discovery of general trends.

The multi-word terms were re-tokenized as single tokens before processing.

The words were grouped into three categories for the frequency studies:

- *Fysiskt handikapp* ('Physical handicap') using the Swedish terms *funktions**[5]*, funktionsneds*, handikap*, invalid*, lyte*/lytt*, rullstolsb*, rörelsehind*, vanför**
- *Psykisk sjukdom* ('Mental disability or illness') using the Swedish terms *mentalsjuk*, sinnessjuk*, dår[ae]*, gale?n*, psykisk rubbning*, psykisk sjukdom*, psykisk störning**
- *Förståndshandikapp* ('Mental retardation') using the Swedish terms *efterbliv*, förståndshandi*, idiot*, imbecill*, utvecklingsstör**

Thus, we are comparing relative frequencies within a cohort of terms assumed to cover roughly the same semantic space over a decade.

[4] For most words, the stem is identical to the look-up form in a Swedish dictionary. For some words, two forms are required due to stem variation, as in *galen* ('mad' singular) vs. *galna* ('mad' plural), *galning* ('mad person').

[5] * means that we use all linguistic forms

Figures 2, 3, and 4 show the frequencies of the various groups of terms. An immediate observation for all three groups is that there have been big changes. Dominant terms from the earliest periods have (almost) disappeared from current official discourse, and the terms that replaced them have given way to even newer terms. An exception to this general pattern is the term *handikapp* ('handicap') which is still current.

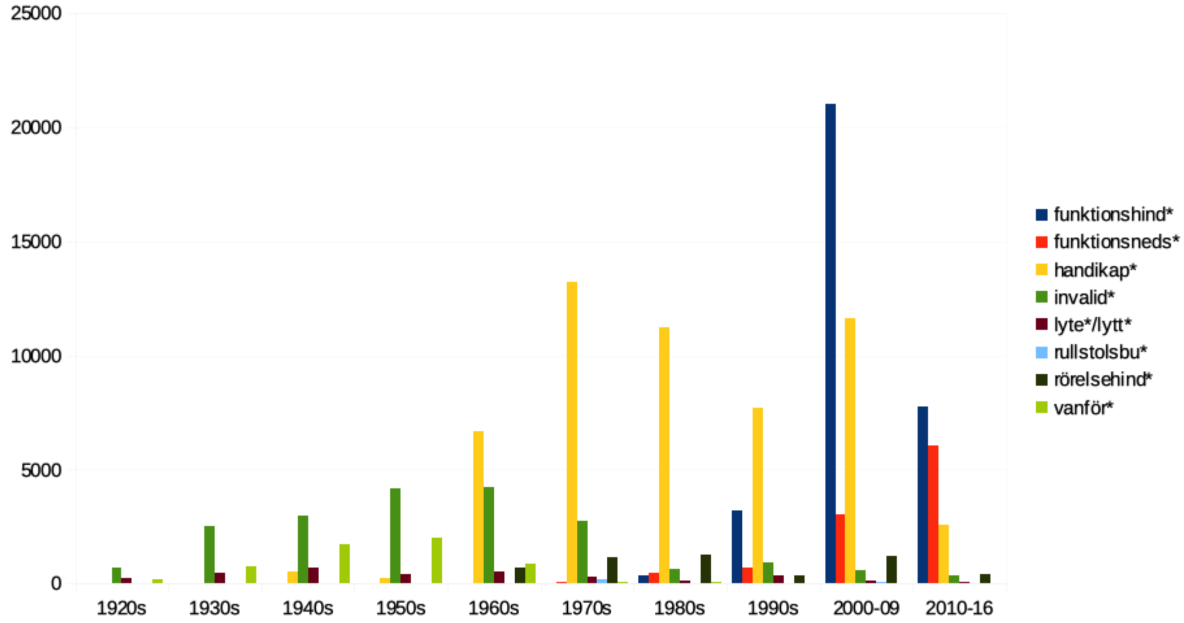


Figure 2: Usage of some Swedish Physical disability terms 1922-2016 by decades. Legend: *funktionshinder* ('disability'), *funktionsnedsättning* ('functional impairment'), *handikapp* ('handicap'), *invalid*, ('invalidity'), *lyte/lytt*, ('deformity', 'crippled'), *rullstolsburen*, ('carried by wheelchair'), *rörelsehindrad*, ('disabled wrt movement'), *vanför*, ('crippled, lame')

For instance, in the domain of Physical disability, Figure 2, we can see clear changes of dominant terms since the 1920:ies. Initially, during the period from 1920-1950 terms like *invaliditet/vanförhet* ('invalidity/lameness') dominate. Then from 1960 the term *handikapp* ('handicap') establishes itself and dominates until the term *funktionshinder/nedsättning* ('disability/functional impairment') is introduced around 2000. *Handikapp* does not disappear, however. Similar changes of dominant terms can be seen in relation to various forms of mental illness/retardation, Figures 3, and 4. For mental illness the total frequency of all terms has decreased.

The changes in dominant terms are largely in agreement with our expectations. However, while the change from *handikapp* ('handicap') to *funktionshinder/nedsättning* ('disability/functional impairment') is to be expected from the adoption of a relational model, it seems to happen quite slowly and with full force much later than one could expect. Also, the change is not abrupt. Thus, several disability models seem to be at work simultaneously.

3.4 Terms in context

By looking at the contexts in which a word is used we can gain an understanding of how people use it. We may use a concordancer or simply look at co-occurrences with words in the context. The Korp concordancer[6], which has a parsed version of the SOU-texts, can display neighbours with different grammatical relations to a word with their frequencies. Korp also enables the generation of concordances for the pair of context word and key word. See Figure 5.

Figure 5 shows that *handikapp* ('handicap') is often accompanied by attributes referring to extent: *svår*, ('hard, difficult'), *grav*, ('grave'), *allvarlig*, ('serious'), *lätt*, ('light'), or kind: *psykisk*, ('mental'),

[6] https://spraakbanken.gu.se/korp

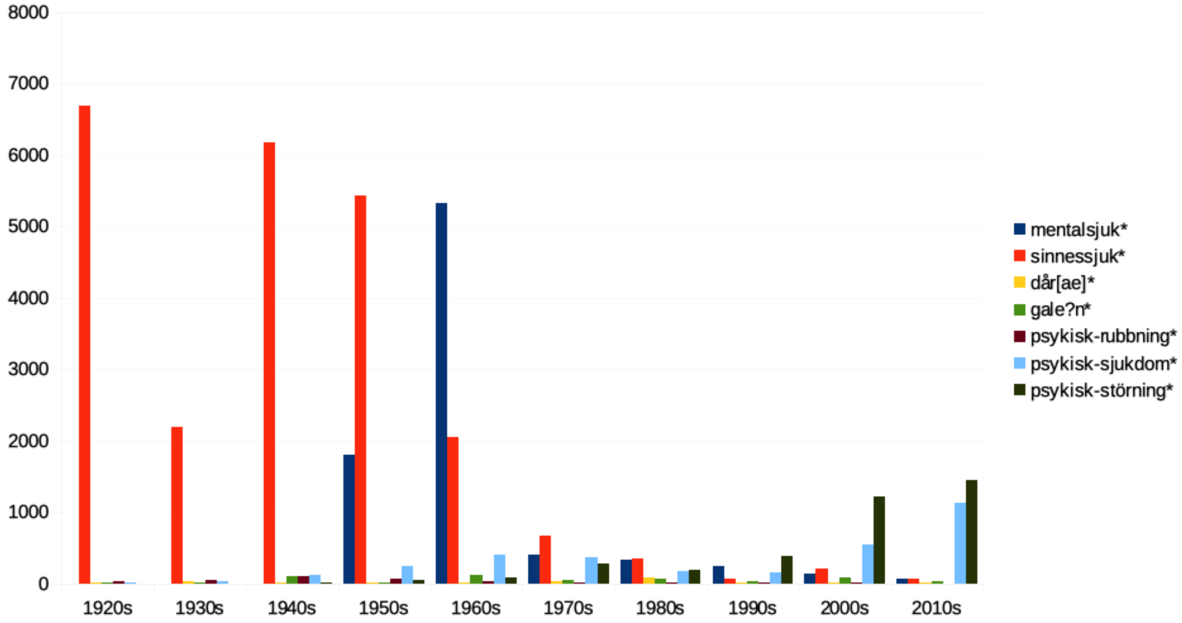


Figure 3: Usage of some Swedish *mental illness* terms 1922-2016 by decades. Legend: *mentalsjuk*, ('mentally ill'), *sinnessjuk*, ('insane'), *dåre*, ('lunatic'), *galen*, ('mad(ness)'), *psykisk-rubbning*, ('mental disorder'), *psykisk-sjukdom*, ('mental illness'), *psykisk-störning*, ('mental disorder')

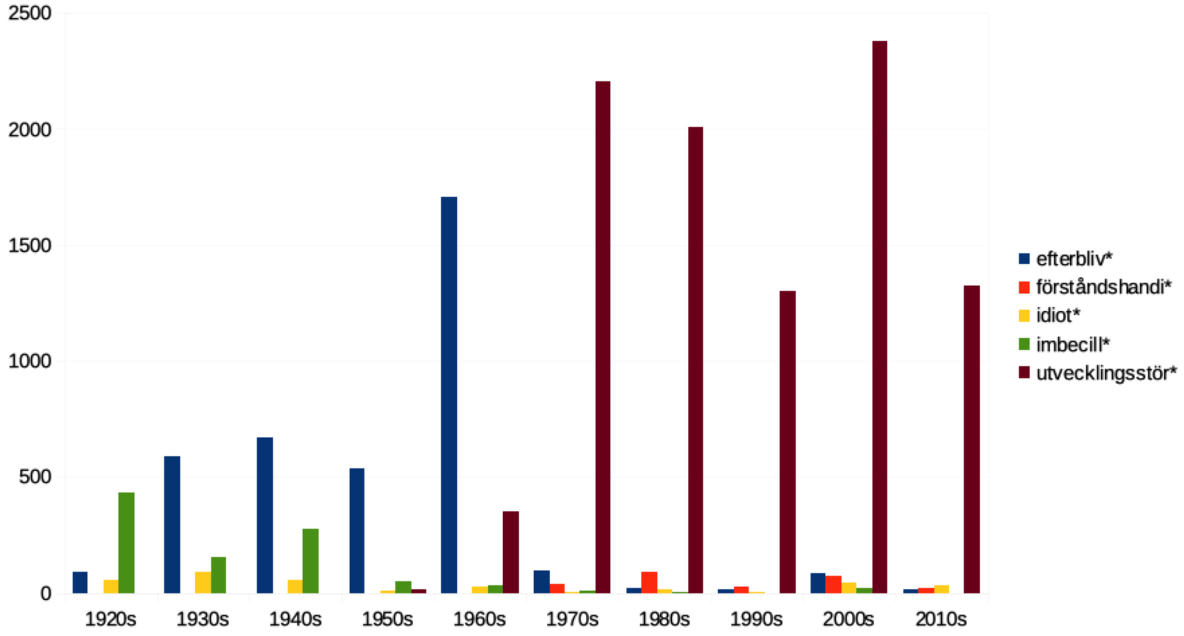
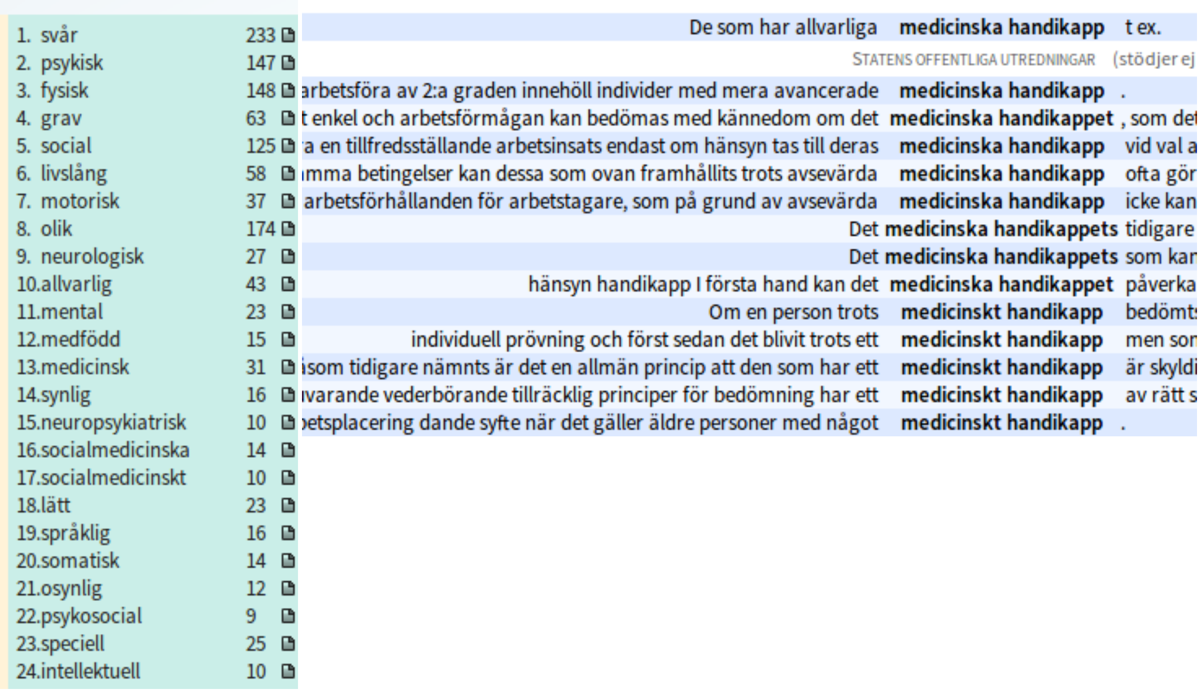


Figure 4: Usage of some Swedish *mental retardation* terms 1922-2016 by decades. Legend: *efterbliven*, ('retarded'), *förståndshandikappad*, ('intellectually handicapped'), *idiot*, ('idiot'), *imbecill*, ('imbecile'), *utvecklingsstörd*, ('mentally retarded')

fysisk, ('physical'), *neurologisk*, ('neurological'), *medfödd*, ('congenital'), *livslång*, ('life-long'). There are many overlaps with the corresponding lists for the words *sjukdom*, ('illness') and *funktionshinder* ('disability'). This clearly gives the impression that the medical model is well represented in the data.



1. svår	233
2. psykisk	147
3. fysisk	148
4. grav	63
5. social	125
6. livslång	58
7. motorisk	37
8. olik	174
9. neurologisk	27
10.allvarlig	43
11.mental	23
12.medfödd	15
13.medicinsk	31
14.synlig	16
15.neuropsykiatrisk	10
16.socialmedicinska	14
17.socialmedicinskt	10
18.lätt	23
19.språklig	16
20.somatisk	14
21.osynlig	12
22.psykosocial	9
23.speciell	25
24.intellektuell	10

De som har allvarliga	**medicinska handikapp**	t ex.
STATENS OFFENTLIGA UTREDNINGAR		(stödjer ej
arbetsföra av 2:a graden innehöll individer med mera avancerade	**medicinska handikapp**	.
t enkel och arbetsförmågan kan bedömas med kännedom om det	**medicinska handikappet**	, som det
a en tillfredsställande arbetsinsats endast om hänsyn tas till deras	**medicinska handikapp**	vid val a
ımma betingelser kan dessa som ovan framhållits trots avsevärda	**medicinska handikapp**	ofta gör
arbetsförhållanden för arbetstagare, som på grund av avsevärda	**medicinska handikapp**	icke kan
Det	**medicinska handikappets**	tidigare
Det	**medicinska handikappets**	som kan
hänsyn handikapp I första hand kan det	**medicinska handikappet**	påverka
Om en person trots	**medicinskt handikapp**	bedömts
individuell prövning och först sedan det blivit trots ett	**medicinskt handikapp**	men som
ısom tidigare nämnts är det en allmän princip att den som har ett	**medicinskt handikapp**	är skyldi
ıvarande vederbörande tillräcklig principer för bedömning har ett	**medicinskt handikapp**	av rätt s
ɔetsplacering dande syfte när det gäller äldre personer med något	**medicinskt handikapp**	.

Figure 5: Adjectival attributes of the noun *handikapp* ('handicap'), from 1 billion tokens in the Swedish Language Bank. Apart from the SOU-files all newspaper data and a corpus of novels have been included. The icons to the right of an attribute provide links to a KWIC concordance, where the search word occurs with this particular attribute.

3.5 Word embeddings

To obtain richer models we have trained word embeddings for the full corpus of SOU-reports, and for each decade. Given a sufficiently large corpus, learned vectors for frequent words give a good representation of their similarity. In the following, we focus on four Swedish words that have been used to try out the techniques. They are: *handikapp* ('handicap'), *funktionshinder* ('disability'), *funktionsnedsättning* ('functional impairment'), and *delaktighet* ('participation'). We used the SOU data from the 1970s and forward, split into decades as our corpus.

We used word2vec (Mikolov et al., 2013) with standard parameter settings as implemented in the GenSim framework[7]. Word2vec models are models created by a shallow, two-layer neural network that is trained on relevant data. As the training algorithm initializes the vectors with random numbers, ten models were trained for each decade, giving us 50 models altogether. This was done to check the stability of the models. The top three neighbouring word vectors for each term of interest was then checked in every model.

Some neighbours appeared in all ten models, but because of the random starting number there were also some variation. The found words were sorted on the basis of similarity and descending frequency.

Results are shown in Table 2. We could see a change in moving from the 1970:ies to the 1980:ies. For the terms *funktionshinder* ('disability'), and *funktionsnedsättning* ('functional impairment'), the term *handikapp* ('handicap') is one of the three closest neighbours only once in the period 1970-79. In the following decade 1980-89, *handikapp* ('handicap') is the closest neighbour for both terms. From this decade on these three terms are close neighbours in the word space. We could see from Figure 2 that the three terms are all used in the 1990:ies and later in the 2000:s, albeit with *handikapp* ('handicap') becoming the least frequent. Table 2 tells us, however, that there is close similarity in usage.

[7] `https://radimrehurek.com/gensim/index.html`

Term	1970-79	1980-89	1990-99	2000-16
handikapp	handikappade	handikappade	funktionshinder	funktionshinder
funktionshinder	funktions-rubbningar	handikapp	funktions-nedsättningar	funktions-nedsättning
funktionsnedsättning	psykiska	handikapp	funktionshinder	funktionshinder
delaktighet	gemenskap	gemenskap	jämlikhet	jämlikhet

Table 2: Nearest neighbours for selected disability terms in different decades

3.6 Temporal analogies and cross-decade projections

Vector spaces for the different decades were compared using the technique of temporal analogies (Szymanski, 2017). The method enables comparisons of one vector space to another by a global transformation or projection. Each ”early” model was paired with each ”later” model, e.g. a model from 1970 was paired with all other 1970 models and all models from later decades. After the transformation, cosine similarity was checked again in each model to see which words appeared. Most analogies were to the same word, but for some words mappings shifted to new words, as listed in Table 3.

Table 4 shows the closest analogy for the term *funktionshinder* (’disability’). The picture we got for this word by considering the neighbours in each decade, is confirmed. In the 1970:ies this term was used differently, analogously to words such as *sjukdomar*, (’illnesses’) in later years. From 1980 onwards it shows more affinity with the terms *handikapp* (’handicap’) and *funktionsnedsättning* (’functional impairment’).

Early	Later	Start term	New word	Nr of occurrences Max is 100
1970s	1980s	funktionshinder	sjukdomstillstånd	91
1970s	1980s	funktionsnedsättning	sjukdom	82
1970s	1990s	delaktighet	gemenskap	70
1970s	1990s	funktionshinder	sjukdomar	93
1970s	1990s	funktionsnedsättning	psykiska	95
1970s	2000s	funktionshinder	sjukdomar	79
1970s	2000s	funktionsnedsättning	sjukdom	98
1970s	2000s	handikapp	funktionsnedsättningar	96
1970s	2010s	funktionshinder	smärttillstånd	79
1970s	2010s	handikapp	funktionsnedsättningar	100
1980s	1990s	funktionshinder	handikapp	73
1980s	1990s	funktionsnedsättning	handikapp	100
1980s	2000s	funktionshinder	funktionsnedsättningar	98
1980s	2010s	funktionshinder	funktionsnedsättningar	100
1980s	2010s	handikapp	funktionsnedsättningar	93
2000s	2010s	funktionshinder	funktionsnedsättning	100

Table 3: Temporal analogies of selected terms. For pairs of periods not listed, the term was mapped onto itself.

4 Discussion

So, what can we learn from our analyses in relation to the two questions asked at the beginning?

As regards term usage, it is hard to find examples of overall terms being used in the early decades. Instead, various impairments are at work that affect a person in different ways. But as early as the 1940s we

	1970-79	1980-89	1990-99	2000-09	2010-16
1970-79	=	sjukdomstillstånd	sjukdomar	sjukdomar	smärttillstånd
1980-89		=	handikapp	funktions-nedsättning	funktions-nedsättning
1990-99			=	funktionshinder	funktions-nedsättning
2000-09				=	funktions-nedsättning

Table 4: Forward temporal analogies for the term *funktionshinder* ('disability').

see various umbrella terms present in the SOU-reports. Both *funktionsnedsättning* ('functional impairment') and *handikapp* ('handicap') are now being seen in the official discourses surrounding disability. But *handikapp* ('handicap') is still used only to a limited extent. These findings indicate that disability in the early period was interpreted foremost according to a medical model understanding of disability with the problem mainly located within the individual. However, it is hard, from this analysis, to decide whether or not this meaning was attached to the *handikapp* ('handicap') term as early as the 1940s.

However, the results raise the question whether some kind of conceptual change had started to occur and whether the SOU investigators had reasons to choose a new terminology when talking about the body. This, in turn, raises questions concerning how and to what extent the social model approach was preceded by the previous conceptual landscape that somehow helped prepare the way for the social model's emphasis on the societal dimension of disability, for instance by introducing an umbrella term that puts more focus on the collective aspect of a phenomenon. What we clearly see is how the introduction of *handikapp* ('handicap') has been broadly established during the 1970s. It is interesting that this change happens in parallel with the emergence of the social model perspective gaining ground in both disability theory and disability activism. The fact that the more relative, and in that sense more progressive, disability concept is being used in official reports suggests that the official discourse to some extent is affected by wider conceptual terminology changes within the society. The results indicate that disability was now being interpreted in another way and that it, broadly, had turned into a more collective phenomenon.

At the same time, the data show that those changes did not happen overnight, but that the process of change was rather slow and that old and new concepts lived side by side for a considerable amount of time. Another thing that is evident in the data is how other concepts become increasingly important concerning disability. Not only do we witness an increase when it comes to summary concepts that underline more of a social model and relativistic approach to disability, but we can also see how concepts linking it to social policies intentions takes more place in the SOU-reports.

As to our second question, what quantitative studies can contribute, it must be understood that in disability theory the concepts that are being used are very much linked to the overall philosophy or principle guiding the way a phenomenon is supposed to be understood and interpreted. In that way, the quantitative analysis not only shows how several vital concepts in relation to disability and the body are represented in a certain text, but how these concepts change over time, which indicates that the phenomenon of disability is being surrounded by changing ideas and interpretations. It is also interesting that various concepts live side by side and that we do not see a uniform language used in the governmental reports and the text corpus under scrutiny. This, in turn, might suggest that some uncertainty how to understand and describe disability, and how a message was supposed to be presented in relation to the logic of a certain text, has prevailed.

In that respect, it is also crucial to understand the meaning of an SOU-investigation. These texts were used not only to depict something but also to perhaps suggest reforms and to reason in relation to substantiate change in some area. Thus, concepts and terminology constituted an important part of that process.

What the data tells us is that disability represents a field in which there were limited consensus on what concepts to be used during those periods, which in turn suggests that representations of disability are movable and open and that they are characterised by the negotiating nature in which they are introduced, maintained and finally abandoned and replaced. This also touches on the notion of language seen from a constructionist point of view, in which language not only reflects reality but also helps to construct and shape it. The analysis conducted here thus suggests how disability and the body, as a phenomenon, are not given in terms of their meaning. The fact that there are different terms used during the same period indicates that the official discourse concerning disability had no fixed borders.

The approach can also assist in problematizing changes over time. Figure 2 shows how the individual based concept of *funktionsnedsättning* ('functional impairment') is highly present in the SOU-reports in the last twenty years and that it even, approaches the more environmentally related concept of *funktionshinder* ('disability') in frequency in the decade 2010-16. One question nourished by this outcome is whether this trend indicates some kind of reaction towards the social model thinking and relativistic understanding of disability, quite well established in the second half of the 1900s? This is just one example of how a quantitative content analysis can be a starting point for analysing discourses and understandings of disability in more depth.

5 Conclusions

Our analysis of the use of Swedish disability terms in resources made available by Swe-Clarin partners indicates that, while recommendations have effects, they seem to be delayed. Moreover, several frames of understanding disability live along side by side; especially the development and political use of different terms in a broader and official political context.

Most language technology methods used in this project utilise publicly available libraries, that we deliberately used without any parameter optimisation to illustrate what can be achieved ”out of the box”. Such optimisation, and more advanced preprocessing, would provide more reliable results, and will be carried out in future research projects on studying the *handikapp* ('handicap') concept. One obvious first improvement would be to lemmatize or stem the corpus to produce more stable word vectors. Further improvements include improved Optical Character Recognition of the SOU-reports used in the study. A comprehensive study would also require that access can be given to the full set of SOU-reports.

Using similarities in vector spaces, as a language technological tool, is a point of departure in analysing how a certain concept has been used in relation to its textual context. It underlines the fact that a certain word or concept never stands alone but derives, for instance, rhetorical meaning and content from its surroundings, which in turn can illustrate how a certain concept has changed its meaning content. The creation of the word2vec models presented above was based on central concepts linked not only to disability concepts as such but also to wider discourses guiding disability policy and the desired outcome of social policies.

From the results obtained, the conclusion can be drawn that quantitative analysis of even fairly noisy textual data can contribute with results that show a more complex picture of the history of disability concepts. This may in turn generate hypotheses to be investigated further by both quantitative and qualitative studies.

References

Handikappförbundens Centralkommitté. 1972. Ett samhälle för alla – princip- och handlingsprogram för handikappförbundens centralkommitté.

Elin Gardeström. 2006. *Handikapprörelsen och forskningen: en studie om relationerna utifrån handikapprörelsens perspektiv*. Sundbyberg: Handikappförbundens samarbetsorgan (HSO).

Marie Gustavsson Holmström. 2005. Den sociala modellen. In Bert Danermark, editor, *Sociologiska perspektiv på funktionshinder och handikapp*, pages 59–81. Lund: Studentlitteratur.

Lotta Holme. 2000. Begrepp om handikapp. en essä om det miljörelativa handikappbegreppet. In Magnus Tideman, editor, *Handikapp: synsätt, principer, perspektiv*, pages 67–78. Lund: Studentlitteratur.

Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Lars Lindberg and Lars Grönvik. 2011. *Funktionshinderspolitik: en introduktion*. Lund: Studentlitteratur.

Olov Andersson Marie-Louise Larsson-Severinsson and Magnus Tideman. 2009. Det relativa handikappbegreppets framväxt och etablering [elektronisk resurs].

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, , Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Franco Moretti. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

Mike Oliver. 2013. The social model of disability: thirty years on. *Disability & Society*, 28(7):1024–1026.

SOU 2019:23. 2019. Styrkraft i funktionshinderspolitiken.

Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the ACL (Short papers)*, pages 448–453.

Mårten Söder. 1982. Handikappbegreppet: en analys utifrån who:s terminologi och svensk debatt.

Mårten Söder. 2013. Swedish social disability research: a short version of a long story. *Scandinavian Journal of Disability Research*, 15(1):90–107.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *CoRR*, abs/1811.06278.

CLARIN Web Services for TEI-annotated Transcripts of Spoken Language

Bernhard Fisseni
Leibniz-Institut für Deutsche Sprache (IDS)
Mannheim, Germany
fisseni@ids-mannheim.de

Thomas Schmidt
Leibniz-Institut für Deutsche Sprache (IDS)
Mannheim, Germany
thomas.schmidt@ids-mannheim.de

Abstract

We present web services which implement a workflow for transcripts of spoken language following the TEI guidelines, in particular ISO 24624:2016 “Language resource management – Transcription of spoken language”. The web services are available at our website and will be available via the CLARIN infrastructure, including the Virtual Language Observatory and WebLicht.

1 Introduction / Recapitulation

Spoken language corpora are an important type of language resource as they exhibit many interesting aspects of language, such as language as a means of social interaction, dialectal, sociolectal or other forms of variation that are not covered by their written counterparts. Spoken language corpora require specialised processing methods. When dealing with audiovisual recordings, these methods can be based on speech technology; when dealing with the transcriptions of such recordings, methods can be used that are also applicable to written language data. The present paper focusses on the latter type of processing. It proposes elements of a workflow with documents in the TEI-based standard ISO 24624:2016 ”Language resource management – Transcription of spoken language” (henceforth **ISO/TEI**, see ISO 2016).

Schmidt, Hedeland and Jettka (2017) sketch, and partly implement, an architecture for making CLARIN webservices usable for transcriptions of spoken language, focusing on ISO/TEI as a pivot format on which web services operate and in which their output annotations can be represented. Schmidt, Hedeland and Jettka (2017) concentrates on a solution with an encoder/decoder pair which, at the entry point to a web service chain, transforms the ISO format to the Text Corpus Format TCF, which has been established as the basis for tool interoperability in WebLicht (see E. Hinrichs, M. Hinrichs and Zastrow 2010), and re-transforms the TCF result of the chain to ISO/TEI at the exit point. Since converters from common tool formats, such as those used by EXMARaLDA (see Schmidt and Wörner 2014), FOLKER (see Schmidt and Schütte 2010), ELAN (see Sloetjes 2014), Transcriber (see Barras et al. 2001) or CLAN/CHAT (see MacWhinney 2000), to ISO/TEI exist and can be prepended to the chain, a large class of language technology tools developed for written data thus becomes accessible to researchers working with spoken language while maintaining interoperability with tools which are commonly used for manual transcription and annotation of audiovisual material.

Schmidt, Hedeland and Jettka (2017) argue in the outlook tat CLARIN’s service-oriented approach could be further leveraged for spoken language data through the development of adequate services. These must take into account the specific characteristics of transcribed spoken data. Important features are the use of forms deviating from standard orthography and the fact that multilinguality is a much more frequent phenomenon in spoken data (see the third paragraph of section 3). Moreover, these services must be adapted to the specific tasks arising in the curation of oral corpora, such as the alignment between transcript and audio. These services could operate directly on the ISO format, which provides features to cater for the aforementioned features of spoken corpora, without a ‘detour’ to TCF. The work reported in the present contribution explores this option further, first, by portraying a use case that typically arises in the curation of interview data (sec. 3), second, by sketching elements of a workflow suitable for this and related use cases and describing details of a CLARIN-conformant implementation of this workflow (sec. 4).

Bernhard Fisseni and Thomas Schmidt 2020. CLARIN Web Services for TEI-annotated Transcripts of Spoken Language. *Selected papers from the CLARIN Annual Conference 2019.* Linköping Electronic Conference Proceedings 172: 172 12–22.

2 Related Work

Workflows for the curation of interview data have been discussed in the CLARIN context on the occasion of several workshops on Oral History data, whose results are documented on a dedicated website (`https://oralhistory.eu/`). The focus of this work is on the use of speech technology (e.g. automatic speech recognition, forced alignment) which operates on the audio signal. By contrast, the current paper concentrates on tools for enriching textual transcription data with language technology. Ideally these two approaches should be combined to complement each other.

Several methods described here were originally developed in the context of the EXMARaLDA system (Schmidt and Wörner 2014), as part of the workflow for compiling the Research and Teaching Corpus of Spoken German (FOLK, see Schmidt 2016) and/or as components of curation workflows at the CLARIN-D B-centres Hamburg Center for Language Corpora (*Hamburger Zentrum für Sprachkorpora*, HZSK)[1] and the Archive for Spoken German at IDS (*Archiv für gesprochenes Deutsch*, AGD, Schmidt 2017).[2] Details on the development of the POS tagging model are described by Westpfahl (2020). Several of the services described in sections 4 reuse and extend these methods (at least conceptually) and put them on a different technological basis thereby integrating them more fully into the CLARIN infrastructure.

Besides Schmidt, Hedeland and Jettka (2017) and the ISO specification itself (ISO 2016), the role of TEI as a suitable basis of a standard for spoken language transcription has been discussed, among others, by Schmidt (2011) and Liégeois et al. (2017). The TEI guidelines' chapter 8 on "Transcriptions of Speech" (TEI Consortium 2019) has also been used in CLARIN resources such as the GOS Corpus of Spoken Slovene (see Verdonik et al. 2013) and as the basis for a CLARIN-wide format for parliamentary data.[3]

3 Use case: Legacy interview corpora

The Archive for Spoken German (*Archiv für Gesprochenes Deutsch*, AGD) at the Leibniz Institute for the German Language is a central point for depositing, archiving, and disseminating corpora of spoken German. AGD hosts more than 80 spoken language corpora with more than 10,000 hours of audio or video recordings. The archive's stock is increasing continuously through internal corpus compilation projects, through collaborations with external partners and through data deposits by completed projects. A substantial part of the archive's work goes into curating such external resources, i.e. putting audio/video recordings, metadata, transcripts and annotations into a state where they can be archived, discovered (= found) and reused (thus conforming to the FAIR principles, cf. Wilkinson et al. 2016), among others through CLARIN services like the Virtual Language Observatory (VLO). The data types which the AGD deals with can be roughly divided into three classes:

(1) *interaction corpora* which document language in interaction (e.g. the FOLK corpus, Schmidt 2016),

(2) *variation corpora* which deal with language variation (dialects, regiolects, etc.) within the German-speaking core countries (e.g. *Deutsch Heute*, Brinckmann et al. 2008) and in German language communities around the world (e.g. Australian German, Lich and Clyne 1984) and

(3) *interview corpora*, which consist of relatively free (mostly narrative or biographic) interviews with selected speaker groups and/or on specific topics. In the present paper, we would like to focus on this corpus type.

Examples for already curated interview corpora at the AGD are Norbert Dittmar's *Berliner Wendekorpus* (see Schmidt 2019)[4] in which speakers from East and West Berlin were asked to relate their personal experiences with the fall of the Berlin wall, Anne Betten's extensive data on German-speaking emigrants to Israel (see Betten 1995),[5] or a recent interview study by Serap Devran (see Devran 2017) which deals with people of Turkish descent who (re)migrated to Turkey from Germany (available in the DGD since January 2020 since January 2020). It should be pointed out that ("language-biographic") interviews are also often an integral part of variation corpora. It is also worth noting that multilingualism plays a central role for a substantial part of these data because the respective studies focus on speakers with migration histories and their specific language varieties which often include code switching or mixing.

[1] see `https://corpora.uni-hamburg.de/`

[2] see `http://agd.ids-mannheim.de/`

[3] see `https://www.clarin.eu/event/2019/parlaformat-workshop`

[4] see `http://hdl.handle.net/10932/00-0332-BD7C-3EF5-0B01-4`, `http://agd.ids-mannheim.de/BW--_extern.shtml`

[5] see `http://hdl.handle.net/10932/00-0332-C3A7-393A-8A01-3`, `http://agd.ids-mannheim.de/ISW-_extern.shtml`

While they cover a wide range of topics, these data have a lot in common in terms of methodology (all of them are semi-structured, narrative or biographic interviews with little interviewer intervention and a high degree of spontaneity) and in terms of structural and technical properties (typically audio recordings in quiet environments with durations up to three or four hours, rich biographic metadata on the speakers, orthographic transcriptions). They also share a high potential for interdisciplinary reuse, mostly because, beyond documenting specific linguistic forms, their contents also make them a valuable source for sociological or oral history studies.

The AGD has recently acquired, or is in the process of acquiring, further such interview corpora. Among them are the 2800 hours of audio recordings from the Bonn Longitudinal study of Aging (*Bonner gerontologische Längsschnittstudie*, see Lehr and Thomae 1987) and an interview study on German refugees in Britain ("Kindertransporte", see Thüne 2019). Other projects or data centres (outside of CLARIN) in Germany dealing with similar data are *Zwangsarbeit-Archiv* at the *Center für Digitale Systeme* (CeDis) in Berlin[6], *Archiv „Deutsches Gedächtnis"* at *FernUniversität in Hagen* (the German distance-learning university)[7] and *QualiService Bremen*[8], a centre for qualitative research data in the social sciences.

Typically, when data from interview studies are deposited at the AGD, they consist of the audio recordings (digitised or not), transcripts in modified orthography (e.g. "zwohunnert" for a spoken form of standard "zweihundert" ('two hundred') ; "dunno" for "don't know" would be a similar example in English) in some word processor format, and more or less structured metadata on interviews and interviewees in spreadsheet, text files, or similar formats. The AGD curates such data with the aims of:

(a) fully digitising the resource, especially the primary audio or video data,

(b) transforming all textual data into structured, machine-readable, interoperable formats which conform to current best practices and/or standards (e.g. for transcripts ISO/TEI; for metadata CMDI, cf. Broeder et al. 2012; CLARIN ERIC 2019),

(c) interconnecting the different data types (e.g. aligning transcripts with recordings, referencing between primary, secondary and metadata),

(d) enriching the data with further information useful for linguistic analysis (e.g. lemmatisation, POS tagging) and

(e) integrating them into the Database for Spoken German (DGD, `https://dgd.ids-mannheim.de`) as the primary dissemination platform and

(f) integrating them into the wider language resource infrastructure through the institute's long term archiving repository thereby associating datasets with PIDs and making metadata available through OAI/PMH.

The workflows needed for transforming the deposited source data into the desired target state may differ with the details of the individual resource, but experience has shown that they are always made up from the same set of building blocks. It is these building blocks that we propose to implement in a set of ISO/TEI-based, CLARIN-conformant web services and which we will describe in more detail in the following sections. While we illustrate them on a specific piece of data from a specific data centre, we argue that the same methods and tools, if properly configured, will be useful and applicable in a much wider range of contexts.

For example, the output of many transcription tools such as ELAN, EXMARaLDA, FOLKER, Transcriber, CLAN or Praat can be directly transformed to ISO/TEI with the help of suitable conversion tools (such as the web services described by Schmidt, Hedeland and Jettka 2017). Skipping the first step in the toolchain, viz. plain text to ISO/TEI conversion, the tools described here can be used on these data, thus applying to a much wider range of TEI documents than those deriving from our use case.

Furthermore, not all methods described here require the input to be fully conformant ISO/TEI transcripts. The language detection service, for example, can be applied to any data which uses the TEI <u> (*utterance*)[9] in a meaningful way as an annotation. Likewise, the only prerequisite for lemmatization and POS-tagging is that texts have been tokenized with TEI <w> (*word*)[10] elements.

[6] see `https://www.zwangsarbeit-archiv.de/index.html`

[7] see `https://deutsches-gedaechtnis.fernuni-hagen.de/`

[8] see `https://www.qualiservice.org/`

[9] see `https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-u.html`

[10] see `https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-w.html`

Normalisation can also be applied to any <w>-level annotated texts, but it is probably only useful in cases of playful writing, e.g. in computer-mediated communication. Even then, it would probably be preferable to make a new dictionary of normalisation and capitalised-only words.

We use the following example from the *Corpus Australian German* (see Clyne 1981; Kipp 2002)[11], gathered by the Australian linguist Michael Clyne in the 1960s and deposited at the AGD in 2014, throughout the text, and show excerpts from the results of step in the toolchain. The backslash indicates that a line is only broken for typesetting purposes.

```
MC: Welche Früchte ham sie (.) hier in der (-) Gegend?
AS: Äh, Apfel.
    Apfel, Birnen, äh, Pflaumen, etwas Feigen, nich su viel und äh, dann hat \
    man auch äh Aprikosen, sehr viel Aprikosen und auch Pfirsiche, ja, \
    und äh, Mandeln sind auch sehr viel vorhanden.
    Mandeln tun eigentlich ganz gut hier.
MC: Und ähm vielleicht könnten wir n bisschen umschalten ins Englische.
    What part of Germany did your forefathers come from?
AS: Eh, our people came from what they call Schlesien.
    I wouldn't know how you pronounce that in English.
```

4 Workflow and Tools

We provide an abstract description of the functionality of the services and an explanation of the motivation and challenges for each step.[12]

The process is conceived of as a pipeline, so that the output of one step can immediately serve as input to the next step. We will also mention some parameters, but we have to refer the reader to the documentation for a detailed description.

All services can be given a default language which will be used if the language of the document cannot be otherwise determined. Contrary to the approach in TCF, ISO/TEI documents, and TEI documents in general, inherently support multilingual texts, that is: not only can a language be specified for the text as a whole, but individual components (here: utterances or words) can be assigned differing language tags.

4.1 Plain text to ISO/TEI-annotated texts (`text2iso`)

As detailed in sec. 3, our use case of legacy corpora starts with documents in word processor format. As we can disregard most of the formatting, we expect input in plain text format for our web services. Hence the first step is to convert plain text transcribed data to a ISO/TEI-conformant format, which serves as input for all further processing steps.[13]

In this step, the main challenge was specifying a plain text input format that is sufficiently expressive to serve in the most common cases of transcriptions that will be subject to the workflow, as outlined above, and sufficiently simple and restricted to be typed and parsed efficiently. Conventions should also be as close as possible to those typically used in the text submitted to the AGD. The latter point was a reason to exclude existing formats such as CHAT. The format is supposed to allow segmentation of the conversation into utterances, assignment of these utterances to speakers. A specification is available at `https://github.com/Exmaralda-Org/teispeechtools/blob/master/doc/Simple-EXMARaLDA.md`.

Building on previous work, we spelt out some restrictions and corner cases and specified a formal language which can be deterministically parsed. Parsing was implemented using the ANTLR 4 parser generator.[14] The format manages simple forms of overlap between utterances as well as the annotation of nonverbal actions accompanying or stepping in for verbal actions. As we did not want to add too much explicit markup, we had to specify limitations with respect to overlap handling. As can be seen from the example, overlaps are indicated by marks occurring in the text. The restriction is that such marks can occur freely in the first utterance containing them, but to avoid complicated temporal alignment structures that might turn contradictory, marks must occur at the beginning of later utterances referencing them.

The result of this step is a transcription file which is split into utterances: an <annotationBlock> for each utterance contains a <u> element as well as <incident> elements containing non-verbal actions and

[11] see `http://hdl.handle.net/10932/00-0332-BCFF-D7B3-7A01-9`, AD--_E_00010

[12] The web services are available at `http://clarin.ids-mannheim.de/teilicht`.

[13] For intergration into WebLicht, see sec. 4.8, `text2iso` and `segmentize` were combined into a service `text2seg`, which takes all the parameters of these services.

[14] see `https://www.antlr.org/`

<spanGrp> elements containing commentaries. A <timeline> is derived from the text, and all annotation is situated with respect to the <timeline>. Elements of the timeline are the beginning and end of each utterance; in case of overlap, the overlap start and end is referenced as an <anchor> within the utterances.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <profileDesc><particDesc>
         <person n="AS" xml:id="AS"><persName><abbr>AS</abbr></persName></person>
         <person n="MC" xml:id="MC"><persName><abbr>MC</abbr></persName></person>
      </particDesc></profileDesc>
      <encodingDesc>...</encodingDesc> <revisionDesc>...</revisionDesc>
   </teiHeader>
   <text xml:lang="de">
      <timeline unit="ORDER">
         <when xml:id="B_1"/> <when xml:id="E_1"/>
         <when xml:id="B_2"/> <when xml:id="E_2"/>
         <when xml:id="B_3"/> <when xml:id="E_3"/>
         <when xml:id="B_4"/> <when xml:id="E_4"/>
         <when xml:id="B_5"/> <when xml:id="E_5"/>
         <when xml:id="B_6"/> <when xml:id="E_6"/>
         <when xml:id="B_7"/> <when xml:id="E_7"/>
         <when xml:id="B_8"/> <when xml:id="E_8"/>
      </timeline>
      <body>
         <annotationBlock start="B_1" end="E_1" who="MC">
            <u>Welche Früchte ham sie (.) hier in der (..) Gegend?</u>
         </annotationBlock>
         <annotationBlock start="B_2" end="E_2" start="B_2" who="AS">
            <u>Äh, Apfel.</u>
         </annotationBlock>
         ...
         <annotationBlock start="B_8" end="E_8" who="AS">
            <u>I wouldn't know how you pronounce that in English.</u>
         </annotationBlock>
      </body>
   </text>
</TEI>
```

4.2 Segmentation according to transcription convention (segmentize)

In the next step, the text is segmented according to transcription conventions. Again, this is implemented deterministically by processing a formal language. We enforce a a tokenisation into words in TEI <w> elements and punctuation in TEI <pc>, and some information is lifted from the plain text of an <u> to the annotation level, mainly pauses (encoded as TEI <pause> with a @type attribute) and unclear or incomprehensible text. The most adequate transcription level for the paradigmatic workflow is the *generic* transcription level, which provides these basic features.[15] More advanced transcription levels follow cGAT conventions.[16]

ISO/TEI allows to use time <anchor> elements also in the middle of words. Keeping the <anchor>s in place while processing the surrounding plain text was one of the challenges of implementing this step, as in this case, XML structure interferes with the abstract structure of the transcription.

```
<annotationBlock start="B_1" end="E_1" who="MC"><u>
  <w>Welche</w> <w>Früchte</w> <w>ham</w> <w>sie</w> <pause type="micro"/>
  <w>hier</w> <w>in</w> <w>der</w> <pause type="short"/>
  <w>Gegend</w> <pc>?</pc>
</u></annotationBlock>
```

4.3 Language detection (guess)

Language detection is an addition to the workflow implemented in EXMARaLDA up to now. The motivation for this step is that, as mentioned in sec. 3, data are often massively multilingual, and it is useful

[15] The specification is available at https://github.com/Exmaralda-Org/teispeechtools/blob/master/doc/Generic-Conventions.md

[16] see http://agd.ids-mannheim.de/gat.shtml

to be able to assign languages to single utterances. In contrast to TCF, the TEI formats allow @xml:lang to specify a language on every structural level of text. We leverage this attribute to annotate language changes.

The service uses the Apache OpenNLP[17] language models and language detector to process single utterances and guess what language they are in. It is possible to constrain the search space to a set of languages to avoid mis-detection of similar languages like German and Low German; the default is German, Turkish and English. Language detection quality deteriorates if too little linguistic material is available, and if the transcription deviates too much from standard orthography. Therefore, a configurable threshold (default: 5 words) can be set to prevent potentially unreliable language detection in utterances that are too short.

In the result, the <u> have been annotated with @xml:lang attributes where the algorithm[18] reached a decision. If languages are equally probable, the document language is preferred. Cases of doubt are reported in XML comments; for debugging purposes, we also report probabilities for the expected languages here.

```
<annotationBlock start="B_5" end="E_5" who="MC">
  <!--deu: 0,07; eng: 0,01; tur: 0,01--><u xml:lang="de">
     <w>Und</w> <w>ähm</w> <w>vielleicht</w> <w>könnten</w> <w>wir</w> ...
  </u>
</annotationBlock>
<annotationBlock start="B_6" end="E_6" who="MC">
  <!--eng: 0,05; deu: 0,01; tur: 0,01--><u xml:lang="en">
      <w>What</w> <w>part</w> <w>of</w> <w>Germany</w> <w>did</w> ...
  </u>
</annotationBlock>
```

The following steps depend on correct language classification, and can hence be facilitated by manual language annotation or by applying guess before they are executed.

4.4 OrthoNormal-like Normalisation (normalize)

EXMARaLDA includes the OrthoNormal tool for transcript normalisation, i.e. the mapping of tokens in modified orthography to their standard orthographic equivalent, e.g. “zwohunnert” to “zweihundert”, “kannste” to “kannst Du”, “hab isch net” to “habe ich nicht”, but also nouns which are non-capitalized according to the transcription convention, but capitalized according to standard orthography (“haus” to “Haus”).

The automated part of normalisation is dictionary-based and only available for German at the moment (see Schmidt 2012). We plan to experiment with other algorithms or languages in the future.

Normalisation works on the <w> elements, which are annotated with a @norm attribute containing the normalised form. The algorithm can be summarised as follows:

1. The most frequent normalisation for a word form in the FOLK corpus is applied.
2. If nothing is found in Step 1, the list of words that occur capitalized-only in DeReKo[19] is consulted and a normalisation is chosen.
3. Out-of-dictionary words are left as is.

On the FOLK corpus, this automatic procedure yields correct normalisations for 93% of all tokens, and for 83% of tokens which require normalisation. Since the FOLK corpus is relatively large and contains data from diverse regions and settings, we can expect the procedure to perform with similar quality on interview data.

```
<annotationBlock start="B_1" end="E_1" who="MC"><u xml:lang="de">
      <w norm="welche">Welche</w> <w norm="Früchte">Früchte</w>
      <w norm="haben">ham</w> <w norm="sie">sie</w>
      <pause type="micro"/>
      <w norm="hier">hier</w> <w norm="in">in</w> <w norm="der">der</w>
      <pause type="short"/> <w norm="Gegend">Gegend</w>
      <pc>?</pc>
</u></annotationBlock>
```

[17] see https://opennlp.apache.org/

[18] see https://opennlp.apache.org/docs/1.9.0/manual/opennlp.html#tools.langdetect

[19] *Deutsches Referenzkorpus*, see http://www1.ids-mannheim.de/kl/projekte/korpora.html

4.5 POS-Tagging with the TreeTagger (`pos`)

POS-tagging and lemmatisation are preferrably done after normalisation, since a notably higher precision is achieved when the tagger is fed normalised forms instead of forms in modified orthography (see Westpfahl 2020). However, it is not a requirement for this step that transcripts be normalised. We use the TreeTagger by Helmut Schmid (1995) for POS tagging, employing the Java wrapper TT4J by Richard Eckart de Castilho.[20]

We use the standard tagging models provided by the TreeTagger, which were mostly trained on and intended for written language. Tagging models trained on and intended for spoken language exist for French and for German (Westpfahl 2020). As Westpfahl (2020) shows for German, tagging models trained on spoken language data and with tag sets optimised for this resource type will yield significantly lower error rates (around 5% as compared to 15%–20%) than tagging models which have not been adapted to this task.[21]

Respecting the language of the current word `<w>`, the correct parser model is chosen by language, and the `@pos` and `@lemma` attributes are set accordingly.

```
<annotationBlock start="B_5" end="E_5" who="MC"><u xml:lang="de">
  <w lemma="und" norm="und" pos="KON">Und</w>
  ...
  <w lemma="in" norm="ins" pos="APPRART">ins</w>
  <w lemma="Englische" norm="englische" pos="NN">Englische</w>
  <pc>.</pc>
</u></annotationBlock>
<annotationBlock start="B_6" end="E_6" who="MC"><u xml:lang="en">
  <w lemma="what" pos="DTQ">What</w> ... <w lemma="come" pos="VVB">come</w>
  <w lemma="from" pos="PRP">from</w> <pc>?</pc>
</u></annotationBlock>
```

Note how in our example, this results in different tag sets being used for `<u>` elements in different languages.

4.6 Pseudo-alignment using Phonetic Transcription or Orthographic Information (`align`)

Another addition to the EXMARaLDA workflow is pseudo-alignment between transcription and recordings using graphemic or phonemic information. Most of the data submitted to the paradigmatic workflow do not contain information on the time when utterances occurred.

A logical step would be to apply *forced alignment* on these. Forced alignment is a speech processing technique that fits a given segmentation, in our case, the transcription, to a speech signal. Several aligners exist; for German, one of the most easily accessible and prominent ones, WebMAUS, is provided by the Bavarian Archive for Speech Signals (BAS), as part of their web services (Kisler, Reichel and Schiel 2017; Draxler, Harrington and Schiel 2017).[22]

We have been experimenting successfully with integrating WebMAUS into our workflow. However, we also found that it would be useful to have an alternative, as in many cases, data cannot be sent to web services such as those provided by the BAS, for three possible reasons. We report on our experiments with the BAS services, in particular. First, often the audio quality is insufficient for speech processing. Secondly, data may be too sensitive to transmit to external services which cannot guarantee encrypted storage; this applies to many of the corpora that can only be made accessible under restricted conditions. Thirdly, BAS web services often have problems if recordings are too long (where the limit may be as short as ten minutes) or contain certain features such as long pauses; both tend to occur in the data relevant to our current use case.

The processing performed in this step permits to estimate the alignment between temporal data (sound, video) and transcriptions relying on the graph(em)ic form of utterances, i.e. counting letters, or the canonical phonetic transcriptions provided by BAS web services, counting phone(me)s. Optionally, the canonical phonetic transcription can be added to the TEI-ISO document using the attribute `@phon` on `<w>`

[20] see `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/` and `https://reckart.github.io/tt4j/`, respectively.

[21] Unfortunately, we cannot give a general figure of accuracy for POS tagging in this context, but have to refer to general papers such as those by Schmid (1995) or Giesbrecht and Evert (2009). We would very much welcome the development of models for spoken language for languages other than German and French.

[22] see `https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface`

elements. As there is as yet no good TEI attribute for this[23], we use a non-standard attribute @phon, as for practical reasons, we try to avoid extra <spanGrp> elements for simple annotations.

The alignment thus achieved can be manually improved, if necessary.

Graph(eme)-based measures are useful because graph(eme)-to-phone(me) conversion is not available for every language, because of difficulties with language tagging ambiguity (see next paragraph) and because of the fact that timeline <anchor>s can occur even in words, and then it is impossible to determine the phone boundary. For instance, if an overlap (marked |) starts in the middle of the word "psycho|logist", it is difficult to guess the correct breaking point. The algorithm regresses to counting letters instead.

A difficulty arises with respect to language handling, as for some languages, BAS web services use fully qualified locales as parameters. The service will do some adjustment to be able to transcribe (e.g., accept ltz and not just the full ltz-LU for Luxemburgish).

```
<timeline><when id="T0" interval="0.0s" since="T0"/>
  <when xml:id="B_2" interval="5.394s" since="T0"/>
  <when xml:id="E_2" interval="6.356" since="T0"/> ... </timeline>
<body>
  <annotationBlock end="E_2" start="B_2" who="AS"><u start="B_2" end="E_2">
    <w lemma="Äh" norm="äh" phon="ʔɛː" pos="ADJA">Äh</w> <pc>,</pc>
    <w lemma="Apfel" norm="Apfel" phon="ʔap.fəl" pos="NN">Apfel</w> <pc>.</pc>
  </u></annotationBlock> ...
```

Starting with guess, all steps are based on heuristics and NLP. Therefore, the results of these steps should be

4.7 Adressable elements (identify and unidentify)

There are two more services, which are only useful in specific cases. Occasionally, it is useful if all structural elements can be addessed with an @xml:id attribute. Hence, identify adds @xml:id attributes to all TEI elements that do not have one, and @unidentify removes such attributes whose form suggests they have been added by identify.

4.8 Integration with WebLicht: Parameters and an Optional Header

WebLicht (E. Hinrichs, M. Hinrichs and Zastrow 2010)[24] has proven successful, especially as a didactic and explorative environment for running webservices for linguistic annotation, and it is an important part of the CLARIN infrastructure. WebLicht's architecture is built around the pivot format TCF (see E. Hinrichs, M. Hinrichs and Zastrow 2010) which is currently in ins fifth version.[25]

The TEILicht services have been integrated into WebLicht and will be integrated into the Language Resource Switchboard (https://switchboard.clarin.eu/). The WebLicht team provided much help, and also a new input type for the plain text transcripts. The integration was not seamless, however. At this point in time, WebLicht's requirement to explicitly list all possible values for a given parameter poses problems for parameters with a large or continuous value set (such as languages, audio duration etc.).

Let us consider the case of language tags at some length. Of course, in the most common cases, a simple language code such as de or nl-BE may be sufficient. However, the language codes suggested by BCP 47, recommended by the TEI guidelines[26], are actually an open class and allow for impromptu tags like de-DE-x-goethe (example taken from BCP 47, page 10). Moreover, even offering a full list of languages that can be selected with two or three letter codes, and even more so offering to select several from this list is problematic. Therefore, the guess webservice was modified to accept four one-language identifiers expected1 to expected4. These are used in addition to the list-valued expected parameter to restrict the search space of languages.

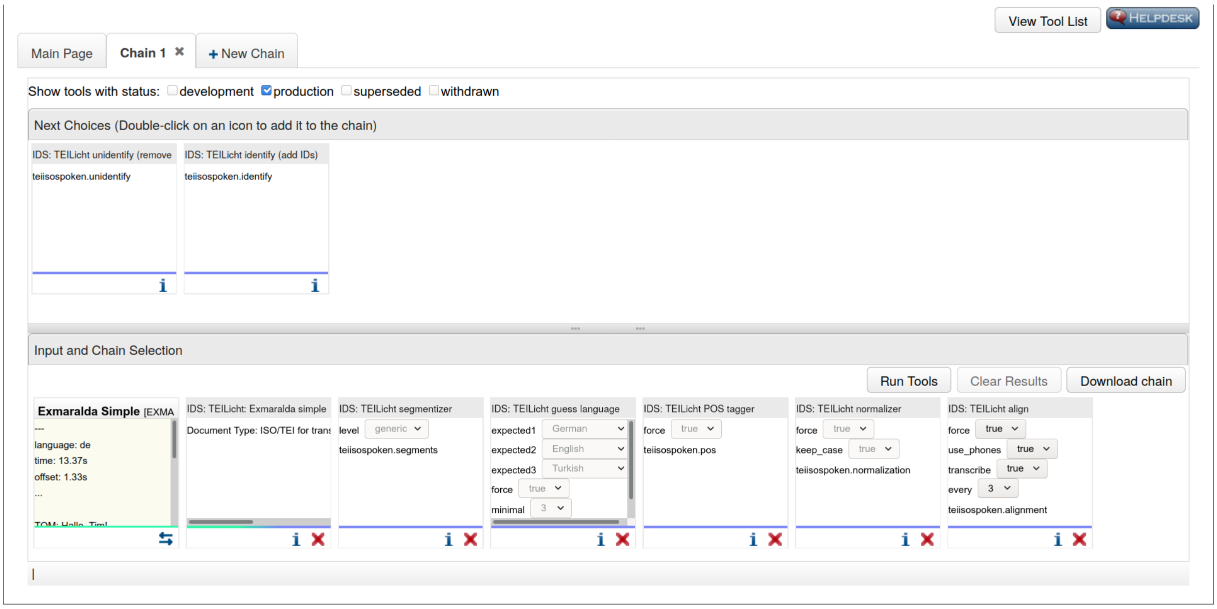
The only way to proceed with values such as positive integers (e.g., the every parameter of align for inserting a time anchor every n words) is to offer a snsible choice of values (e.g. 3, 5 and 10). For values such as positive floating point numbers such as the duration of a transcript or the offset parameter of the align service, it is not possible to find a good representation in WebLicht. However, these are important

[23] The integration of @phon has been submitted as a request to TEI.

[24] see https://weblicht.sfs.uni-tuebingen.de/

[25] The current specification is avaiblable at https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format.

[26] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-teidata.language.html


View Tool List
HELPDESK
Main Page
Chain 1 ✖
+ New Chain
Show tools with status: □ development ☑ production □ superseded □ withdrawn
Next Choices (Double-click on an icon to add it to the chain)
Input and Chain Selection
Run Tools
Clear Results
Download chain

IDS: TEILicht unidentify (remove
teiisospoken.unidentify

IDS: TEILicht identify (add IDs)
teiisospoken.identify

Exmaralda Simple [EXMA
---
language: de
time: 13.37s
offset: 1.33s
...
TOM: Hallo, Tim!

IDS: TEILicht: Exmaralda simple
Document Type: ISO/TEI for trans

IDS: TEILicht segmentizer
level: generic
teiisospoken.segments

IDS: TEILicht guess language
expected1: German
expected2: English
expected3: Turkish
force: true
minimal: 3

IDS: TEILicht POS tagger
force: true
teiisospoken.pos

IDS: TEILicht normalizer
force: true
keep_case: true
teiisospoken.normalization

IDS: TEILicht align
force: true
use_phones: true
transcribe: true
every: 3
teiisospoken.alignment

Figure 1: WebLicht integration, example chain combining most services in a sensible order

for the correct operation of the `align` service. To allow the processing, the plain text format was enhanced with an optional header where one can specify the main document language, the duration until end of the last utterance, and the offset of the first utterance can be specified, the example below. All specifications are optional, but an offset is only accepted if the duration is also specified.

```
---
lang: de
duration: 43s
offset: 0.0
---
```

4.9 Command line version and web services

The functionality of all web services is also available as a Java library and command line tool, see `https://github.com/Exmaralda-Org/teispeechtools/`. On bulk data, the command line tool is easier to use than the web services. Moreover, the command line invocation is generally free of privacy concerns, as no data are sent through the web.[27]

The commands for the command line tools have the same name as the services described above.

The parameters of the web service and the command line version generally have the same name, e.g. `--lang` (or `--language`) on the command line and `lang` in the web services, but with dashes swapped for underscores, e.g. `--minimal-length` (alternatively, `--minimal`) on the command line and `minimal_length` in the web service, and some shorter option names provided for the command line tool.

The following simulates an example run with the provided wrapper script `spindel.sh`. The `--indent` parameter causes the output XML file to be pretty-printed, mainly useful for debugging.

```
spindel.sh segmentize --lang=de -i 0-text2iso.xml --indent -o 1-segmentize.xml
spindel.sh guess --input=1-segmentize.xml --indent --output=2-guess.xml
spindel.sh normalize --input=2-guess.xml --indent --output=3-normalize.xml
spindel.sh pos --input=3-normalize.xml --indent --output=4-pos.xml
spindel.sh identify --input=4-pos.xml --indent --output=5-identify.xml
spindel.sh unidentify --input=5-identify.xml --indent --output=6-unidentify.xml
spindel.sh align -i 6-unidentify.xml --indent -o 7-align.xml --time 43 --every 5 -t
```

In the last step, `-t` causes transcription via the BAS web service to be added.

[27] But note that the `align` service may call the BAS transcription web service!

5 Conclusion

We have presented web services which implement a workflow for transcripts of spoken language which follow the TEI guidelines and in particular ISO 24624:2016 “Language resource management – Transcription of spoken language”. These web services were illustrated with respect to a use case that occurs frequently in our daily work at the IDS. We hope to have shown that these web services are useful for a broader public, and form a useful addition to the CLARIN universe.

6 Outlook

The web services are currently available from IDS, and have been integrated into the CLARIN infrastructure, so that they can be found in the Virtual Language Observatory and can also be used in WebLicht.

For tagging, we shall have to evaluate whether it is useful to offer a direct choice of the tagger models for specific languages, as we now prefer models for spoken language where they are available, i.e. in the case of French and German, and use one of three models in the case of Portuguese.

It may also be worthwhile to test whether language detection with moving windows can be applied to longer utterances in a way that detects language shifts like code switching.

As regards alignment, we intend to evaluate pseudoalignment more than impressionistically, and we intend to evaluate further forced alignment tools.

In the long run, we will also be able to evaluate in which form such tools are best distributed. While WebLicht is very useful as a didactic showroom and can help to quickly and easily explore how a given tool works, it is probably not the tool of choice for batch operation in curation work on larger datasets. Standalone webservices may serve this purpose better, but can still bring a considerable overhead, or even constitute an obstacle when legal restrictions do not permit sending data over the internet. At least as long as data curation remains an expert job carried out in specialised data centres, plain command line tools as described in section sec. 4.9 are likely to remain a candidate for the most adequate option.

References

Barras, Claude et al. 2001. “Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production”. In: *Speech Communication* 33.1–2, pp. 5–22.

Betten, Anne, ed. 1995. *Sprachbewahrung nach der Emigration – Das Deutsch der 20er Jahre in Israel. Teil I: Transkripte und Tondokumente. unter Mitarbeit von Sigrid Graßl*. Phonai 42. Tübingen: Niemeyer.

Brinckmann, Caren et al. 2008. “German Today: a really extensive Corpus of Spoken Standard German”. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*. Marrakech, Morocco: European Language Resources Association (ELRA). URL: `http://www.lrec-conf.org/proceedings/lrec2008/pdf/806_paper.pdf`.

Broeder, Daan et al. 2012. “CMDI: a component metadata infrastructure”. In: *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*.

Clyne, Michael. 1981. *Deutsch als Muttersprache in Australien. Zur Ökologie einer Einwanderersprache. In Zusammenarbeit mit dem Centre for Migrant Studies, Monash University.* Wiesbaden: Franz Steiner.

CLARIN ERIC. 2019. *Component Metadata*. URL: `https://www.clarin.eu/content/component-metadata`.

Devran, Serap. 2017. *Deutsch-türkische Migration: Die Darstellung narrativer Identitäten von Studentinnen in Istanbul. Eine biografie- und interaktionsanalytische Pilotstudie*. amades. Mannheim: Institut für Deutsche Sprache.

Draxler, Christoph, Jonathan Harrington and Florian Schiel. 2017. “Towards the next generation of speech tools and corpora”. In: *Computer Speech and Language* 46, pp. 175–178.

Giesbrecht, Eugenie and Stefan Evert. 2009. “Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus”. In: *Web as Corpus Workshop (WAC5)*.

Hinrichs, Erhard, Marie Hinrichs and Thomas Zastrow. 2010. “WebLicht: Web-Based LRT Services for German”. In: *Proceedings of the ACL 2010 System Demonstrations*. ACLDemos ’10. Uppsala, Sweden: Association for Computational Linguistics, pp. 25–29.

ISO. 2016. *ISO 24624:2016 Language resource management – Transcription of spoken language*. Tech. rep. Genève: ISO.

Kipp, Sandra Joy. 2002. "German-English Bilingualism in the Western District of Victoria". PhD thesis. Department of Linguistics and Applied Linguistics. The University of Melbourne.

Kisler, Thomas, Uwe D. Reichel and Florian Schiel. 2017. "Multilingual processing of speech via web services". In: *Computer Speech and Language* 45, pp. 326–347.

Lehr, Ursula and Hans Thomae, eds. 1987. *Formen seelischen Alterns*. Stuttgart: Enke.

Lich, Glen E. and Michael Clyne. 1984. *Deutsch als Muttersprache in Australien: zur Ökologie einer Einwanderersprache. In Zusammenarbeit mit dem Centre for Migrant Studies, Monash University*. Wiesbaden: Franz Steiner.

Liégeois, Loïc et al. 2017. "Vers un format pivot commun pour la mutualisation, l'échange et l'analyse des corpus oraux". In: *FLORAL*. Orléans, France.

MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk: Transcription format and programs*. 3rd ed. Mahwah, NJ: Lawrence Erlbaum.

Schmid, Helmut. 1995. "Improvements In Part-of-Speech Tagging With an Application To German". In: *In Proceedings of the ACL SIGDAT-Workshop*, pp. 47–50.

Schmidt, Thomas. 2011. "A TEI-based approach to standardising spoken language transcription". In: *Journal of the Text Encoding Initiative* 1, pp. 1–22.

Schmidt, Thomas. 2012. "EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language". In: *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12)*. Ed. by Thierry Declerck, Khalid Choukri and Nicoletta Calzolari. European Language Resources Association (ELRA), pp. 236–240.

Schmidt, Thomas. 2016. "Construction and dissemination of a corpus of spoken interaction – tools and workflows in the FOLK project". In: *Journal for Language Technology and Computational Linguistics* 31.1. Ed. by Marc Kupietz and Alexander Geyken, pp. 127–154.

Schmidt, Thomas. 2017. "DGD – die Datenbank für Gesprochenes Deutsch. Mündliche Korpora am Institut für Deutsche Sprache (IDS) in Mannheim". de. In: *Zeitschrift für germanistische Linguistik* 45.3. Ed. by Vilmos Ágel et al., pp. 451–463. URL: `http://nbn-resolving.de/urn:nbn:de:bsz:mh39-68145`.

Schmidt, Thomas. 2019. "Das Berliner Wendekorpus am Archiv für gesprochenes Deutsch". In: *Sprechen im Umbruch. Zeitzeugen erzählen und argumentieren rund um den Fall der Mauer im Wendekorpus*. Ed. by Norbert Dittmar and Christine Paul. Mannheim: Leibniz-Institut für Deutsche Sprache (IDS), pp. 23–27.

Schmidt, Thomas, Hanna Hedeland and Daniel Jettka. 2017. "Conversion and annotation web services for spoken language data in CLARIN". In: *Selected papers from the CLARIN Annual Conf. 2016*. Ed. by Lars Borin. Linköping University Electronic Press, pp. 113–130.

Schmidt, Thomas and Wilfried Schütte. 2010. "FOLKER: An Annotation Tool for Efficient Transcription of Natural, Multi-party Interaction". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). URL: `http://www.lrec-conf.org/proceedings/lrec2010/pdf/18_Paper.pdf`.

Schmidt, Thomas and Kai Wörner. 2014. "EXMARaLDA". In: *The Oxford handbook of corpus phonology*. Ed. by Jacques Durand, Ulrike Gut and Gjert Kristoffersen. Oxford: Oxford University Press.

Sloetjes, Han. 2014. "ELAN: Multimedia Annotation Application". In: *The Oxford handbook of corpus phonology*. Ed. by Jacques Durand, Ulrike Gut and Gjert Kristoffersen. Oxford: Oxford University Press.

TEI Consortium. 2019. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Tech. rep. Version 3.5.0. Last updated on 29th January 2019. TEI Consortium.

Thüne, Eva-Maria. 2019. *Gerettet. Berichte von Kindertransport und Auswanderung nach Großbritannien*. Berlin, Leipzig: Hentrich & Hentrich.

Verdonik, Darinka et al. 2013. "Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS". In: *Language Resources and Evaluation* 47.4, pp. 1031–1048.

Westpfahl, Swantje. 2020. "POS-Tagging für Transkripte gesprochener Sprache. Entwicklung einer automatisierten Wortarten-Annotation am Beispiel des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)". PhD thesis. Tübingen.

Wilkinson, Mark D. et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3, p. 160018. URL: `https://doi.org/10.1038/sdata.2016.18`.

Enriching and Increasing the Usability of Lexicographical Data for Less-Resourced Languages

Dirk Goldhahn
Natural Language Processing Group, University of Leipzig, Germany
Saxon Academy of Sciences and Humanities, Leipzig, Germany
goldhahn@saw-leipzig.de

Thomas Eckart
Natural Language Processing Group, University of Leipzig, Germany
Saxon Academy of Sciences and Humanities, Leipzig, Germany
teckart@informatik.uni-leipzig.de

Sonja Bosch
Department of African Languages,
University of South Africa, South Africa
boschse@unisa.ac.za

Abstract

This paper presents a use case for enriching lexicographical data for less-resourced languages employing the CLARIN infrastructure. Newly prepared lexicographical data sets for under-resourced Bantu languages spoken in southern regions of the African continent form the basis of the presented work. These datasets have been made digitally available using well-established standards of the Linguistic Linked Open Data (LLOD) community. To overcome the insufficient amount of freely available reference material, a crowdsourcing web portal for collecting textual data for less-resourced languages has been created and incorporated into the CLARIN infrastructure. Using this portal, the number of available text resources for the respective languages was significantly increased in a community effort. The collected content is used to enrich lexicographical data with real-world samples to increase the usability of the entire resource.

1 Introduction

The availability of contemporary text material is a prerequisite for a variety of applications and research scenarios, especially including studying recent developments in language use. Projects such as *An Crúbadán*[1] offer text freely available on the web to enable such studies for languages with small numbers of speakers. Resources in *An Crúbadán* are typically added manually, have a limited scope but are available for over 2,000 languages.

To expand the amount of available textual data, crawling and processing web content is now a standard procedure to acquire those needed resources. As a positive consequence of the vast amount of available online content, preselection of highly specific material is now possible for many languages and allows examination of all sorts of linguistic phenomena for specific domains and genres. Automatically generating valuable data sources from online resources requires specific means of text acquisition and pre-processing of the gathered material. Subsequently, different systems that simplify the crawling and processing of web pages for end users were developed and are in active use. One of the most popular services is the SketchEngine (Kilgarriff et al., 2014) which has a focus on lexicography.

On the other hand, for many less-resourced languages, significant amounts of material are now available for the first time. Using standards of the growing Linguistic Linked Open Data (LLOD) community,

[1] http://crubadan.org/

Dirk Goldhahn, Thomas Eckart and Sonja Bosch 2020. Enriching and Increasing the Usability of Lexicographical Data for Less-Resourced Languages. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 23–32.

connecting – so far isolated – datasets of all kinds, helps creating substantial resources for these languages in a federated infrastructure. One of the benefits of these interconnections is the ability of building bridges between lexicographical entries and concrete, real-world usage examples. The resulting resources have a high value for all kinds of use cases and user groups, including being essential for first- and second-language acquisition.

This paper focuses on a specific use case, facilitated by the federated research infrastructure CLARIN (Hinrichs and Krauwer, 2014), in which newly created lexicographical datasets for some Bantu languages were enriched using a new web crawling portal that focuses on the acquisition of text material for less-resourced languages.

2 Crawling Under-Resourced Languages

The situation concerning the availability of digital language resources is satisfactory only for a small number of languages. Even for most of the languages with more than one million speakers, no reasonably sized textual resources or tools like POS taggers are available. This points to a widespread need for digital language resources for many languages of the world. Therefore, the CURL (**C**rawling **U**nder-**R**esourced **L**anguages) web portal[2] for corpus collection with a focus on languages with more than one million speakers, has been initiated (Goldhahn et al., 2016) as a service in the CLARIN infrastructure. It relies on native speakers with knowledge of web pages in their respective language. The initiative gives interested scholars and language enthusiasts the opportunity to contribute to corpus creation or extension by simply entering a URL into a web interface.

In the backend, Heritrix (Mohr et al., 2004), the crawler of the Internet Archive[3], is used in combination with a well-established corpus processing chain that was adapted to append newly added web pages to continuously growing corpora. This enables us to collect larger corpora for under-resourced languages by means of a community effort. These corpora are made publicly available within the CLARIN infrastructure, both directly in the created portal and as part of the Leipzig Corpora Collection project (Goldhahn et al., 2012). Figure 1 gives an overview of the general structure of the application.

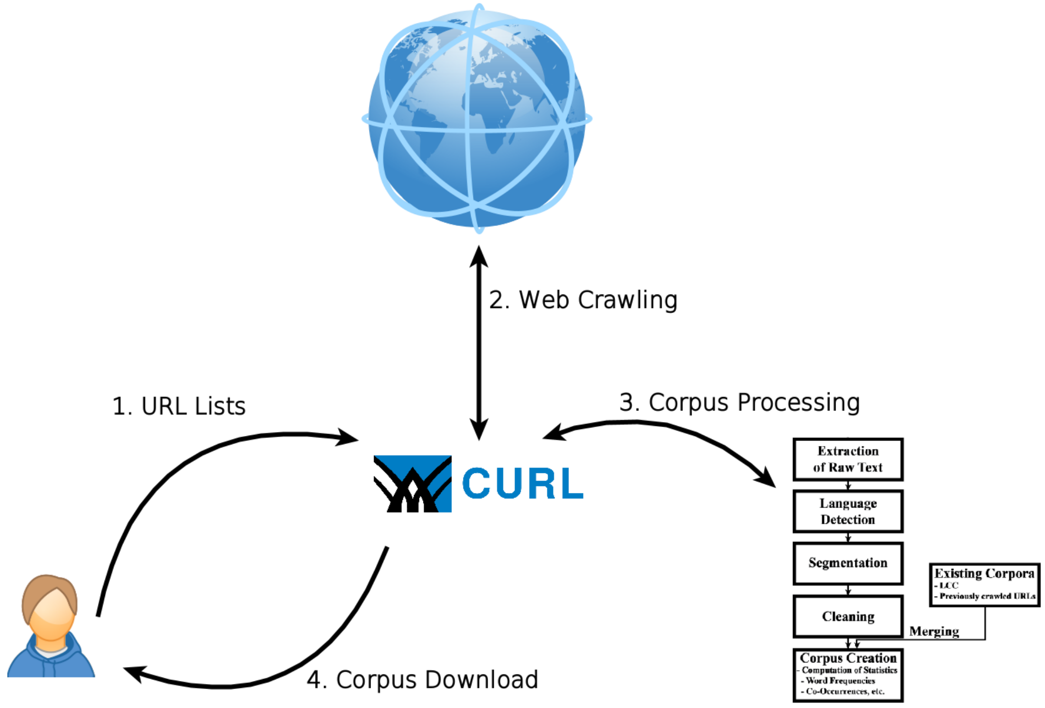


Figure 1: Schematic overview of the CURL portal

[2] https://curl.corpora.uni-leipzig.de/

[3] https://archive.org

The data provided do not only include the crawled text material with information about their source of origin and date of crawling; in addition, statistical analysis of word frequencies and word relationships based on statistical word co-occurrences is also generated and included in the download files. This kind of information has proven to be a valuable resource when working with language material, where manually created and evaluated resources on word semantics like thesauri are hardly available.

Since its establishment, the web portal has helped creating initial text resources for several languages and also facilitated the expansion of available text collections. All in all, more than 10,000 URLs were submitted to the system in 127 crawling jobs for 62 languages. Concrete efforts for two languages belonging to the Bantu language family - Xhosa (ISO 639-3:xho) and Kalanga (ISO 639-3:kck)[4] - will be described in the following sections.

3 Bantu Languages

The Bantu languages are a family of languages spoken in sub-Saharan Africa with around 240 million speakers spread across 27 African countries (Nurse and Philippson, 2003:1). The exact number of Bantu languages cannot easily be determined because it is often difficult to draw the line between a language and a dialect. However, taking this into consideration, Nurse and Philippson's (2003:3) estimate is about 300 Bantu languages two of which are under discussion, namely Xhosa and Kalanga. Xhosa, with approximately 8.1 million speakers, is spoken predominantly in the Eastern Cape and Western Cape regions of South Africa, and is classified as a member of the Nguni group of languages (Nurse and Philippson, 2003:649). The cross-border language Kalanga is spoken in eastern Botswana and western Zimbabwe and has a total of 338,000 users[5]. It is classified as a member of the larger Shona group of languages (Nurse and Philippson, 2003:609).

Many linguistic features are shared by these two languages. Morphologically, they are of an agglutinating nature with verbs exhibiting an intricate set of affixes, while nouns are assigned to classes by means of so-called class prefixes. The number of class prefixes per language varies, but in general Bantu languages have between 12 and 20 class prefixes. Each class is characterised by a distinct prefix, a particular singular/plural pairing and agreement that branches out to other word categories, namely verbs (subject and object markers), adjectives, possessives and so on. Exceptions to the singular/plural rule also occur, for example mass nouns such as 'water' in so-called plural classes do not have a singular form; plurals of class 11 nouns are found in class 10, while a class such as 14 is usually not associated with a number at all.

4 Dictionary Data

Like most Bantu languages, Xhosa and Kalanga are considered resource scarce languages, implying that linguistic resources such as large annotated corpora and machine-readable lexicons are not available.

Moreover, academic and commercial interest in developing such resources is limited. In the following section, available sources for lexicographical data for Bantu languages used in this publication are described in more detail.

4.1 Xhosa Dictionary

Xhosa lexical data were taken from a resource compiled by J.A. Louw (University of South Africa - UNISA) which is available under a Creative Commons (CC) license. This Xhosa lexicographical data set consists of morphological information accompanied by English translations. It was created and made available by the authors for purposes of further developing Xhosa language resources (Bosch et al., 2018). The data were compiled with the intention of documenting Xhosa words and expanding existing bilingual Xhosa dictionaries by means of – among others – botanical names, animal names, grammar terms, modern forms and so on, as well as lexicalisations of verbs with extensions. The publication process involved digitisation into CSV tables and several iterations of quality control in order to make the data reusable and shareable.

[4] The speakers of the two languages use the names isiXhosa and Ikalanga.

[5] https://www.ethnologue.com/language/kck

In its current state, the data set contains approximately 6,800 lexical entries and is already published in the CLARIN infrastructure[6] and available via a dedicated web portal[7]. Table 1 gives a short summary of its current inventory.

Dataset feature	Value
Number of noun lexemes	4020
Number of verb lexemes	2763
Number of noun classes	15
Number of English translations	7807

Table 1: Characteristics of the Xhosa dataset (as of 2020-01-12)

However, the compilation process is not completed yet and the data are still subject to quality assurance measures. The final dataset is expected to contain approximately 10,000 lexical entries and will also be available in the context of CLARIN via the South African Centre for Digital Language Resources (SADiLaR[8]).

4.2 Kalanga Dictionary

Lexicographical data for the Kalanga language was extracted from the Comparative Bantu OnLine Dictionary (CBOLD[9]). The project started in 1994 to create open source lexicographical data for Bantu languages. The amount and range of available data, and its quality and format vary from dictionary to dictionary. The CBOLD dictionary for Kalanga was created in 1994 by Joyce Mathangwane and is provided as a plain text file. The dictionary contains 2960 lexemes with information about the part of speech, tone, noun classes and prefix/stem structure for the nouns. Additionally, English translations are provided. The resulting data set is currently used primarily for testing different approaches of dictionary alignment (Eckart et al., 2019). However, it also shows the high relevance of openly available data as a starting point for building up structured data sets – including their extension and improvement – in cases where no comparable resources are available. In the case of the CBOLD project, this includes material for dozens of less-resourced languages.

4.3 Bantu Language Model

The lexical resources introduced were transformed into a unified schema to simplify all relevant data enrichment and quality assurance procedures and to form a basis for future applications and user interfaces. The Bantu Language Model (BLM) (Bosch et al., 2018) is an ontology of the Linguistic Linked Open Data (LLOD) community that ensures semantic and structural interoperability. The BLM is based on the MMoOn ontology (Klimek, 2017) and allows for the representation and interrelation of lexical, morphological and translational elements but also common grammatical meanings as well as noun class elements of Bantu languages.

A summary of the chosen data model is depicted in Figure 2. In the context of Bantu languages, both the nominal classifier system – as a (language family) specific *LinguisticCategory* – and the support of their rich morpheme structure and morph-related predicates (*isAllomorphTo, isHomonymTo*) are of particular relevance.

The benefit of using an LOD-based format is the simplicity to enrich existing datasets with additional information. In the context of this contribution, the focus lies on connecting lexicographical data based on full forms with real-world samples. The resulting interconnected resources are helpful for a variety

[6] https://hdl.handle.net/11022/0000-0007-C655-A
[7] https://rdf.corpora.uni-leipzig.de
[8] https://www.sadilar.org
[9] http://www.cbold.ish-lyon.cnrs.fr

of user groups and topics, including support of first and second language acquisition, or as a general resource for all types of text-producing activities.

A more detailed discussion of the specific problems when representing these interconnections and the current state of helpful data models is given in the following section.

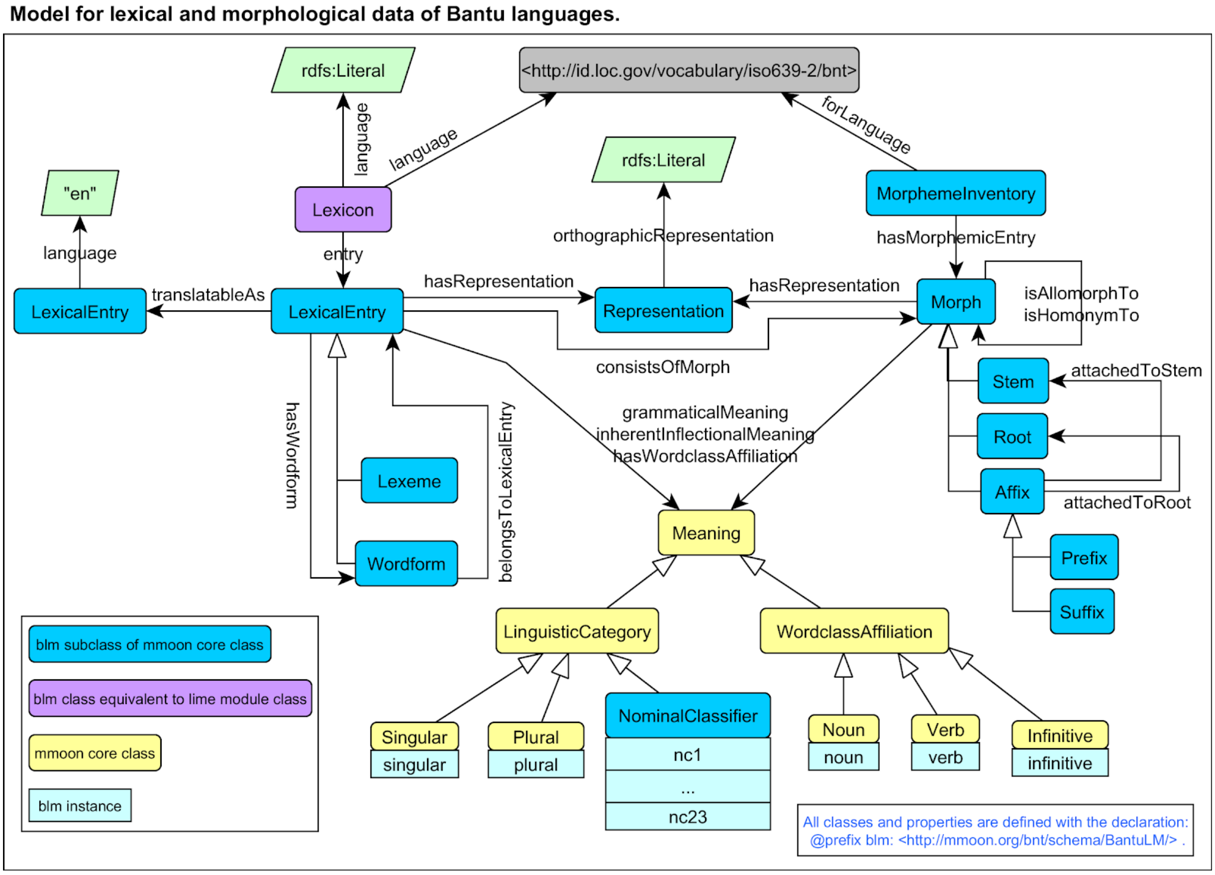


Figure 2: Summary of the Bantu Language Model

4.4 Representing Examples in Current Lexicographical Data Models

Besides standard lexicographical data types focusing on pragmatic, morphosyntactic and similar information with their established means of representation and publication, the growing utilisation of statistical analysis in the field extends the lexicographical focus significantly. Unfortunately, many established data models and formats (like ISO 24613:2008 LMF or the dictionary module of the TEI guidelines) allow the incorporation of these data types only in small parts, if at all.

Current standardisation endeavours try to reduce this problem by extending established standards with new modules or by extending established data models. One candidate currently under progress in the LLOD community is *OntoLex-FRAC*[10] (Frequency, Attestations and Corpus Data) as a new model of the established OntoLex ontology (McCrae et al., 2017). In the context of this paper, it seems to be a fit solution as it will both support referencing concrete usage examples and statistical results. However, the OntoLex-FRAC model is still under development. The incorporation of references to the described data sets based on included full forms will be started as soon as its standardisation is completed.

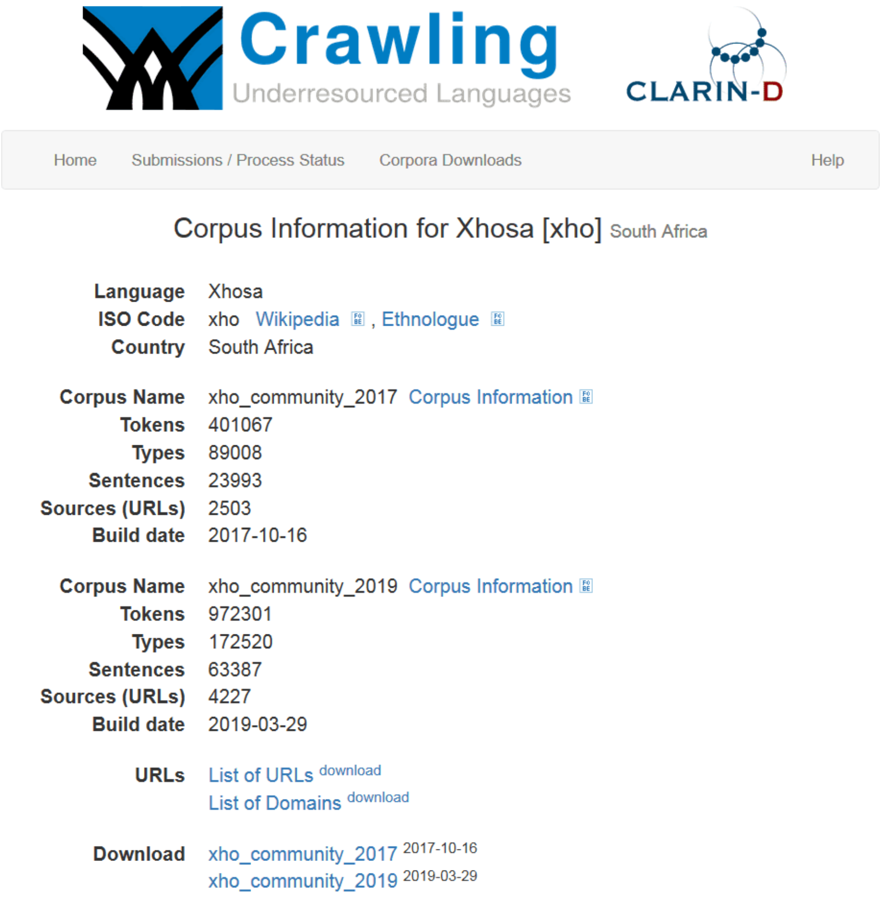
The inclusion of examples by external references makes good use of the CLARIN architecture that strongly relies on persistent identifiers and distributed resources. For example, corpora provided by the Leipzig Corpora Collection – including the data generated by the CURL portal – provide a granularity down to the sentence level (Boehlke et al., 2012), can be addressed using handles with part identifier, and are therefore easy to use for direct reference in a Linked Data environment.

[10] https://github.com/acoli-repo/ontolex-frac

5 Language Data

5.1 Collecting and Processing of Language Data

In a next step, the newly created lexical data were enriched with additional information: the availability of sample sentences proves to be valuable for users of lexical resources since they provide real-world usage examples of the lexical units. In the beginning of the project, available text data for the respective languages was very limited, both in the Leipzig Corpora Collection (LCC) and in other freely available digital resources. For Xhosa, fewer than 18,000 sentences could be found in the LCC. For Kalanga, the situation was even worse with only about 600 sentences. By advertising the initiative to some researchers in the respective communities, we were able to collect 180 seed URLs for Xhosa and one web domain for Kalanga. Crawling resulted in 45,585 additional unique sentences for Xhosa and 996 for Kalanga, increasing available resources significantly. Figure 3 depicts these text collecting efforts for Xhosa.



Corpus Information for Xhosa [xho] South Africa

Language	Xhosa
ISO Code	xho Wikipedia , Ethnologue
Country	South Africa
Corpus Name	xho_community_2017 Corpus Information
Tokens	401067
Types	89008
Sentences	23993
Sources (URLs)	2503
Build date	2017-10-16
Corpus Name	xho_community_2019 Corpus Information
Tokens	972301
Types	172520
Sentences	63387
Sources (URLs)	4227
Build date	2019-03-29
URLs	List of URLs download List of Domains download
Download	xho_community_2017 2017-10-16 xho_community_2019 2019-03-29

Figure 3: Overview of crawling activities in the CURL portal for Xhosa since 2017

The textual resources were processed to serve as a basis for assigning sample sentences to dictionary entries. For Xhosa, this resulted in sample sentences for about 25% of the lexical entries available. Since the coverage is significantly higher for more frequent words and these words are typically queried more often, this will result in a higher sample sentence coverage for actual queries. Additionally, the integration of tools for lemmatisation and morphological decomposition can increase the number even further.

The textual data are made available in the CLARIN infrastructure via the Leipzig repository. This allows for download of the data sets, for searching in the data via the Federated Content Search (FCS) or the local Leipzig Corpora Collection portal, for sustainably citing textual resources on sentence level and for further processing using web tools such as WebLicht. In a next step, extending the Bantu Language Model dataset is planned to allow for a direct linking of lexical entries and sample sentences using LLOD formats and therefore for easier integration, as depicted above.

Although extracting authentic examples from corpora is often a contentious issue, in particular for the purpose of language learners, there are opinions such as that of Frankenberg-Garcia (2012:290) that several corpus examples per lexeme can offer learners concentrated patterns of language that would encourage appropriate generalisations. Hanks (2012:431) is also of the opinion that corpora reflect language as communicative behaviour since "meanings reside only partly in individual words; meanings also reside in the phraseology (or constructions) in which words are used."

5.2 Sample Results

In this section, identified sample sentences for lexical entries in Xhosa and Kalanga will be presented. For Xhosa, the lexemes *umfazi* and *abafazi* were chosen. Their lexical data is summarized in Table 2.

Lexeme	**Prefix**	**Root**	**Gram. number**	**Class**	**POS**	**Gloss**
umfazi	um	fazi	singular	1	noun	wife
abafazi	aba	fazi	plural	2	noun	wives

Table 2: Lexical information available for the lexemes *umfazi* and *abafazi*

For both lexical entries, sample sentences (with corresponding English translations[11]) can now be provided by matching lexemes with word forms occurring in the crawled text material.

Sample sentences for *umfazi*:

(1) *Abazali bam abagulayo kwaye bafuna ukutyelela US kwaye nzima kuhlala* ***umfazi*** *wam.*
"My parents are sick and want to visit the US and it is difficult for my **wife** to stay."

(2) *Ingaba* ***umfazi*** *uzakuziva njani xa indoda yakhe ingaphangeli?*
"How would the **wife** feel if her husband was unemployed?"

(3) *Ndinabantwana abathathu kwaye* ***umfazi*** *wam akaphangeli.*
"I have three children and my **wife** is unemployed."

(4) *'Uza kugalela ntoni kuyo?' kwabuza omnye* ***umfazi****.*
"'What will you add to it?' asked another **woman**."

(5) *Baya efanayo baya oonyana bam xa ufune* ***umfazi****.*
"The same goes for my sons when you find a **wife**."

Sample sentences for *abafazi*:

(6) *Ndiza kuqala bathi* ***abafazi*** *kufuneka uyeke nangegunya ngokwabo, kodwa nam ndiya umngeni amadoda ukwandisa ku nembasa & ukuqala imbeko* ***abafazi****.*
*"*I will start by saying that **women** need to give up the power themselves, but I also challenge men to increase in honor & start respecting **women**."

(7) *Ngenxa yokuba kwakusele amadoda ambalwa kuloo mmandla, kwanyanzeleka ukuba* ***abafazi*** *baluthathele kubo uxanduva lokuloba.*
*"*With so few men left in the area, **women** had to take responsibility for their fishing."

(8) *USankara ukwa ngumongameli wokuqala eAfrika ukuphuhlisa amalungelo* ***abafazi****; esithi le nto ingumfazi nayo ingumlingani na le nto iyindoda.*
*"*Sankara is also the first African president to promote **women**'s rights; saying that this woman is also a partner is this man."

[11] The translations provided are automatically generated based on Google Translate results: https://translate.google.com

(9) *Iyomhla kuphela yaye wayezeke* ***abafazi*** *abamnyama ndatshata ngakumbi kuxanduva (umntwana) nolindelo yenkcubeko ngaphezu umdla ofanayo.*
“Only date and had married black **women** I married more to the (child) responsibility and cultural expectations than the same interest.”

(10) *Beyonce and ezinye iimvumi ababhinqileyo' iingoma kukhokelela* ***abafazi*** *evakalisa ubuni babo yaye ke objectified.*
“Beyonce and other female singers' songs lead **women** to express their identity and are therefore objectified.”

The identical procedure was applied to the Kalanga inventory. Lexical information for the sample lexeme *ikombo* can be found in Table 3.

Lexeme	**Prefix**	**Root**	**Tone**	**Class**	**POS**	**Gloss**
ikombo	i	kombo	HH	7	noun	navel

Table 3: Lexical information for the lexeme *ikombo*

Identified sample sentences for *ikombo* include the following[12]:

(11) *Atitjaka dwilila taka lingilila baka tatana bakajalo,* ***ikombo*** *tjibe tji shomoka mu liboko gwe mdala tjino wila mbeli kwe mtshana.*

(12) *Atitji gele towana shango ya pituka; mdala wabe mbeli, mtshana wabe iye ushule ne* ***ikombo****, kufiwa tate.*

(13) *Ebe atji nunga* ***ikombo*** *mtshana ndokubudza, ebe e shanduka e lingisana ne mdala.*

(14) *Koti zhulo tili gele kusi kwe mpani pa khisimusi tobona mdala e pinda aka tatamila mtshana elitsha* ***ikombo*** *to come nice.*

(15) *Mtshana a ka amuchila kwa ka nlingisana (****ikombo****).*

(16) *Mtshana alishule ne* ***ikombo*** *akabata; “andibilo mubudza tate ati muletje ndideelela.”*

(17) *Mtshana ebe e ntumba ne* ***ikombo*** *atenti.*

(18) *Yaka bobola ikano ngina ka mai ne mabilo titjara alishule ne* ***ikombo*** *atenti.*

(19) *Yeela, tjakalila* ***ikombo*** *ilelo zhuba abona kuti ya, kwiba kuna zwibili.*

For nouns, sample sentences can be attributed easily since the lexical entries are usually available for the singular and/or plural form (see *umfazi* and *abafazi* in Xhosa). For verbs, the situation is more challenging. Lexemes as found in the dictionary do not appear unchanged in the sample sentences; affixation has to be taken into account.

By using tools such as a lemmatiser[13], morphological decomposer[14] and POS-tagger[15] for Xhosa, this gap between the lexicon and crawled full texts can be overcome. The usage of these tools will be investigated in the near future.

5.3 Lexical Ambiguity

Besides lemmatisation, aspects of disambiguation play a role when attributing sample sentences to dictionary entries. In this section, cases of homonymy and polysemy will be discussed. Currently these cases are handled manually. Applying methods of automatic sense disambiguation are limited due to the

[12] Due to missing support in Google Translate, English translations had to be omitted for Kalanga sample sentences and are left as future work.
[13] https://repo.sadilar.org/handle/20.500.12185/310
[14] https://repo.sadilar.org/handle/20.500.12185/311
[15] https://repo.sadilar.org/handle/20.500.12185/323

small number of available text samples.

An example of the sense relation homonymy is demonstrated by means of the noun *ithanga* (plural *amathanga*) which has two unrelated meanings, namely “pumpkin” (a type of vegetable) and “colony” (a geographical area politically controlled by a distant country). The following sentences from the crawled text material (with their English translations) illustrate disambiguation in context:

(20) *Ilizwe alisawulwa ngamandla* ***amathanga*** (colonial powers).
“The country is not ruled by colonial powers (literally – powers of **colonies**).”

(21) *Isitiya sakhe semifuno sasisoloko siyokozela zizinto ezimnandi zokutya, kodwa ngamanye amaxesha ayede adlulise ngobuninzi* ***amathanga,*** *umbona, iitapile nezinye iintlobo zemifuno.*
“His vegetable garden was always full of delicious things to eat, but sometimes even heaped large quantities of **pumpkins**, corn, potatoes and other vegetables.”

The sense relation polysemy, is illustrated in the case of the noun *inyanga* (plural *izinyanga*) which has two related meanings “moon” (astronomy) and “month” (time period). The two sense relations, as disambiguated in context, are illustrated in the following example sentences (accompanied by English translations) extracted from the crawled text material:

(22) *Ndibona* ***inyanga*** *iphuma ndiselapha.*
“I can see the **moon** rising from here.”

(23) *Ngobunye ubusuku,* ***inyanga*** *eyayikhanya yenza kwakho izithunzi edlelweni.*
“One night, a bright **moon** made shadows in the fields.”

(24) *Yayibubusuku obubanda kakhulu, kukhanyise luzizi* ***inyanga*** *eliceba.*
“It was a very cold night, with a clear **moon**light.”

(25) *Wahlala naye* ***inyanga*** *iphela.*
“And he stayed with him for a **month**.”

(26) *Ukuba umsebenzi unqunyanyisiwe okanye utshintshelwe kwenye indawo, umqeshi makenze konke okusemandleni akhe okanye aqukumbele ukuthethwa kwetyala ingadlulanga* ***inyanga*** *umsebenzi lowo enqunyanyisiwe okanye etshintshelwe kwenye indawo*.
“If an employee is suspended or transferred, the employer must do everything in his power or conclude a hearing within one **month** of the employee's termination or transfer.”

6 Conclusion and Further Work

This paper presented a use case for enriching lexicographical data for less-resourced languages with sample sentences. The basis was recently added resources and services of the CLARIN infrastructure such as the Xhosa lexicographical data based on the Bantu Language Model[16] and a portal for crawling under-resourced languages (CURL[17]). Results are made available via the CLARIN infrastructure to allow for wide applicability. They include text corpora for Xhosa[18] and Kalanga[19].

Future work will focus on deeper integration of the CURL portal into the CLARIN infrastructure. Advanced options for format conversion (e.g. TCF or plain text) are planned to be implemented. This will allow for direct processing of crawling results in environments such as WebLicht by employing the Language Resource Switchboard. Support for CLARIN’s private work space solution will increase usability even further. In addition, the RDF datasets will be extended to allow direct reference of sample sentences of which many are already integrated in CLARIN and available for reference using persistent identifiers. The integration of statistical outcomes in the data sets for enhanced usage scenarios is currently under investigation. This has the potential to enhance their utilisation in a mobile dictionary application that is presently under development.

[16] https://hdl.handle.net/11022/0000-0007-C655-A
[17] https://hdl.handle.net/11022/0000-0007-D369-5
[18] https://hdl.handle.net/11022/0000-0007-D396-1
[19] https://hdl.handle.net/11022/0000-0007-D395-2

References

Volker Boehlke, Torsten Compart and Thomas Eckart 2012. Building up a CLARIN resource center – Step 1: Providing metadata. In: Workshop on Describing Language Resources with Metadata at 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul 2012.

Sonja Bosch, Thomas Eckart, Bettina Klimek, Dirk Goldhahn and Uwe Quasthoff 2018. Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment, Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki (Japan) 2018.

Thomas Eckart, Sonja Bosch, Dirk Goldhahn, Uwe Quasthoff and Bettina Klimek 2019. Translation-based Dictionary Alignment for Under-resourced Bantu Languages, OpenAccess Series in Informatics (OASIcs), Vol. 70: Language Data and Knowledge LDK 2019.

Ana Frankenberg-Garcia 2012. Learners' Use of Corpus Examples. International Journal of Lexicography, Vol. 25 No. 3, pp. 273–296. doi:10.1093/ijl/ecs011

Dirk Goldhahn, Thomas Eckart and Uwe Quasthoff 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012) 2012.

Dirk Goldhahn, Maciej Sumalvico and Uwe Quasthoff 2016. Corpus Collection for Under-Resourced Languages with more than One Million Speakers, CCURL 2016 Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity 2016.

Patrick Hanks 2012. The Corpus Revolution in Lexicography. International Journal of Lexicography, Vol. 25 No. 4, pp. 398–436. doi:10.1093/ijl/ecs026

Erhard Hinrichs and Steven Krauwer 2014. The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014) 2014.

ISO 24613:2008 - Language resource management - Lexical markup framework (LMF). Iso.org.

Adam Kilgarriff, Vit Baisa, Jan Buta, Milos Jakubicek, Vojtech Kova, Jan Michelfeit, Pavel Rychly and Vit Suchomel 2014. The Sketch Engine: ten years on, Lexicography, pp. 7–36, Springer 2014.

Bettina Klimek 2017. Proposing an OntoLex-MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models, Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets 2017.

John P. McCrae, Julia Bosque Gil, Jordi Gràcia, Paul Buitelaar and Philipp Cimiano 2017. The OntoLex-Lemon Model: Development and Applications 2017.

Gordon Mohr, Michele Kimpton, Michael Stack and Igor Ranitovic 2004. Introduction to heritrix, an archival quality web crawler, Proceedings of the 4th International Web Archiving Workshop (IWAW'04) 2004.

Derek Nurse and Gérard Philippson 2003. The Bantu languages. London: Routledge.

Using DiaCollo for Historical Research

Bryan Jurish
Berlin-Brandenburgische Akademie der Wissenschaften
Berlin, Germany
jurish@bbaw.de

Maret Nieländer
Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung
Braunschweig, Germany
nielaender@leibniz-gei.de

Abstract

This article presents some applications of the open-source software tool DiaCollo for historical research. Developed in a cooperation between computational linguists and historians within the framework of CLARIN-D's discipline-specific working groups, DiaCollo can be used to explore and visualize diachronic collocation phenomena in large text corpora. In this paper, we briefly discuss the constitution and aims of the CLARIN-D discipline-specific working groups, and then introduce and demonstrate DiaCollo in more detail from a user perspective, providing concrete examples from the bi-weekly German-language newspaper "Die Grenzboten" (1841-1922) ("messengers from the borders") and other historical text corpora. Our goal is to demonstrate the utility of the software tool for historical research, and to raise awareness regarding the need for well-curated data and solutions for specific scientific interests.

1 Introduction

Ever since their establishment in 2011, German CLARIN centers have worked together with discipline-specific working groups to develop and improve their services in close dialogue with the needs of philologies, history, social science, etc.[1] The German CLARIN initiative, CLARIN-D, has strong roots in computational linguistics. With the help of the working groups, it has been possible to curate and integrate corpus data that is important to different fields of the humanities and social sciences, as well as to disseminate knowledge of the usefulness of computational linguistic methods for other disciplines.

Developed in a collaboration between historians and computational linguistics within the context of the CLARIN-D working groups, DiaCollo (Jurish; 2015, 2018) is an open-source software tool for exploration and interactive visualization of diachronic change with respect to collocation behavior in large collections of (historical) text. In addition to the technical, implementation-oriented issues common to all software development projects on the one hand and the various challenges of source criticism characteristic for historical research on the other, interdisciplinary collaborations of this kind present challenges all their own, ranging from lack of established shared terminology (e.g. "term", "concept", "query", "type/token", "collocant/collocate", "relevance") to fundamentally different approaches to what constitutes "research activity" as such (analytic/stipulative vs. hermeneutic/interpretive). Over the course of the collaboration, DiaCollo underwent several iterations of the software development lifecycle phases of "planning", "implementation", and "evaluation" – in the latter case relying on extensive feedback from the working group's historians to identify missing functionality and potentially useful new features.

After its initial release, DiaCollo was integrated into the corpus administration framework of the CLARIN service center at the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW). At the time of writing (March, 2020), DiaCollo indices for 78 distinct curated text corpora comprising a total of over 23,000,000,000 (23G) source tokens have been indexed and deployed at the BBAW,

[1] https://www.clarin-d.net/en/disciplines

Bryan Jurish and Maret Nieländer 2020. Using DiaCollo for Historical Research. *Selected papers from the CLARIN Annual Conference 2019.* Linköping Electronic Conference Proceedings 172: 172 33–40.

where they enjoy a modicum of popularity with an average of about 500 queries per day over the past 12 months. Of these curated corpora, 18 indices are publicly accessible, 3 require registration with the www.dwds.de web platform, and 3 more can be queried using CLARIN credentials.[2] The BBAW corpus infrastructure is strongly focused on German-language sources – both historical and contemporary – but also includes DiaCollo indices for several English and French corpora, developed in cooperation with the respective corpus providers. DiaCollo itself is language-agnostic and can be applied to any source corpus, provided that it has been appropriately pre-processed. Documentation and demos are available in German and English.

2 Background

In linguistics, collocations are sets of words or terms that frequently occur in one another's vicinity, presumably because they belong to the same "semantic field" and thus shape their respective meanings, as suggested by J.R. Firth's (1957) well-known assertion that "you shall know a word by the company it keeps" and Wittgenstein's (1953) famous *"die Bedeutung eines Wortes ist sein Gebrauch in der Sprache"* ("the meaning of a word is its use in the language"). For example, the fact that the words "smoke" and "fire" tend to occur near one another in a text corpus suggests that there is indeed a semantic relation between them – in this case, a causal one.

Previous work in computational linguistics has established a number of methods for unsupervised discovery of collocations in text corpora, based on distributional properties of the collocated terms alone (see e.g. Evert, 2008). Informally, distributional collocation discovery procedures identify those word-pairs as potential collocations which occur together substantially more often than would be expected under "chance" conditions. Collocation profiling is a related technique which requires the user to provide one or more search terms of interest (the "collocant"), and searches for those terms in the corpus which associate most strongly with the collocant (i.e. the "collocates"). The association strength of a particular candidate collocate is estimated with regard to its own independent frequency in the corpus as well as that of the collocant, and should provide a quantitative approximation of the "relevance" of the respective collocate for the given collocant. To illustrate, a simple collocation profiling procedure would investigate all words in a pre-defined neighborhood of the search term, e.g. within a window of 5 words to the left and right. The more often one of these words occurs together with the collocant, compared to its frequency in the corpus overall, the stronger its association with the search term will be.

Synchronic collocation analysis has long been employed to provide evidence for typical usage(s) of words/concepts in the corpus as a whole, i.e. (distributional) semantics. It is also possible to compare collocation-profiles of different words, to look at differences and similarities in usage (e.g. for lexicography). "Ready-to-use" implementations include both the DWDS "*Wortprofil"* database[3] and Cyril Belica's co-occurrence database "CCDB"[4]. More complex user queries are possible (and familiarity with the associated software tools and interfaces required) when using the *Deutsches Referenzkorpus* (DeReKo) with the "COSMAS II" interface.

When analyzing historical text, synchronic collocation analysis can be a part of departure for comparing the usage of certain terms in historic source material with their use in the contemporary reference corpora. Historical corpora (if existent and accessible in sufficient quality and quantity for the culture/time period of interest) can provide empirical data regarding the use of specific terms during a certain epoch. By comparing historical and synchronic reference corpora, researchers can identify linguistic divergences, which in turn may lead to hypotheses (and perhaps even conclusions) about an author's or issuing institution's specific intentions as realized by their "linguistic framing" of the phenomena under discussion.

In order to be truly useful for historical research, collocation analysis should also provide methods that reveal changes in language use over time (in specific corpora), allowing users to trace phenomena such as semantic shifts, discourse trends, history of concepts, introduction of neologisms, etc. DiaCollo has been specifically developed for this purpose. As a free, open-source, language-agnostic software package[5], it can also be integrated into other project contexts and corpus infrastructures.

[2] http://kaskade.dwds.de/~jurish/diacollo/corpora/

[3] http://www.dwds.de/d/ressources#wortprofil

[4] http://corpora.ids-mannheim.de/ccdb/

[5] http://metacpan.org/release/DiaColloDB/

DiaCollo corpus data must be pre-tokenized and each document must be assigned a characteristic date (e.g. year of publication) to represent the diachronic axis. If provided by the corpus, DiaCollo can also make use of additional token-level attributes such as lemmata or part-of-speech tags as well as document-level metadata such as author or genre to enable finer-grained queries and aggregation of result profiles (Jurish, 2018). As with any other data-driven procedure, DiaCollo is subject to "garbage-in / garbage-out" phenomena: "messy" corpora containing abundant OCR or annotation errors, mistokenizations, and/or incorrect document metadata are less likely to produce satisfying results for humanities researchers than "tidy", well-curated corpora with accurate metadata and reliable linguistic annotations (Nieländer & Weiß, 2018).

3 Output, Visualization, and Usability

The usefulness of computer-aided text analysis for historical research and the acceptance of such methods in the discipline are influenced by a number of factors. For this reason, historians working in the CLARIN-D working group "Contemporary History" (2014-2016) and "History" (2016-present) were involved in the development of DiaCollo from the very beginning. Historians traditionally had to investigate any diachronic and synchronic differences in an individual's or a group's language use by personally "learning" these different usages, as one would learn different languages, and then comparing them. This method – as large parts of historic methodology in general – focussed on finding meaningful anomalies (as opposed to overall patterns), and required careful and close reading of original source texts or true-to-text reproductions.

The goal was to design and implement the software tool in such a way that "distant reading" – a bird's-eye view of large digital text corpora – would be possible. At the same time, the algorithm's functions should be made transparent and its parameters should remain flexible enough to accommodate the requirements of the users' respective research interests. The result displays should be linked to the sources in order to enable differentiated interpretation with the help of direct source study.

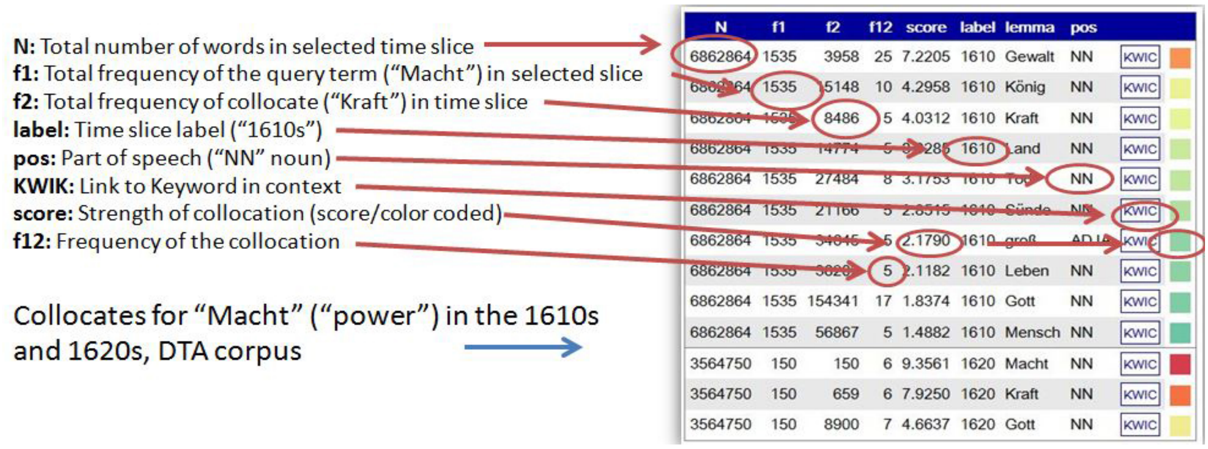
3.1 Output

The output of DiaCollo should be as complete, correct, and reproducible as possible. These requirements derive from a projection of quality standards of analog historical research onto computer-aided research on the one hand, and from generic criteria for software development on the other. Completeness of output is favored by historians because their source materials are usually sparse, fragmentary, and biased. Every detail is important in order to keep the gaps (which have to be reconstructed) as small as possible, so as to be able to draw as well-founded a conclusion as possible from the particular evidence to the bigger picture. So ideally, digital corpora should encompass all sources relevant to a given subject, and an analysis tool should find every instance of a sought-after linguistic phenomenon in the corpus, and include that instance in subsequent calculations. In contrast to techniques like topic modeling, collocation analysis can yield complete, correct and reproducible results in this sense, provided that the corpora are of high data quality and are have undergone appropriate and sufficiently accurate NLP preprocessing (e.g. tokenization, lemmatization, etc.).

Paradoxically, this preprocessing both guarantees and compromises the correctness-criterium. Individual- or time-specific peculiarities such as orthographic features, foreign-language insertions, illegible typescript or handwriting, the choice of words, a gap, an addition, a typographical or spelling error – can be decisive in assessing the authenticity and significance of a source, and might indeed be precisely the phenomenon sought after in a specific research design. Consequently, this type of information should not be lost when working with digital tools. But in order to support search functions in the text, the peculiarities and "quirks" must be leveled and the text normalized, which is diametrically opposed to the requirements of source criticism. The *Basisformat*[6] ("base format") of the *Deutsches Textarchiv* ("German Text Archive", DTA) provides a set of effective and field-tested guildelines for corpus annotation recommended by the *Deutsche Forschungsgemeinschaft* ("German Research Foundation", DFG). Digital source criticism must encompass an understanding of these labor-intensive data curation efforts, both of their limits and of the opportunities they provide by enabling complex queries with search engines such as DDC.[7]

[6] http://www.deutschestextarchiv.de/doku/basisformat/

[7] http://www.deutschestextarchiv.de/doku/software#ddc , http://kaskade.dwds.de/dstar/dta/diacollo/help.perl#queries


N: Total number of words in selected time slice
f1: Total frequency of the query term (“Macht”) in selected slice
f2: Total frequency of collocate (“Kraft”) in time slice
label: Time slice label (“1610s”)
pos: Part of speech (“NN” noun)
KWIK: Link to Keyword in context
score: Strength of collocation (score/color coded)
f12: Frequency of the collocation
Collocates for “Macht” (“power”) in the 1610s and 1620s, DTA corpus

N	f1	f2	f12	score	label	lemma	pos		
6862864	1535	3958	25	7.2205	1610	Gewalt	NN	KWIC	🟧
6862864	1535	15148	10	4.2958	1610	König	NN	KWIC	🟨
6862864	1535	8486	5	4.0312	1610	Kraft	NN	KWIC	🟨
6862864	1535	14774	5	[illegible]	1610	Land	NN	KWIC	🟩
6862864	1535	27484	8	3.1753	1610	Tod	NN	KWIC	🟩
6862864	1535	21166	5	2.8515	1610	Sünde	NN	KWIC	🟩
6862864	1535	34045	5	2.1790	1610	groß	ADJA	KWIC	🟩
6862864	1535	3820	5	2.1182	1610	Leben	NN	KWIC	🟩
6862864	1535	154341	17	1.8374	1610	Gott	NN	KWIC	🟩
6862864	1535	56867	5	1.4882	1610	Mensch	NN	KWIC	🟩
3564750	150	150	6	9.3561	1620	Macht	NN	KWIC	🟥
3564750	150	659	6	7.9250	1620	Kraft	NN	KWIC	🟧
3564750	150	8900	7	4.6637	1620	Gott	NN	KWIC	🟨

Figure 1: Annotated screenshot of DiaCollo's tabular HTML display format

3.2 Visualization

In addition to the citation forms of the *k*-best collocates discovered per epoch, DiaCollo's default tabular HTML output (Figure 1) includes a number of additional data columns for each collocate row. This supplementary data represents the empirical basis by means of which DiaCollo computes the association strength between the user’s search term(s) and the collocate item in question. In particular, each collocate row contains the minimum date (“label”) for the current corpus epoch, the total size of the epoch (“N”), the total frequencies of both the user’s search term (“f1”) and the current collocate item (“f2”) in the current epoch, and the final association score itself (“score”) as computed by the selected scoring function.

DiaCollo also offers several interactive visualization formats for diachronic collocation data, including animated bubble charts and tag-clouds, in which each collocate item’s association score is mapped directly to visually salient properties (size and color) of the corresponding display element, and the “raw” empirical corpus frequencies are not displayed by default. In general, such visualizations can facilitate comprehension by reducing the cognitive workload involved in interpreting search results, but may fail to adequately capture all relevant aspects of the underlying data, leading to oversimplification and the danger of “jumping to conclusions” (Jurish, 2016b). DiaCollo’s visualization formats were developed in close collaboration with and in response to the needs of the contributing humanities researchers. Color-coding and differences in font-size allow “intuitive” comparisons of large datasets at a single glance. In the interactive visualizations, detailed information (f1, f2, f12, etc.) about individual collocation pairs can be displayed in a popup window by clicking on a collocate item of interest. At the same time, some of the pitfalls of distant reading (e.g. extractions of single words “solely” on the basis of distributional properties alone) are ameliorated by establishing a direct connection to the underlying text sources. A KWIC (“Keywords in Context”) view of corpus hits for a given collocation pair as well as links to the original texts are offered (provided the requisite preprocessing has been done). DiaCollo thus attempts to allow smooth transitions from distant to close reading in all supported visualization formats.

3.3 Usability

The statistics and algorithms underlying tools such as DiaCollo, as well as corpus query languages, are only starting to be included in the academic curriculum of historians. They are neither trivial nor easy to grasp without this kind of specialized training. DiaCollo allows for free choice of up to 15 different parameters influencing the final result and its method of presentation. The flexibility offered by this large selection of choices is what makes it useful for historians – but not necessarily easy to use. Lifting the obligation to explicitly specify all parameter values by providing sensible defaults and suppressing “raw” frequency data (f1, f2 etc.) from the output presentation might increase users’ initial comfort with the tool, but hinder understanding and interpretation of the results obtained. Thorough documentation, examples, and use-cases are needed for users with different levels of experience. DiaCollo comes with documentation, use cases and tutorials in German and English that aid digital source

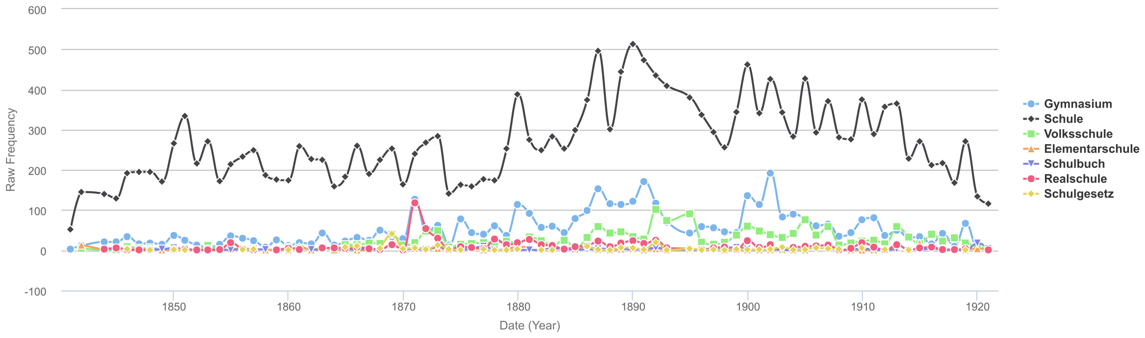


Figure 2: Grenzboten corpus frequencies for school-related terms

criticism[8]. Examples range from simple single-word queries to more elaborate constructions using regular expressions, thesaurus expansion, Boolean operators, and document metadata filters. Even so, there is further demand for finer-grained tutorials, for introductory workshops as provided by members of the CLARIN-D working group "History"[9], and for one-to-one feedback and support as provided by the experts at the BBAW.

4 Use Case: Debates on Education in *Die Grenzboten*

As an introduction to DiaCollo's functionality, we will consider the collocates of a simple search term "*Schule*" ("school") in the largest historical corpus available at the BBAW, the *Deutsches Textarchiv*[10] ("German Text Archive", DTA). Presentation of results in HTML format displays up to the specified number (kbest) of collocates (by default 10) for "*Schule*" discovered within the chosen time slice (e.g. a decade) in the form of a table. Each row of the table includes a color-code indicating the strength of the collocate's association preference as well as links to (close approximations of) the underlying corpus evidence for the corresponding collocation pair as Keywords-in-Context (KWIC), allowing the user to focus her attention more closely on the original text source. Additional visualization modes such as the "bubble" and "cloud" formats display changes in the collocates on an interactive timeline. For the example query, the collocates give quite obvious evidence that the term "school" associated with words within the semantic field of the institution of the church in the earliest documents queried (e.g. in the 1560s: *Kloster* ("cloister"), *Pfarrherr* ("pastor"), and *Kirche* ("church")). The findings imply that the influence of this institution on the school system begins to mingle with worldy institutions in texts from the 1710s, where collocates include *Kirche* ("church"), as well as *Inspektor* ("inspector"), *preußisch* ("Prussian"), and *Universität* ("university"); the term "church" disappears from the lists of top-10 collocates from the 1770s onwards (but re-occurs in the 1840s and 1890s).

We will further demonstrate the use of DiaCollo by looking at German education policy as discussed in a historical periodical. This is a typical use case for historical research: determining to what extent this particular corpus is relevant for specific research questions, e.g. for historical textbook research as conducted by researchers at institutions like the Georg-Eckert-Institute (GEI). *Die Grenzboten* was a German-language national-liberal magazine published from 1841 to 1922, covering a wide range of subjects in politics, literature, and the arts throughout the 'long' nineteenth century (Werner, 1922). Its "messengers from the borders" did not limit their reports to German-speaking territories, but also addressed issues in other European countries and the rest of the world. Its coverage of civic life, opinions, and debates before and after the revolution of 1848, the restoration period, industrialization, the German Empire (*Kaiserreich*), and the First World War makes this periodical a valuable source for a broad range of disciplines.

Originally published as a (bi-)weekly periodical, the 311 volumes (roughly 180,000 pages) of *Die Grenzboten* were first digitized by the Staats- und Universiätsbibliothek Bremen[11] with funding from

[8] Tutorial in German: http://kaskade.dwds.de/diacollo-tutorial; DiaCollo help in English: http://kaskade.dwds.de/dstar/dta/diacollo/help.perl; Short guide in English and German: https://www.clarin-d.net/en/diacollo-en, https://www.clarin-d.net/de/kollokationsanalyse-in-diachroner-perspektive

[9] https://www.clarin-d.net/en/disciplines/history

[10] http://kaskade.dwds.de/~jurish/cac2019/Schule-dta

[11] University of Bremen: Grenzboten project, http://www.suub.uni-bremen.de/ueber-uns/projekte/grenzboten/

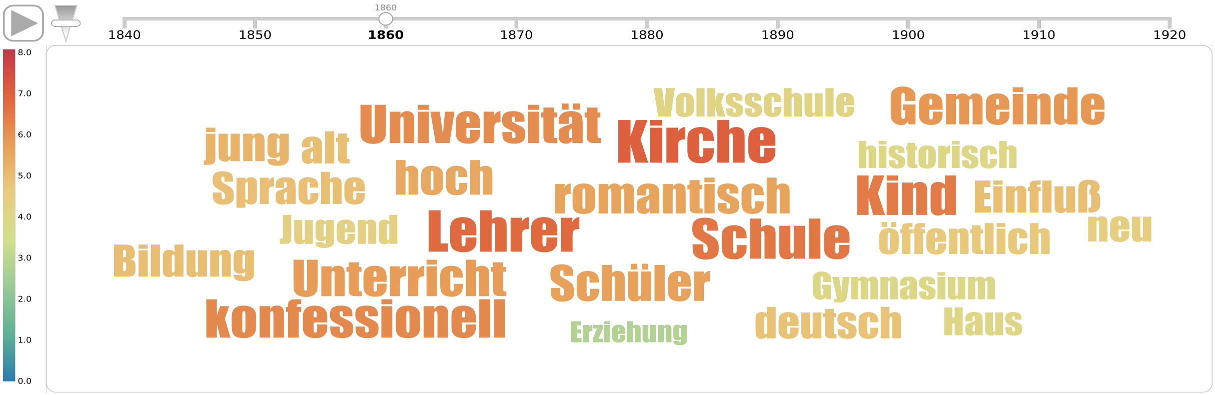


Figure 3: DiaCollo 'cloud' visualization of "Schule" collocates in the Grenzboten corpus (1860-1869)

the German Research Association (DFG), and have been integrated into the BBAW CLARIN service center's corpus infrastructure. Over the course of its publication, *Die Grenzboten* was witness to several changes and attempted reforms of school systems in German-speaking territories. Its political agenda changed over time as well: the publication's original focus on Austria shifted to Prussia when its founder Ignaz Kuranda (1811-1884) left the editorial team in 1848. After 1870, *Die Grenzboten* increasingly became a voice for conservative views. But throughout its lifetime, the magazine self-identified as part of the liberal opposition which favored German unity under Prussian leadership, the so-called "small German solution" (von Wurzbach-Tannenberg, 1865; Werner, 1922). Using DiaCollo, we will now explore *Die Grenzboten*'s stance on education policy.

4.1 Is the corpus a source for research into the history of education?

A time series analysis of the absolute frequency of selected relevant terms such as *Schule* ("school"), *Schulgesetz* ("school law"), *Schulbuch* ("textbook"), and other terms denoting various types of German schools shows that the lemma "school" was indeed mentioned in every year of *Die Grenzboten*'s publication[12] (Figure 2). Its raw frequency peaked at more than 500 tokens in 1890, and its relative frequency in the *Die Grenzboten* corpus is twice as high[13] as in the corresponding texts (1840–1920) from the aggregated DTA and *Digitales Wörterbuch der deutschen Sprache* (DWDS)[14] "core" corpus. A DiaCollo search[15] for collocates of *Schule* in ten-year epochs beginning at 1840 provides ample results from which to explore the school-related topics discussed in *Die Grenzboten*. Of the top-10 collocates per decade, most are nouns, some adjectives and one a finite verb (*gehören,* "to belong").

4.2 Are all findings relevant? Disambiguation by targeted close reading

DiaCollo's KWIC facility allows one to quickly check whether the results are applicable to a particular research question. In this case, strong adjective collocates of *Schule* are often associated with the sense of "school" as "doctrine", e.g. an artistic school or school of thought, which is irrelevant when looking at education policy. Another adjective collocate of interest is the lemma *hoch* ("high"). In DiaCollo's interactive 'cloud' visualization for this query[16], it becomes evident that the adjective already appeared among the ten best collocates per epoch after 1870, and was strongly associated with *Schule* throughout the entire corpus (Figure 3). Examination of the corresponding KWIC hits reveals that these collocates refer almost exclusively to secondary ("higher") schools, both historic and contemporary, in German-speaking countries and elsewhere. Quite often, the encompassing articles deal with access granted or denied to higher education. The query results lack an antonym – this may be a result of it being distributed over a larger field of words (such as "basic", "elementary", or "primary"), but also supports the impression that *Die Grenzboten* was on the whole more concerned with higher education than with the *Volksschule* which provided basic primary (rural) education.

[12] http://kaskade.dwds.de/~jurish/cac2019/Schule-ts
[13] http://kaskade.dwds.de/~jurish/cac2019/hist-gb
[14] http://kaskade.dwds.de/~jurish/cac2019/hist-dta+dwds
[15] http://kaskade.dwds.de/~jurish/cac2019/Schule-gb
[16] http://kaskade.dwds.de/~jurish/cac2019/Schule-cloud

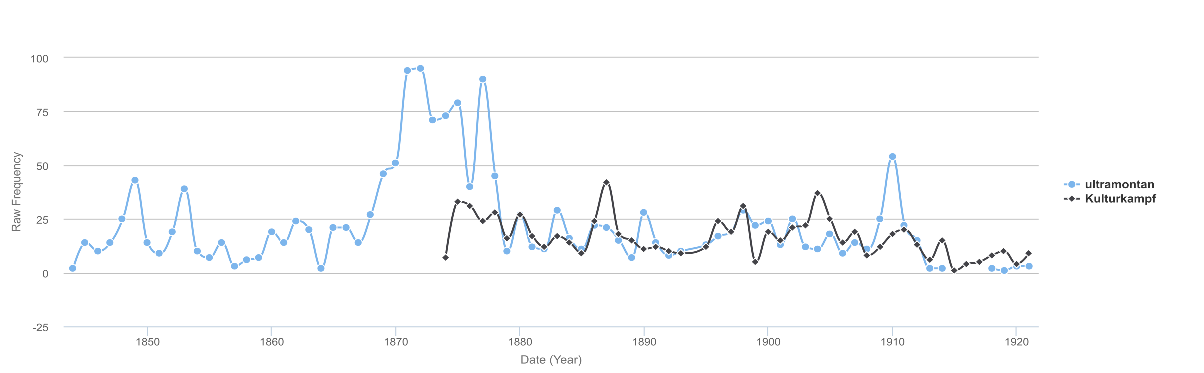


Figure 4: Grenzboten corpus frequencies for "ultramontan" and "Kulturkampf"

4.3 Do the findings offer tracks to specific discourses/debates?

Finally, the adjectives *konfessionell* ("denominational") and *öffentlich* ("public") were examined. These collocates appear among the top ten between 1860 and 1879, as do the nouns *Gemeinde* ("parish"/"congregation") and *Kirche* ("church") – the latter being as prominent and persistent as more expected noun collocates such as *Kind* ("child") or *Lehrer* ("teacher"). Using DiaCollo's on-the-fly filtering function to restrict our attention to adjective collocates only[17], the 1860s and 1870s documents reveal the adjectives *protestantisch* ("protestant") and *evangelisch* ("evangelical") as well as *katholisch* ("catholic") as strong collocates of *Schule*.

We may assume that the prominence of this terminology involving religious denominations at that particular time was caused by the contemporary debates – since referred to as the *Kulturkampf* ("cultural struggle") – concerning the rights and spheres of influence of state (Prussia) and church (Pope Pius IX) which started in some German territories in the 1860s and reached their peak in the 1870s. The debates involved the issue of who should be in charge of education and curricula, and how to deal with different religious denominations in schools. Loyal supporters of the Roman Catholic Church were referred to as *ultramontan* ("ultramontane") during this period. A simple frequency query[18] (Figure 4) shows that this kind of terminology is indeed present in the *Grenzboten* corpus, the former peaking and the latter beginning in the 1870s.

Among the strong collocates of *Kulturkampf* [19]and *ultramontan*[20] are no terms that would hint at education, though. A manual check of the sources can be time consuming; in this case they do indeed yield results relevant for debates on education. Even in cases where the interpretation seems straightforward – as in *Kulturkampf … entbrannt* ("cultural struggle … erupted") – a look at the sources is not superfluous. In this case, a parliamentary debate is commented upon. According to the author, the cultural struggle erupted "in the master's mansion" (*im Herrenhaus*) too, after a heated parliamentary discussion of several ultramontan petitions concerning religious education in different school types and a discussion of the slogan "*Trennung von Kirche und Schule*" ("separation of church and school"; *Die Grenzboten*, "Vom Preußischen Landtage", p. 237f).

The connection of the cultural struggle with debates on education only becomes clear at the level of collocates if we turn our attention to all co-occurrences of *ultramontan* and GermaNet (Hamp & Feldweg, 1997; Henrich & Hinrichs, 2010) hyponyms of the synset *Bildungseinrichtung* ("educational institution") or compounds matching a simple regular expression and using a rather broad paragraph-wide search window. Through closer reading of the corpus hits, we find evidence for anti-Catholic opinions in debates about education emanating from various sources.[21] In an article celebrating 30 years of the Gustav-Adolf Verein (today 'Gustav-Adolf-Werk') – which aided protestants living away from larger congregations (in diaspora) – we find for example the following "activity report":

> "*Tausende von Kindern evangelischer Eltern wurden durch ihn [den Verein] der katholischen Schule, in die sie nothgedrungen gehen mußten, entnommen und so vor den Nachstellungen der ultramontanen Propaganda bewahrt.*" [*Die Grenzboten,* "Der Gustav Adolf Verein", p. 503f]

[17] http://kaskade.dwds.de/~jurish/cac2019/Schule-gb-adj
[18] http://kaskade.dwds.de/~jurish/cac2019/ultramontan-freq
[19] http://kaskade.dwds.de/~jurish/cac2019/Kulturkampf-collocates
[20] http://kaskade.dwds.de/~jurish/cac2019/ultramontan-collocates
[21] http://kaskade.dwds.de/~jurish/cac2019/ultramontan-germanet

> ("Thousands of protestant parents' children were taken [by the Gustav-Adolf-Society] from the Catholic schools they had been forced to attend, and thus spared from the harassments of ultramontane propaganda.")

So even if this important part of the debates on education policy was not immediately apparent in the results of our initial DiaCollo queries, the subsequent indications combined with informed curiosity and further investigation (by means of focused queries and close reading of sources) produces more satisfying results.

5 Conclusion

DiaCollo serves as an effective automatic tool for the analysis of semantic change with respect to terms and concepts in diachronic perspective. Designed and optimized for the needs of humanities researchers, DiaCollo's expressive query language and flexibility make it a useful aid for corpus exploration and research. Thorough documentation, tutorials, and references to previous work as well as user-oriented dissemination in the form of workshops and lectures by the CLARIN-D working group "History" make it easier for the inexperienced to learn and provide a useful resource for more experienced users. Actively maintained and supported as part of the ongoing development cycle, DiaCollo continues to evolve and adapt in response to and in co-operation with its user community. Our use cases have shown the necessity of constant shifts between close and distant reading methods, which DiaCollo facilitates. Although ensuring interoperability between tools and maintaining the high standards of data curation necessary for reliable results requires considerable effort across all disciplines, we believe the prospective gain for the scientific community will justify the endeavor.

References

Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, pages 1212–1248. Mouton de Gruyter, Berlin.

John Rupert Firth. 1957. *Papers in Linguistics 1934–1951*. Oxford University Press, London.

1862. Der Gustav-Adolf-Verein. *Die Grenzboten*, 21:502–515. http://brema.suub.uni-bremen.de/grenzboten/periodical/titleinfo/114291

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, Spain.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT – the GermaNet editing tool. In *Proceedings LREC 2010*, pages 2228–2235, Valletta, Malta.

Bryan Jurish, Alexander Geyken, and Thomas Werneke. 2016. DiaCollo: diachronen Kollokationen auf der Spur. In *Proceedings DHd 2016: Modellierung – Vernetzung – Visualisierung*, pages 172–175.

Bryan Jurish. 2015. DiaCollo: On the trail of diachronic collocations. In Koenraad De Smedt, editor, *Proceedings CLARIN Annual Conference 2015*, pages 28–31, Wrocław, Poland.

Bryan Jurish. 2018. Diachronic collocations, genre, and DiaCollo. In Richard Jason Whitt, editor, *Diachronic Corpora, Genre, and Language Change*, pages 42–64.

Maret Nieländer and Andreas Weiß. 2018. »Schönere Daten« – Nachnutzung und Aufbereitung für die Verwendung in Digital-Humanities-Projekten. In Maret Nieländer and Ernesto William De Luca, editors, *Digital Humanities in der internationalen Schulbuchforschung*, pages 91–116. V&R unipress, Göttingen.

1878. Vom preußischen Landtage – Berlin, 27. Januar. *Die Grenzboten*, 37:234–239. http://brema.suub.uni-bremen.de/grenzboten/periodical/pageview/139535

Constantin von Wurzbach-Tannenberg. 1865. Kuranda, Ignaz. In *Biographisches Lexikon des Kaiserthums Oesterreich*, volume 13, pages 407–416. Staatsdruckerei, Vienna.

Fritz Werner. 1922. Die Grenzboten: aus der Geschichte einer achtzigjährigen Zeitschrift nationaler Bedeutung. *Die Grenzboten*, 81:448–452. http://brema.suub.uni-bremen.de/grenzboten/periodical/titleinfo/178709

Data Collection for Learner Corpus of Latvian: Copyright and Personal Data Protection

Inga Kaija
Institute of Mathematics and Computer Science,
University of Latvia;
Riga Stradiņš University, Latvia
inga.kaija@rsu.lv

Ilze Auziņa
Institute of Mathematics and Computer Science,
University of Latvia,
Riga, Latvia
ilze.auzina@lumii.lv

Abstract

Copyright and personal data protection are two of the most important legal aspects of collecting data for a learner corpus. The paper explains the challenges in data collection for the learner corpus of Latvian "LaVA" and describes the procedure undertaken to ensure protection of the texts' authors' rights. An agreement / metadata questionnaire form was created to inform the authors of the ways their texts are used and to receive the authors' permission to use them in the stated way. The information, permission, and the metadata questionnaire are printed on one side of an A4 size paper sheet, and the author is supposed to write the text on the other side by hand, thus eliminating the need to identify the author of the text separately. After scanning and adding to the corpus, the text originals are returned to the authors.

1 Introduction

Learner corpora have become increasingly popular, and the demand for such corpora to become available to a wider scope of researchers is growing. However, the creation of publicly available learner corpora includes dealing with personal data protection and copyright issues. A learner corpus of Latvian "LaVA" (Latvian Council of Science Grant Development of Learner corpus of Latvian: methods, tools and applications. No. lzp-2018/1-0527) is being created, and it will be publicly accessible, so these legal issues have to be addressed while still enabling researchers to collect relevant metadata about possible factors impacting language learning outcomes.

The "LaVA" creation is divided into several stages: 1) data collection, 2) data digitization; 3) text correction; 4) automated NLP analysis (morphological analysis); 5) original and corrected text alignment; 6) automatic error annotation and manual review. At least 1000 essays on different topics from students with different language backgrounds are planned to be included in the LaVA corpus.

The initial stage of the project covers development of a methodology for data collection and digitization, development of methodology and guidelines for error annotation, and corpus platform development. Among the most important tasks of this phase were the legal and ethical solutions for the text collection process.

Copyright and personal data protection are two of the most important legal aspects that should be resolved before data collection for the learner corpus is started. Therefore, an agreement and metadata questionnaire form was developed to inform the authors of the inclusion of their works in the corpus and to obtain authors' permission.

There have been efforts to create templates for contracts to help deal with the copyright issues when collecting data for research.[1] While they can be extremely helpful, in the case of creating a learner corpus a more specific compact document is useful where the exact aims and rules of using the texts are

[1] For example, see http://www.meta-net.eu/meta-share/licenses

Inga Kaija and Ilze Auziņa 2020. Data Collection for Learner Corpus of Latvian: Copyright and Personal Data Protection. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 41–47.

described. The copyright issues might be similar over all kinds of corpora, but the very nature of learner texts makes also anonymity particularly important – not only that of the people mentioned in the texts, but also that of the authors. In many cases, the learners feel self-conscious about their language skills and want their identity to be protected, especially knowing that the data will be available to the public. This, in turn, makes it necessary to specifically agree on the kinds of data the learners provide and the ways they are used.

To protect learners' rights when collecting their texts, an agreement / metadata questionnaire and the procedure of text collection was developed. The present paper lists the main legal and ethical principles considered and describes how the data collection process is carried out.

2 Regulations

The learners, i.e., authors of the texts collected for the corpus, come from various backgrounds and belong to various countries in Europe and outside of it. However, their studies of Latvian (including text writing process) and corpus creation take place in Latvia, so the legislature of Republic of Latvia applies. The regulations regarding personal data protection and copyright issues that concern learner corpus creation in Latvia have been previously described in comparison with the relevant regulations in Lithuania (Znotiņa, 2016). We further list the main legal documents and principles to be observed in each of those areas.

2.1 Copyright

The main document regulating copyright protection in Republic of Latvia is the Copyright Law (AL, 2000), and it states that:

- texts written as a part of study process are protected by copyright, unless otherwise stated in the study agreement between the author and the study institution;
- in order to make the text (or part of it) available to the public, a written permission must be received from the author;
- the author has the right to decide to be recognized as an author and to decide when, how many times etc. the work can be accessed.

In order to comply with the regulations, the corpus creators have to make it possible for the authors to express their decision explicitly. However, providing the authors with various choices would make the corpus creation process extremely complicated, as all of the different choices would have to be taken into account, especially considering that each of the submitted texts is added a separate consent. Therefore, it was decided that a standardized form for all authors of the texts in corpus must be created. Those authors who would not agree with the common terms could opt out of participating in the project altogether.

2.2 Personal data protection

Protection of personal data in Republic of Latvia is regulated by the Personal Data Processing Law (FPDAL, 2018) as well as one of the most influential regulations regarding personal data protection in European Union, the European Union's new General Data Protection Regulation (Regulation EU 2016/6791), enforced on May 25 2018 (GDPR 2016). Both of them emphasize the ability to identify a person as a criterion for defining personal data. GDPR states that personal data "means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" (GDPR 2016).

The corpus "LaVA" is a beginner learner corpus, and the topics offered for writing to the beginner students often inherently include telling one's own or other people's data (e. g. "Me and my family"). Thus, it is of utmost importance to eliminate the possibility that such personal data would be made publicly available. It can be done by anonymizing the stored data containing personal information, or by avoiding the disclosure of any personal data. In this case, the topics of the texts often require a significant amount of personal data to be included, and anonymization could quickly become a daunting task. The corpus creators decided to avoid it by requesting the authors of the texts to not include any real personal data and to replace it with imaginary ones instead. This request also has to be included in the form signed by the authors as a part of the conditions for including the texts into the corpus.

Some concerns have been expressed that handwriting can also be seen as personal data as it can be used to recognize the writer. Latvian legislation does not seem to address this issue specifically but it has to be considered when a corpus includes scanned copies of a handwritten text. Here are a few important aspects at play:

- There are many authors of the texts which leads to many examples of similar handwritings. The corpus is expected to contain at least 1000 texts, and each student is offered to submit a text no more than once per semester, for two semesters at most. Therefore, even if all students participated in the project twice (which is not the case), there would still be at least 500 different authors altogether. It makes recognizing someone by handwriting alone highly unlikely.
- The only real information provided about the author is the metadata: gender, other languages spoken, and age at the time of writing (time of writing is between 2018 and 2020, but it is not specified more precisely for any of the texts). In case unusual combinations are found, the corpus creators discuss not including the text into the corpus because unique metadata (such as a rare mother tongue) may give little quantifiable insight into the language learning process in general. This minimizes the possibility of recognizing a person by their handwriting and metadata combination,

During the corpus creation process, the text is seen by the author, the teacher, and no less than three people of the corpus creator team. If any of those people express doubts about the possibility to recognize the author by their handwriting (e. g. the handwriting looks unusual, distinguishable from most), the possibility to not include the text is considered. The amount of texts provided by the participating higher education institutions is large enough that it does not add any pressure on the corpus creators to try including as many texts as possible at the expense of authors' rights protection.

3 Data collection for the learner corpus of Latvian

The main principles of the agreement / questionnaire form are the same ones already used in the learner corpus of the second Baltic language "Esam"[2] (Znotiņa, 2018), but data collection is carried out in a different way. In "Esam", the permissions to use the data were acquired long after the texts were written (in some cases, several years), and all texts were additionally anonymized. In the case of "LaVA", the learners know the texts are going to be included in the corpus when they write them. Besides, the texts in "LaVA" are not further anonymized by the project team, and the data is collected by various people, so the procedure is regulated more strictly in order to maintain uniformity in the received data and information given to the authors.

3.1 Contents of the agreement / questionnaire

An agreement / questionnaire form was created for data collection of the corpus "LaVA". It is written in English because English is used as an intermediary language in studies of Latvian as a foreign language in the higher education institutions of Latvia, so all authors speak this language well. The form is offered to all authors of the texts expected to be included into the corpus, and every text is only included into the corpus after a signed copy of the form is received from the author. The texts are collected from the learners of Latvian in the 1st or 2nd semester of their Latvian language course. If one author submits more than one text (one text during the 1st semester, another one in the 2nd semester), each texts needs to have its own questionnaire filled.

The form is printed on one side of an A4 size paper sheet (for layout, see Picture 1) and includes three parts – an information letter, a permission form, and a metadata collection questionnaire (information about the author).

[2] Available online: http://www.esamkorpuss.lv

The former consists of:

- basic information about the project, the institutions that are carrying it out, and contact information;
- brief instructions for the learner;
- information about the security of data on the server used for the corpus and privacy;
- explanation on expressing one's will regarding participation in the project (i.e. what to do if the author decides they no longer want their texts to be used in the corpus).

The permission includes seven statements the author agrees to comply with by signing the form:

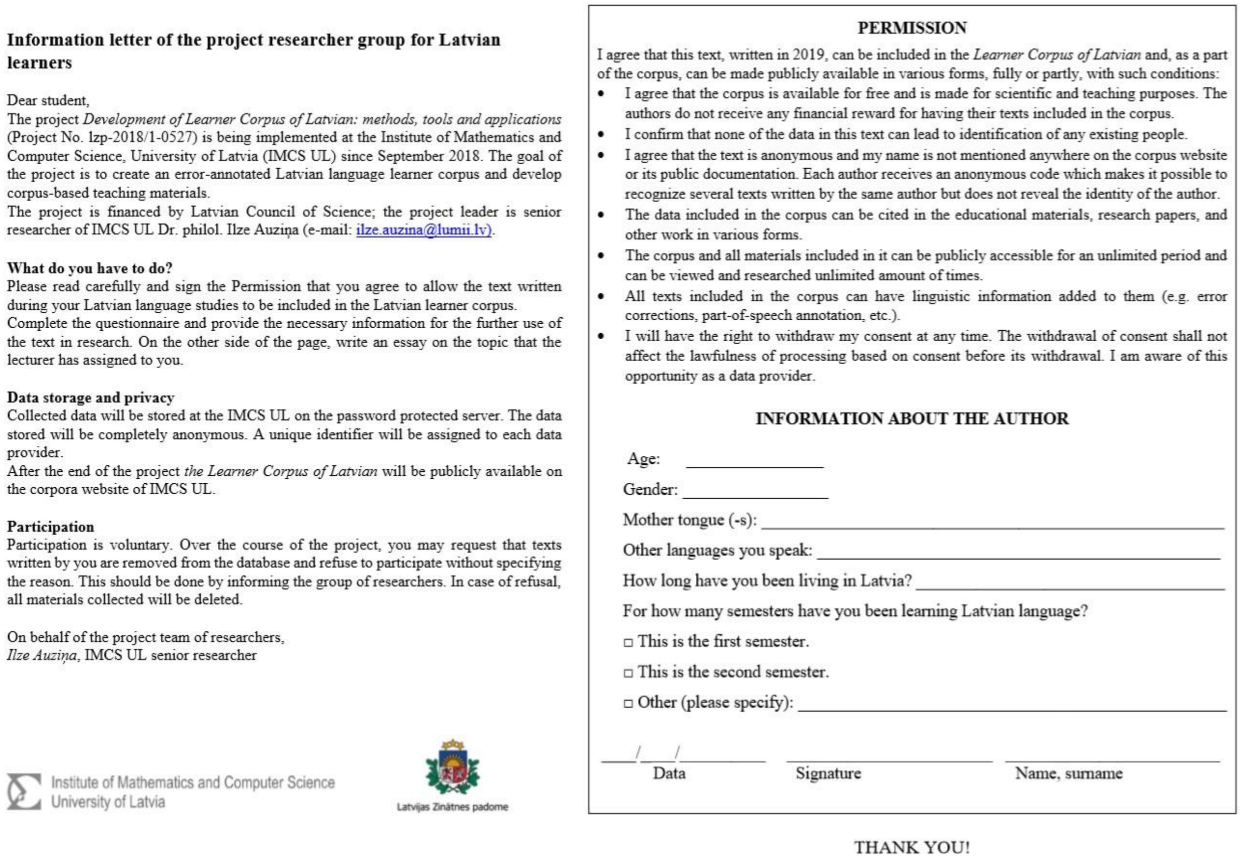
- The author agrees that the corpus is available for free and is made for scientific and teaching purposes. The authors do not receive any financial reward for having their texts included in the corpus.
- The author confirms that none of the data in this text can lead to identification of any existing people. This condition is also particularly stressed when instructing the students; not all of them have a clear understanding what personal data is, so teachers who participate in the project sometimes explain the concept and help deciding what kind of data must be replaced.
- The author agrees that the text is anonymous and their name is not mentioned anywhere on the corpus website or its public documentation. While copyright issues are often solved by crediting the author, the standard solution was decided to be anonymity. There may be some authors who would not mind their names to be associated with their texts, but, the more authors are known, the easier it is to recognize the others who do not want it. The questionnaire also states that each author receives an anonymous code which makes it possible to recognize several texts written by the same author but does not reveal the identity of the author. However, it was later decided that associating several texts with the same author would not give enough research possibilities. Moreover, it would potentially enable one to recognize an author based on the combined contents of the texts, thus undermining the anonymity and personal data protection factors in play.
- The data included in the corpus can be cited in the educational materials, research papers, and other work in various forms.
- The corpus and all materials included in it can be publicly accessible for an unlimited period and can be viewed and researched an unlimited amount of times.
- All texts included in the corpus can have linguistic information added to them (e.g. error corrections, part-of-speech annotation, etc.).
- The author will have the right to withdraw their consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. The author is aware of this opportunity as a data provider.

Finally, the metadata collection questionnaire requests the author to provide some information about factors that may influence their target language production: age, gender, mother tongue(-s), other spoken languages, the length of residence in Latvia, and the number of semesters studying Latvian language in a higher education institution.

The date, signature, name, and surname of the author is needed to ensure the author's full agreement with the aforementioned statements in the permission, but is not included in the metadata of the corpus. In case any of the authors later decided to revoke their consent, their name could also be used to find the text that should be deleted.

It is important to note that the information, permission agreement, and the metadata questionnaire are integrated into one document which is then printed on one side of an A4 size paper sheet. The other side of the form is blank, and authors are requested to hand-write an essay there. This eliminates the need to have any identifying information in or around the text. Since the texts are given back to their authors to ensure educational feedback, it is important for teachers to know who should receive which one of the texts. If the text were written on a separate piece of paper, it would therefore require some kind of identification that would complicate the process of avoiding personal data inclusion. The length of texts normally does not exceed the amount that is easily fitted on an A4 size paper sheet because the 1st and 2nd semester (level A1 or A2) students are usually not writing extended essays yet. This approach may not be suitable if longer texts (probably in higher language skill level) are collected. Any alternative that does not complicate matters is possible; in "LaVA", some students who prefer to write on different paper or who needed more than one sheet of paper, stapled the text (on one piece of paper) and form (on another one) together before submitting.

When the questionnaire is completed and the essay is written, both sides of the page are scanned, and the data is further used in building the corpus. The scanned copy of the written text becomes an integral part of the corpus.



Information letter of the project researcher group for Latvian learners

Dear student,
The project *Development of Learner Corpus of Latvian: methods, tools and applications* (Project No. lzp-2018/1-0527) is being implemented at the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL) since September 2018. The goal of the project is to create an error-annotated Latvian language learner corpus and develop corpus-based teaching materials.
The project is financed by Latvian Council of Science; the project leader is senior researcher of IMCS UL Dr. philol. Ilze Auziņa (e-mail: ilze.auzina@lumii.lv).

What do you have to do?
Please read carefully and sign the Permission that you agree to allow the text written during your Latvian language studies to be included in the Latvian learner corpus.
Complete the questionnaire and provide the necessary information for the further use of the text in research. On the other side of the page, write an essay on the topic that the lecturer has assigned to you.

Data storage and privacy
Collected data will be stored at the IMCS UL on the password protected server. The data stored will be completely anonymous. A unique identifier will be assigned to each data provider.
After the end of the project *the Learner Corpus of Latvian* will be publicly available on the corpora website of IMCS UL.

Participation
Participation is voluntary. Over the course of the project, you may request that texts written by you are removed from the database and refuse to participate without specifying the reason. This should be done by informing the group of researchers. In case of refusal, all materials collected will be deleted.

On behalf of the project team of researchers,
Ilze Auziņa, IMCS UL senior researcher

Institute of Mathematics and Computer Science
University of Latvia

Latvijas Zinātnes padome

PERMISSION

I agree that this text, written in 2019, can be included in the *Learner Corpus of Latvian* and, as a part of the corpus, can be made publicly available in various forms, fully or partly, with such conditions:

- I agree that the corpus is available for free and is made for scientific and teaching purposes. The authors do not receive any financial reward for having their texts included in the corpus.
- I confirm that none of the data in this text can lead to identification of any existing people.
- I agree that the text is anonymous and my name is not mentioned anywhere on the corpus website or its public documentation. Each author receives an anonymous code which makes it possible to recognize several texts written by the same author but does not reveal the identity of the author.
- The data included in the corpus can be cited in the educational materials, research papers, and other work in various forms.
- The corpus and all materials included in it can be publicly accessible for an unlimited period and can be viewed and researched unlimited amount of times.
- All texts included in the corpus can have linguistic information added to them (e.g. error corrections, part-of-speech annotation, etc.).
- I will have the right to withdraw my consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. I am aware of this opportunity as a data provider.

INFORMATION ABOUT THE AUTHOR

Age: ______________

Gender: ______________

Mother tongue (-s): ____

Other languages you speak: ____

How long have you been living in Latvia? ______________________________

For how many semesters have you been learning Latvian language?

□ This is the first semester.

□ This is the second semester.

□ Other (please specify): ____

____/____/__________ ______________________ ______________________
Data Signature Name, surname

THANK YOU!

Picture 1: The layout of the agreement / questionnaire form

3.2 Data collection procedure

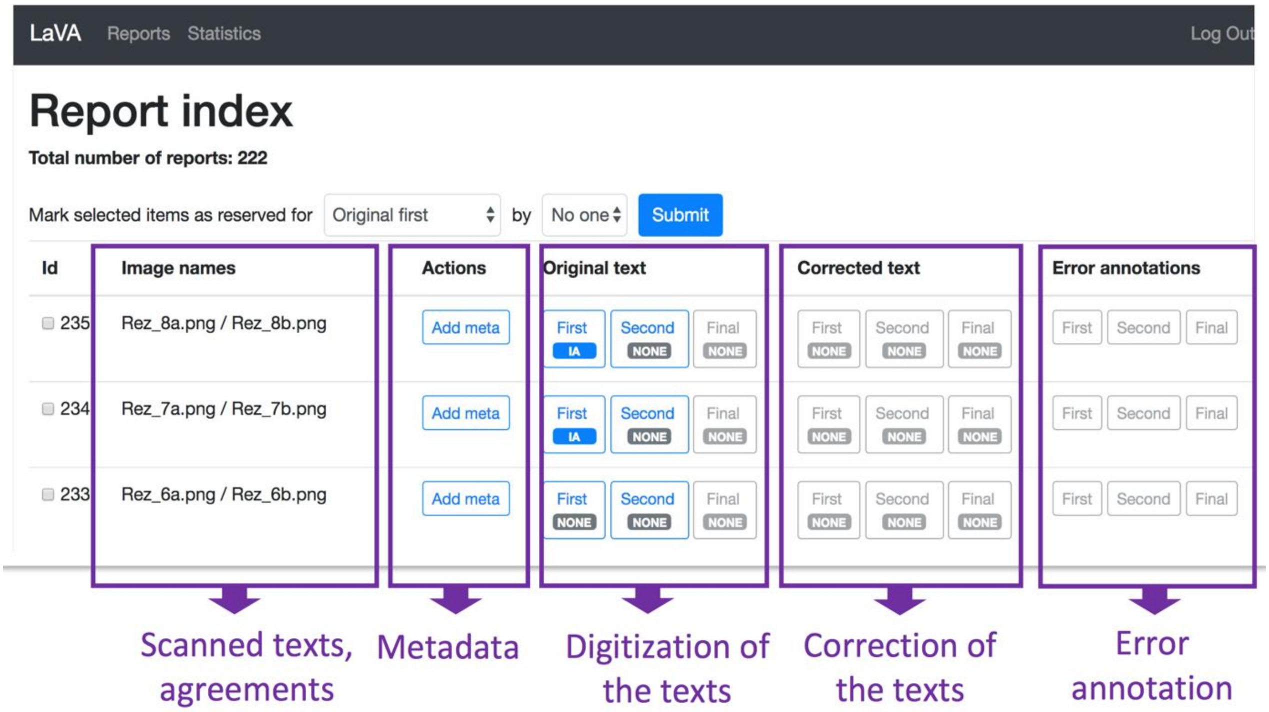
The authors of the texts are all higher education students who have been living in Latvia for a relatively short time and are learning Latvian language at the beginner level for the first or second semester. Teachers are allowed to choose the desired topic and length of the text, and study materials may be used when writing. The teachers who collect the texts instruct the students about the copyright and personal data protection system used in the project, and remind them particularly that regardless of the topic no real personal information should be included in the text. If the topic contradicts this idea (e. g. “My friends and my family”), students are instructed to write about imaginary people or replace the real information with false one.

The preferred text length of each individual text is at least 100 words, as this was decided to be long enough for the learners to be able to use various phrases and constructions which demonstrates their skills of using vocabulary and grammar, as well as other aspects of language use. The maximum length of a text has not been set but rarely exceeds ~270 words.

After the texts are digitized for inclusion in the corpus, the originals are given back to the teacher who corrects them according to the needs of the pedagogical process, and then hands the texts back to the students, once more reminding them about the possibility to revoke the permission if need be (such as accidental inclusion of real personal data etc.).

The corpus is built on an integrated multifunctional platform (Figure 2) that provides a single interface for uploading, digitizing, annotating and search. At the same time, the web platform can also be used for storing scanned copies of essays, comparing texts entered and corrected by two independent digitizers, editing automatically morphological annotated and error-annotated texts, and making inter-annotator agreement.

Collected essays with metadata are handwritten; therefore, they need to be digitized for further data processing steps. The digitization is being carried out in three steps: (1) scanning of the assignments and essays; (2) metadata input; (3) text rewriting in digital format. Scanned images of the assignments help to validate data correctness if any concerns arise. Metadata is entered manually, and the authors' names are not included to retain anonymity.



Report index

Total number of reports: 222

Mark selected items as reserved for Original first by No one Submit

Id	Image names	Actions	Original text	Corrected text	Error annotations
☐ 235	Rez_8a.png / Rez_8b.png	Add meta	First IA Second NONE Final NONE	First NONE Second NONE Final NONE	First Second Final
☐ 234	Rez_7a.png / Rez_7b.png	Add meta	First IA Second NONE Final NONE	First NONE Second NONE Final NONE	First Second Final
☐ 233	Rez_6a.png / Rez_6b.png	Add meta	First NONE Second NONE Final NONE	First NONE Second NONE Final NONE	First Second Final

Picture 2: An integrated multifunctional platform for data uploading, mark-up, annotating and search.

4 Conclusions

The agreement / metadata collection questionnaire form used in the learner corpus "LaVA" is relatively simple and it helps minimise the amount of additional paperwork involved in the creation of the corpus and gives learners a chance to exercise their rights. If any text is suspected to include any real personal data, the author is contacted once more by the teacher / data collector.

The form can be used as a basis for agreements in data collection for other learner corpora in countries which have similar personal data and copyright protection regulations.

Acknowledgements

The work reported in this paper is part of the project *Development of Learner Corpus of Latvian: methods, tools and applications* (Project No. lzp-2018/1-0527) that is being implemented at the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL) since September 2018. The project is financed by Latvian Council of Science.

This work is also a part of the Latvian State Research Programme “Latvian Language” (No. VPP-IZM-2018/2-0002) subproject “Acquisition of Latvian Language” and the European Structural Funds project No. 1.1.1.5/18/I/016 that are being implemented at IMCS UL.

References

[AL 2000] Autortiesību likums, 48/150 (2059/2061), 27.04.2000. [Viewed on April 29, 2019]. Available online: https://likumi.lv/doc.php?id=5138

[FPDAL 2018] Fizisko personu datu apstrādes likums, 132 (6218), 04.07.2018. [Viewed on April 29, 2019]. Available online: https://likumi.lv/ta/id/300099-fizisko-personu-datu-apstrades-likums

[GDPR 2016] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

Inga Znotiņa. 2016. Valodas apguvēju korpuss Latvijā un Lietuvā: autortiesības un personas datu aizsardzība. *Vārds un tā pētīšanas aspekti* 20 (2): 219–227.

Inga Znotiņa. 2018. Otrās baltu valodas apguvēju korpuss: izveides metodoloģija un lietojuma iespējas. Doctoral dissertation. Liepaja : Liepaja University.

Liability of CLARIN Centres as Service Providers: What Changes with the New Directive on Copyright in the Digital Single Market?

Pawel Kamocki
Leibniz-Institut für Deutsche Sprache, Germany
kamocki@ids-mannheim.de

Erik Ketzan
Brikbeck, University of London, United Kingdom
eketza01@mail.bbk.ac.uk

Julia Wildgans
Leibniz-Institut für Deutsche Sprache/ Mannheim University, Germany
j.wildgans@googlemail.com

Andreas Witt
Leibniz-Institut für Deutsche Sprache/ CLARIN ERIC
witt@ids-mannheim.de

Abstract

Providing online repositories for language resources is one of the main activities of CLARIN centres. The legal framework regarding liability of Service Providers for content uploaded by their users has recently been modified by the new Directive on Copyright in the Digital Single Market. A new category of Service Providers, Online Content-Sharing Service Providers (OCSSPs), was added. It is subject to a complex and strict framework, including the requirement to obtain licenses from rightholders for the hosted content. This paper provides the background and effect of these changes to law and aims to initiate a debate on how CLARIN repositories should navigate this new legal landscape.

1 Introduction

One of the main activities of CLARIN centres is to provide online services, such as online repositories, to their users. However, the content uploaded by users of such services can sometimes be of infringing nature. Researchers are well aware of the fact that language resources may violate many rules from copyright and related rights (such as the sui generis database right) through data protection, to rules on defamation and hate speech.

The question of liability for hosting content (such as language resources) uploaded by users of scientific repositories has not attracted the attention that it deserves, despite it being occasionally brought up at conferences (Kamocki, 2014). This may be due to the assumption by scientists that someone who merely provides an online service (e.g. stores data) should not be liable for illegal acts of the service's users. Under current rules, this statement is largely true, and this common-sense point of view has been reflected in the normative framework for almost the past twenty years (i.e. from the beginning of the participative Web).

However, under pure law (be it Roman or common law), service providers could be found liable for prejudice caused by the users. In fact, at the most fundamental level liability requires three elements: breach, prejudice and a causal link between the two (causation). If there is a breach of law (e.g. copyright infringement or unlawful processing of personal data) that causes prejudice, this prejudice can causally be linked to the actions of a service provider. For example, if sensitive information related to a person's health or sexual orientation is communicated to millions of Internet users via an Internet service (e.g. Facebook), the prejudice suffered by the victim is in fact directly caused by Facebook who made this information available to its users.

This does not mean that the user is not liable for his actions — he or she can also be sued for damages, but from the victim's perspective, the service provider would usually be a much better target. Not only is the service provider easier to identify, but also, as a company or an institution, it is expected to be more solvent than an individual user, and possibly also more inclined to settle to avoid damage to its reputation.

In order to promote the development of online services, in the last years of the twentieth century, legislators in both the United States (cf. the Digital Millennium Copyright Act 1998) and the European Union (see below) adopted special rules (called Safe Harbors) protecting service providers from liability for illegal ac-

Pawel Kamocki, Erik Ketzan, Julia Wildgans and Andreas Witt 2020. Liability of CLARIN Centres as Service Providers: What Changes with the New Directive on Copyright in the Digital Single Market?. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 48–52.

tions of their users. Without these rules, services like Facebook, Twitter or YouTube, as well as countless others, could not have been developed. Nowadays, however, things are slowly beginning to evolve, especially in Europe, where users feel that they live in a world dominated by huge, seemingly omnipotent service providers such as Google, Amazon or Facebook. For some, it is time to revise the Safe Harbor provisions in order to protect user interests. This tendency has been visible in the case law of the Court of Justice of the European Union (CJEU) since at least 2016.[1] Another big step in this direction has been made by the recently adopted Directive on Copyright in the Digital Single Market (see below).

In the following sections, we will discuss whether and how this situation may affect smaller service providers, such as CLARIN centres.

2 Liability of Service Providers from e-Commerce to the Digital Single Market

In EU law, 'service provider' is defined as any natural or legal person providing an information society service (Article 2(b) od the Directive 2000/31/EC). An information society service is defined as "any service normally provided for remuneration, at a distance, by electronic means and at the individual request of a recipient of services" (art. 1, Directive 2015/1535). A service is provided "at a distance" if the provider and the recipient are not simultaneously present; it is provided "by electronic means" if the sent and received by means of electronic equipment; finally it is provided "at the individual request of the recipient" if the transmission of data is initiated on such request.

Regarding the "normally provided for remuneration" requirement, recital 18 of the e-Commerce Directive further specifies that information society services "in so far as they represent an economic activity, extend to services which are not remunerated by those who receive them, such as those offering on-line information or commercial communications, or those providing tools allowing for search, access and retrieval of data". The CJEU also opts for a broad interpretation of this requirement.[2]

It is therefore safe to assume that the definition of a service provider covers not only commercial providers, but also e.g. Wikipedia (which is 'paid for' by donators, cf. Angelopoulos, p. 10) or publicly-funded research data repositories.

As explained above, service providers can, under certain conditions, benefit from liability exemptions, which allowed online services to thrive in the first two decades of the 21st century. These exemptions were harmonised by the Directive 2000/31/EC of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (commonly referred to as the e-Commerce Directive). In 2019, the framework was modified by the Directive on Copyright in the Digital Single Market.

2.1 Liability of Service Providers in the e-Commerce Directive

The e-Commerce Directive concerns specifically three types of information society services: mere conduit (art. 12; defined as transmission of information in or provision of access to a network), caching (art. 13; defined as intermediate and temporary storage of information, performed for the sole purpose of making its onward transmission more efficient) and hosting (art. 14). This last category is most relevant for CLARIN Centres.

Hosting is defined as storage of information provided by the user. According to the relevant Safe Harbour provision, a hosting provider is not liable for the stored content if two conditions are cumulatively met:

a) the provider does not have actual knowledge of the illegality of the content AND

b) upon obtaining such knowledge, he acts expeditiously to remove or disable access to the content.

The liability exemption relies therefore on a notice-and-take-down approach: once the provider receives notification (typically from a user) about illegal nature of the content that she hosts, she should act promptly to remove the content from the service. A notification, however, is not the only way (albeit the most common in practice) of acquiring such knowledge - it is also possible that the provider discovers the illegal content through her own investigation, in which case she is also required to remove it promptly.

At the time when the e-Commerce Directive was adopted (2000), hosting providers were mostly merely offering storage space. Nowadays, however, they often play a much more 'active' role in presenting the content - a visitor certainly knows that this particular piece of content is hosted by Facebook, YouTube or Ama-

[1] CJEU, C-160/15, GS Media, 8 September 2016, where it was ruled that providing links to works constituted communication to the public, and therefore copyright infringement, if the provider knows or ought to know about the fact that the works were placed online without permission from the rightholders - if a link is provided for a profit, this knowledge is presumed.

[2] CJEU, C-291/13, Papasavvas, 11 September 2014; in this case, the provider was not remunerated by the recipient, but by income generated by advertisements posted on the website.

zon. While it is possible for an ordinary user to overlook the identity of the content provider, it is usually the service provider who attracts viewers. In other words, with the participative Web, the distinction between the content providers and the service providers became blurred. It is not surprising, therefore, that the liability of the latter is often sought. Faced with the issue of Web 2.0 providers, the CJEU decided to opt for a narrow interpretation of the Safe Harbour provision.

For the CJEU,[3] in order to qualify for the liability exemption, the Service Provider has to meet the criterion set forth in recital 42 of the e-Commerce Directive, according to which its activity has to be "of a mere technical, automatic and passive nature, which implies that the information society service provider has neither knowledge of nor control over the information which is transmitted or stored". This condition is difficult to apply in practice, and the results of its application may be surprising to some. For example, Google was declared to be eligible for the exemption with regards to the GoogleAds service, despite the fact that it actively assists users in the choice of keywords;[4] on the other hand, eBay was declared ineligible for the exemption inasmuch as it exercised some control over the content, actively optimised its presentation and provided tailored services to some users.[5] A publisher of an online newspaper was also denied the benefit of the liability exemption, as it could not claim lack of awareness of the hosted content.[6] Therefore, even though the neutrality requirement set forth by the CJEU may initially seem difficult to meet, its practical interpretation seems to be rather liberal.

This requirement to remain passive can also be criticised for discouraging service providers from taking preventive actions against infringement, for fear of loosing their 'neutrality' and therefore the Safe Harbor privilege (this situation is described as the 'Good Samaritan paradox'; cf. van Ecke).

However, approaches to applying the liability limitation to service providers differ among EU Member States. Germany is, or at least was, one of the most strict. In 2009, the Federal Court of Justice (BGH) developed the doctrine of 'adoption' of content, according to which a provider of a service whose 'core value' is constituted by contents provided by users, and who editorially checks and approves the contents, tags posted contents with his logo and requires the users to grant him extensive re-use rights to the contents, 'adopts' the content as his own. In such a case, the liability exemption cannot apply. Recently, however, the Hamburg court found that YouTube does not 'adopt' the content provided by the users - BGH subsequently referred the case to the CJEU,[7] who should soon deliver its opinion on the question whether YouTube can qualify for the liability exemption or not. In making this decision, the CJEU is not unlikely to be influenced by the new, stricter framework introduced by the DSM Directive (see below), and especially by previous case law which is quite strict for actors who commercially host or provide access to copyright-protected content.

It is important to keep in mind that the Safe Harbour provision only shields hosting providers from liability claims (claims for damages), and not from injunction claims (Article 14(3) of the e-Commerce Directive). Therefore, the providers can still be ordered by a court to remove content.

It should also be noted that Article 15 of the e-Commerce Directive further states that service providers have no general obligation to monitor the content that they store or transmit, or to actively seek facts or circumstances indicating illegal activity.

2.2. Liability of Service Providers in the New Directive on Copyright in the Digital Single Market

The new Directive on Copyright in the Digital Single Market (hereinafter: the DSM Directive) introduced a new category of Service Providers called 'online content-sharing Service Providers'. They are defined as providers of services "of which the main or one of the main purposes is to store and give the public access to a large amount of copyright-protected works or other protected subject matter uploaded by its users, which it organises and promotes for profit-making purposes" (art. 2(6)). YouTube is a typical example of such a service.

The liability of online content-sharing service providers (OCSSPs) is subject to very complex and much stricter rules (art. 17, several pages long). Because the DSM Directive is of very recent vintage, there is no consensus yet on how to interpret these new rules, and their detailed analysis would greatly exceed the allowed length of this paper; only some basic observations can be made here.

[3] CJEU, joined Cases C-236/08 to C-238/08, Google France, 23 March 2010

[4] *idem*

[5] CJEU, C-324/09, L'Oréal v. eBay, 12 July 2011; see also Cour de cassation, eBay v. Dior, 11-10.508, 3 May 2012.

[6] CJEU, C-291/13, Papasavvas, 11 September 2014.

[7] I ZR 140/15

Under the new Directive, the acts of the OCSSPs qualify as communication to the public within the meaning of copyright rules (art. 3 of the Directive 2001/29/CE), and therefore the OCSSP is in principle required to obtain authorisation (a license) from the rightholder (or rightholders) (art. 17(1)). A license obtained by the OCSSP automatically (*ex lege*) covers subsequent communication to the public by the users of the service, provided that it is carried out for non-commercial purposes (art. 17(2)). From the user point of view this seems to mean that anything found e.g. on Youtube can lawfully be shared for non-commercial purposes, which may potentially affect the creation of language resources (although it remains to be seen how this will be implemented in national laws of the Member States). On the other hand, from the OCSSP perspective, especially those hosting content with multiple rightholders (e.g. language resources), the obligation to obtain a license will be very difficult to fulfil. It is also relevant that the DSM Directive also allows Member States to introduce extended collective licensing mechanisms (art. 12), which could facilitate the process of obtaining licenses, but it is too early to say which Member States will adopt this, and how.

The liability limitation for hosting providers under the e-Commerce Directive does not apply to OCSSPs (art. 17(3)). If the OCSSP fails to obtain a license from rightholders, she is liable for copyright infringement, unless she demonstrates that she made 'best efforts' to obtain the license; to make sure that any content for which she obtained a specific notification from rightholders will not be available via her service; upon receiving a notification from rightholders, to act expeditiously to remove and/or disable access to the notified content; and to prevent future uploads of this content (art. 17(4)). This would probably require close cooperation with rightholders, sophisticated mechanisms of content notification with human review (art. 17(9)), as well as screening of uploaded content (which is why, during the adoption process, the opponents of this solution, originally in art. 13, referred to it as 'censorship machines' (Reda)). This may be seen as a contradiction of the rule of art. 15 of the e-Commerce Directive, which expressly states that service providers shall have no general obligation to monitor content; however, art. 17(8) of the DSM Directive expressly states the new framework "shall not lead to any general monitoring obligation".

It is not yet clear what would constitute 'best efforts' that OCSSPs have to make and demonstrate in order to avoid liability. It seems that the standard will be flexible, taking into account the size of the service, the target audience and the type of material uploaded by users, and the costs of implementation of preventive solutions (art. 17(5)). A CLARIN Centre, if it qualifies as an OCSSP (see below), would probably be held to a significantly lower standard than YouTube.

Importantly, art. 17 of the DSM Directive takes into account some copyright exceptions and limitations (art. 17(7)). For example, if a video was made and uploaded by a user within the limits of parody, then the content, even if notified by the rightholder, should not be removed. Limiting the users' rights to rely on exceptions for parody and pastiche, as well as quotation and criticism would indeed seriously impair their freedom of speech. It remains to be seen how this is going to be implemented in practice: screening algorithms obviously are not designed to identify whether 'blacklisted' content (i.e. content that they should prevent from uploading) is used for parody or criticism. If a human reviews and assesses the content, then according to which standards (given that parody cases regularly require to be decided in court)? How long will it take (in a world where, when it comes to criticism and parody, sometimes minutes matter)? Finally, the research exception is not listed among the exceptions that users of online content-sharing services may rely upon, so it seems that research purposes (regardless of how broadly they are defined in the applicable law) cannot be an excuse (at least not *a priori*) for uploading content in online content-sharing services.

Paradoxically the new framework, targeted at huge international OCSSPs like YouTube, will likely provide them with a competitive advantage over smaller rivals. Compliance with the new obligations would require very significant means - means that large OCSSPs may have, but not necessarily their smaller competitors (despite the mechanisms introduced to protect startups — art. 17(6)). It remains to be seen how these new rules will transform the Web.

3 Possible Impact of the New Framework on CLARIN Centres

CLARIN Centres are faced with the difficult task of balancing between the goal of providing quality content and avoiding excessive legal burdens related to liability for the content. Therefore, they should aim at organising their functioning in such a way as to preserve 'neutrality' with regards to the language resources that they host, and not exercising (excessive) control over them in order not to lose the Safe Harbour privilege of the e-Commerce Directive. The national interpretation of the 'neutrality' criterion should be taken into account by each consortium individually. However, under the new legal framework the difficulty does not stop here. As demonstrated above, the new framework regarding liability of OCSSPs is particularly strict and complex, to the point of potentially having a chilling effect on online content-sharing services.

According to art. 2(6) of the DSM Directive, some categories of Service Providers are expressly excluded from the definition of OCSSPs. This is the case of non-for-profit online encyclopaedias (such as Wikipedia), open source software developing and sharing platforms (such as GitHub), online marketplaces (such as OLX or even, arguably, Amazon) as well as "not-for-profit educational and scientific repositories".

It seems that most CLARIN repositories are indeed concerned by this last exclusion, and so they are not OCSSPs and can still qualify for the liability limitation for hosting providers in the e-Commerce Directive.

However, the situation becomes more complicated if a CLARIN repository is used for some sort of commercial (for-profit) purposes, such as charging (even only some categories of users) for access, or use in public-private partnerships. Sometimes it can indeed be very difficult to draw a line between what is 'not-for-profit' and 'for-profit', but crossing this invisible line may have significant consequences as far as liability is concerned.

It may be tempting for CLARIN repositories to use contractual clauses to shield themselves from liability. Indeed, it is good practice to include in the Deposition License Agreement (DLA) an appropriate warranty or liability clause by which the depositor guarantees that the deposition is made lawfully, and assumes liability for damages caused by the content. However, such clauses may be of limited practical significance - they do not liberate the hosting provider from any obligations vis-à-vis the right holder (who is often a legal entity and not a registered user of the repository, so it has no contractual relation with the repository), and the depositor may simply not be solvable enough to compensate the hosting provider for resulting damages. Furthermore, the presence of such clauses in the DLA may have a chilling effect on some potential depositors. It is therefore erroneous to think that the liability conundrum can be solved with simple contractual mechanisms, if they are not accompanied with appropriate technical and organisational measures, such as notice-and-take down procedures, and diligent risk management. Similarly, any formal declaration that the repository remains neutral and passive with regards to the hosted content will not shield against liability, if it does not correspond to reality.

The purpose of this paper is to stimulate debate between CLARIN centres, the Board of Directors and legal experts on how CLARIN repositories should be organised in order to best fit within the existing legal framework regarding liability of service providers. Such a debate does indeed seem necessary.

References

Christina Angelopoulos. 2017. *On Online Platforms and the Commission's New Proposal for a Directive on Copyright in the Digital Single Market*.

Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market.

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

European Parliament. 2017. Providers Liability: From the eCommerce Directive to the future. In-Depth Analysis for the IMCO Committee. Available at: http://www.europarl.europa.eu/RegData/etudes/IDAN/2017/614179/IPOL_IDA(2017)614179_EN.pdf. Accessed 10.02.2020.

Pawel Kamocki. 2014. The Liability of Service Providers in e-Research Infrastructures. Killing the messenger? *Proceedings of the 9th Language Resources Evaluation Conference*, *Reykjavik*.

Julia Reda. Official Blog, https://juliareda.eu/eu-copyright-reform/censorship-machines/. Accessed 10.02.2020.

Patrick van Eecke, Online Service Providers and Liability: A Plea for a Balanced Approach. *Common Market Law Review*, 1455:48.

The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies

Aleksei Kelli
University of Tartu,
Estonia
aleksei.kelli@ut.ee

Arvi Tavast
Institute of the
Estonian Language,
Estonia
arvi@tavast.ee

Krister Lindén
University of Helsinki,
Finland
krister.linden@helsinki.fi

Kadri Vider
University of Tartu,
Estonia
kadri.vider@ut.ee

Ramūnas Birštonas
Vilnius University,
Lithuania
ramunas.birstonas@tf.vu.lt

Penny Labropoulou
ILSP/ARC, Greece
penny@ilsp.gr

Irene Kull
University of Tartu
Estonia
irene.kull@ut.ee

Gaabriel Tavits
University of Tartu
Estonia
gaabriel.tavits@ut.ee

Age Värv
University of Tartu
Estonia
age.varv@ut.ee

Pavel Straňák
Charles University,
Czechia
stranak@ufal.mff.cuni.cz

Jan Hajic
Charles University,
Czechia
hajic@ufal.mff.cuni.cz

Abstract

The authors address the legal issues relating to the creation and use of language models. The article begins with an explanation of the development of language technologies. The authors analyse the technological process within the framework copyright, related rights and personal data protection law. The authors also cover commercial use of language models. The authors' main argument is that legal restrictions applicable to language data containing copyrighted material and personal data usually do not apply to language models. Language models are generally not considered derivative works. Due to a wide range of language models, this position is not absolute.

Aleksei Kelli, Arvi Tavast, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Värv, Pavel Straňák and Jan Hajic 2020. The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies. *Selected papers from the CLARIN Annual Conference 2019.* Linköping Electronic Conference Proceedings 172: 172 53–65.

1 Introduction

The development of language technologies (LTs) relies on the use of language data (LD). Language data is often covered with several tiers of rights (copyright, related rights, personal data rights). Their use can be based on a contractual (e.g. licence, contract, terms of use/service, etc.) or exception model as regards IPR and a consent or exemption model concerning personal data.

The current paper discusses the impact of language data's legal regime on LTs. The question is whether legal restrictions applicable to language data apply to the language technologies that are developed using them as well. The authors analyse how far, in the pipeline of developing language technologies, the original copyright and personal data protection regulations apply. If we take a recorded phone call, for instance, it is evident that copyright and data protection apply to a copy of that recording. At the other extreme, it is equally apparent that they do not apply to the Voice UI (User Interface) of a new fridge, even though the latter was trained on a corpus containing the former. The line where the original rights cease to apply has to be somewhere between these points, and researchers and developers need to know where.

The authors present arguments that copyright and personal data restrictions covering language data usually do not affect language models.[1]

The article develops further the previous legal research conducted in the field of language technologies (see Eckart de Castilho et al. 2018; Kelli et al. 2016; Kamocki et al. 2019; Kelli et al. 2018a; Ilin and Kelli 2019; Klavan et al. 2018).

2 From language data to language technologies

The development of data-driven/data-based language technologies contains:

1. Collection of raw data (written texts, speech recordings, photos, videos, etc.). These often include copyrighted material and personal data. Their development usually does not involve any other activities than the actual recording, initial cleaning and sanity-checking of the data.

Dangers for both copyright and personal data implications can be real: re-publication of copyrighted works, infringement of privacy by governments or insurance companies, etc.

It is almost impossible to anonymise data entirely so that it would become impossible to identify any persons.

2. Compilation of datasets, or collections of data (raw text corpora like Google News, Common Crawl[2] or Open Subtitles[3], speech corpora like the Prague Database of Spoken Czech, etc.). The above, but collected and organised with a specific criterion in mind (e.g. speech recordings of a particular topic by residents of a specific region to capture the accent of the region). These datasets usually come in such quantities that any individual piece of data constitutes a negligible part of the whole, and could in principle be removed without affecting the usability of the dataset.

For personal data purposes, data collections are not different from raw data. The main practical difference is that the sheer volume of data may make it technically difficult for an individual to become aware that their data has been included in the dataset.

At the IPR side, the original rights of the individual pieces of data remain as is when included in the dataset. For instance, the copyright of a photo or speech recording is carried over so that the copyright of the dataset consists of the copyright of the individual items. IPR of the individual items must first be cleared to attach a single licence to a dataset. Also, the creation of a dataset often involves a nontrivial contribution in gathering, organising, indexing, presenting, hosting etc. of the data. This reflects on the *sui generis* database (SGDB) right, which governs the *structure* rather than the *contents* of the dataset[4].

[1] The analysis is limited to models containing speech and text.

[2] See also http://commoncrawl.org/

[3] see also https://www.opensubtitles.org/

[4] In fact, it can be argued that data-sets qualify for database protection (for further discussion, cf. Eckart de Castilho et al. 2018; Kelli et al. 2012).

3. Creation of annotated datasets (POS-tagged corpora of written texts like the web13 (etTenTen)[5], syntactically parsed corpora like the Universal Dependencies[6] treebanks, etc.). The above, augmented with some analysis.

Again, annotated datasets are not different from raw data in terms of copyright and personal data, although the copyright holders of the raw data and the annotations may be different. The annotation layers may be stored separately and may even have some use on their own. Still, standard practice is to include copies of the original data together with the annotation layers so that the resulting dataset contains all of the original data.

Creation of an annotated dataset includes analysis of the data, either manual, semi-automatic or automatic. When this analysis is performed manually, it can be argued that the copyright of the annotations belongs to the annotators (or the organisation that has commissioned the task). On the other hand, we can argue that a strictly automated annotation of a dataset does not create new rights either on the part of the person(s) that have run the annotation tool nor on the part of the person(s) that have developed the annotation tool.

4. Models. Data products developed from some processing on the above, but not necessarily containing the above, which try to model, i.e. represent or describe, language usage. Examples, in this broad sense, include dictionaries, wordlists, frequency distributions, n-gram lists like Google n-grams, pre-trained word embeddings (cf. Grave et al. 2018), pre-trained language models (cf. Devlin et al. 2018).

Creation of a model involves significant amounts of work, expertise and (computational) resources. Steps include at least creation and/or selection of the algorithm, implementation of the algorithm in software, hardware setup (may even include custom hardware development), hyperparameter optimisation, model validation.

Some model types may be consumer products of their own (e.g. dictionaries). Mainly, however, models are used in downstream tasks to create other products.

5. Semi-finished products (text-to-speech engine or a visual object detector) and finished products (talking fridge). Out of scope for the current analysis, because their independent status should be beyond doubt.

3 Copyright perspective on the creation of language models

From the copyright perspective, there are three relevant issues. Firstly, whether copyrighted material is used. Secondly, if it is used, whether there is a legitimate ground for this use. Thirdly, how to define models themselves within the copyright framework.

The requirements for copyright subject matter should be briefly outlined before explaining the copyright law impact on models. The primary and long-established requirement is that of originality. A work is protected if, and only if, it is original. Therefore, the originality requirement defines the copyright status of the input data. Oddly enough, this general requirement was never defined in international treaties or European *acquis*[7]. The task to define the legal meaning of originality for copyright purposes was mainly taken by the Court of Justice of the European Union (CJEU). As was explained in the seminal decision in the *Infopaq* case (C-5/08), originality means the author's intellectual creation. Another relevant explanation in the *Infopaq* case was that an extract consisting of eleven words could constitute an original work (C-5/08 para 48). The Court has also explained that a single word cannot be regarded as an original and protectable work.

In the context of the current paper, the originality requirement is important from two different perspectives. First, if originality is missing from the dataset used for the creation of the model, the pre-text contained in a dataset is not protected and can be used without authorisation. Therefore, even if parts of this text are reproduced in the model, they are not protected either. Second, even if a text as a whole is original and, therefore, protected, the question remains, whether the fragments used in the model are

[5] http://doi.org/10.15155/1-00-0000-0000-0000-0012EL

[6] See also https://universaldependencies.org/

[7] Although it was defined in several EU directives with regard to specific categories of works, such as computer programs or photographic works.

original on their own. If they are not, then again, they can be used without authorisation. Thus, originality must be established not only concerning the original work but also as regards the parts used.

To answer the question of whether models are copyright protected, we must establish whether they meet the requirement of originality also on their own (irrespective of the input dataset).

One of the criteria that can be used for assessing originality has to do with the degree of human intellectual effort invested in the process: how far is the model a unique product, the result of the intellectual creation of the author? Building a model (as presented in Section 2) includes several choices and actions on the part of the developer: choice/creation of the dataset, choice/creation of the programme to be used for the training and development of the model and various cycles of testing and validation by tuning the parameters of the training programme.

Text can be too short or trivial or limited in creative choices to qualify as an original work. Some models (like a simple frequency list) may also be too simple or too limited in options (cf. Eckart de Castilho et al. 2018). In nontrivial cases, the *de facto* situation is that models are made available together with the research papers describing them and the software tools used in their creation. Standard licenses applied to models by their creators include Creative Commons (CC) Attribution ShareAlike 4.0 International (e.g., Grave et al. 2018), Apache License 2.0 (Devlin et al. 2018; Yang et al. 2019) and Public Domain Dedication and License v1.0 (e.g., Pennington et al. 2014).

It is also crucial to answer the question about the copyright status of models. The problem is whether they can be considered "derivatives" or "adaptations" of the original (primary or underlying) work. There is no uniform definition of derivative work at EU and international levels. Different jurisdictions have their approaches (for further discussion, see Birštonas and Usonienė, 2013; Eckart de Castilho et al. 2018).

The Berne Convention does not name derivative works but refers to them. According to the convention "[t]ranslations, adaptations, arrangements of music and other alterations of a literary or artistic work shall be protected as original works without prejudice to the copyright in the original work" (Art. 2 (3)). The EU case law concerning derivative works makes a reference to the concept of substantial similarities (T‑19/07 para 259).

Some examples are provided below to describe national approaches. For instance, although the Estonian Copyright Act provides that derivative works are protected by copyright (§ 35 (1)), it does not define the derivative work. Instead, it says that "*translations, adaptations of original works, modifications (arrangements) and other alterations of works*" are considered derivative works (§ 4 (3) clause 21). The Estonian Copyright Act provides that the derivative work has to be "*derived from the work of another author*" (§ 35 (1)). The Act sets forth non-exhaustive examples of what constitutes the creation of derivative works: "*the transformation of a narrative work into a dramatic work or a script, the transformation of a dramatic work or a script into a narrative work, the transformation of a dramatic work into a script, and the transformation of a script into a dramatic work*" (§ 35 (2)).

The Finnish Copyright Act contains the following regulation "A person who translates or adapts a work or converts it into some other literary or artistic form shall have copyright in the work in the new form, but shall not have the right to control it in a manner which infringes the copyright in the original work" (Art. 4 (1)).

According to the Lithuanian Copyright Act, the subject matter of copyright also includes "derivative works created on the basis of other literary, scientific or artistic works (translations, dramatisations, adaptations, annotations, reviews, essays, musical arrangements, static and interactive Internet homepages, and other derivative works)" (Art. 4 (3) clause 1).

The Czech Copyright Act provides that "A work which is the outcome of the creative adaptation of another work, including its translation into another language, shall also be subject to copyright" (Art. 2 (4)).

Finally, the Greek Copyright Act under 'economic rights' (Art. 3) gives the rightholder of a work the right to authorize or prohibit what we usually term 'derivative works'. In fact, although the term 'derivative' is often used in the Greek related legal literature, it is not used in the law, which prefers to refer to "*the arrangement, adaptation or other alteration of their works*", using the same exact wording as in Article 12 of the Berne Convention. We should also mention here the clause "*the translation of their works*" which is usually deemed a derivative work. The Greek law does not further specify any criteria for assessing what an "*adaptation or alteration*" is, yet similarity to the original work and originality of the new work are typically used for this purpose (Marinos 2018).

Based on the referred copyright laws, it can be concluded that to qualify as a derivative work, it has to include substantial copyright-protected parts of the used primary work.

In addition to national laws, it is useful to look at standard licenses such as Creative Commons (CC) which are used as tools to make copyrighted content (language data) available. For instance, CC Attribution-NoDerivs (CC BY-ND) does not allow to share adapted materials (derivative works). According to CC BY-ND "*Adapted Material means material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor*". A key issue here is whether permission is required to create a model.

The authors consider that the creation of the model is done through a data mining activity. The Digital Copyright Directive (DCD) defines text and data mining (TDM) as "*any automated analytical technique aimed at analysing text and data in digital form to generate information which includes but is not limited to patterns, trends and correlations*" (Art. 2 (2)).

The Digital Copyright Directive has two mandatory TDM exceptions. One is meant for research and cultural heritage institutions (Art. 3) and the other for everyone (Art. 4). Since the focus of the current article is on the research context and due to limited space, the authors concentrate on TDM for research purposes.

According to the Digital Copyright Directive research organisations and cultural heritage institutions[8] are entitled to rely on this exception. The Directive defines research organisations extensively. The requirement is that research is conducted "*on a not-for-profit basis or by reinvesting all the profits in its scientific research; or pursuant to a public interest mission recognised by a Member State in such a way that the access to the results generated by such scientific research cannot be enjoyed on a preferential basis*" (Art. 2 (1)).

The Digital Copyright Directive Art. 3 (1) allows making copies of works[9], objects of related rights (e.g., performances), press publications[10] and extractions from *sui generis* databases for TDM for scientific research. The key issue here is that access to the material has to be lawful.

There are remedies in case rightholders adopt measures limiting the TDM exception. According to 7 (1) of the Digital Copyright Directive, any contractual provision contrary to the exception is unenforceable. The situation is more nuanced with technological measures.[11] The Digital Copyright Directive Art. 3 (3) allows rightholders "*to apply measures to ensure the security and integrity of the networks and databases where the works or other subject-matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective*". The question is, what happens if rightholders go beyond what is allowed by the Directive. According to the InfoSoc Directive Art. 6 (4), Member States shall "*take appropriate measures to ensure that rightholders make available to the beneficiary of an exception or limitation*". It should be mentioned that the practical application of this requirement is not so smooth. There are few efficient mechanisms to compel rightholders to adopt technological measures to allow the free use prescribed by law.

A key issue for language research relates to the use of compiled datasets exploited for TDM. The question is, what can be done with datasets. The Digital Copyright Directive Art. 3 (2) provides that "*Copies of works or other subject-matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results*". The Directive does not say clearly whether datasets can be shared among researchers. This is a genuinely crucial issue since research and research infrastructures such as CLARIN are based on the ideology of sharing research data. It remains to be seen how the national

[8] The Digital Copyright Directive defines cultural heritage organisations as "*a publicly accessible library or museum, an archive or a film or audio heritage institution*"(Art. 2 (3)).

[9] The quotation right also allows to copy parts of a work. However, according to the EU case law "the user of a protected work wishing to rely on the quotation exception must therefore have the intention of entering into 'dialogue' with that work" (C-476/17 para. 71). Since the development of language technology relies on works as language data, then there is no 'dialogue' and the quotation right is not applicable.

[10] The right to press publications is introduced with the Digital Copyright Directive Art. 15.

[11] The InfoSoc Directive Art. 6 (3) defines technological protection measures as "any technology, device or component that, in the normal course of its operation, is designed to prevent or restrict acts, in respect of works or other subject-matter, which are not authorised by the rightholder".

legislators implement the provision. The research community should use all possible measures to introduce a regulation which allows at least limited sharing.

The TDM exception is not limited to non-commercial activities. The Directive allows for public-private partnerships. This means that research organisations can collaborate with private partners to carry out the TDM (Recital 11 of DCD).

To say whether models constitute derivative works, we should classify and analyse all possible model types, the processes and resource types and modalities they have been built upon. It is not feasible within the limits of this article. It can be argued though that models by definition try to capture *generalities* of language use and *abstract* from the original texts as far as possible, producing mainly lists of words or phrases and patterns with statistical measures. Therefore, they cannot be usually considered derivative works. This conclusion is supported by other researchers as well (see Eckart de Castilho et al. 2018).

4 Database right perspective

Another set of rights which could encumber the initial material and (possibly) the models, is the *sui generis* database makers right (database right). This type of protection was introduced by the Database Directive of 1996 (DD) and mostly remains a peculiarity of European countries. The database right is different from copyright in several important respects. First of all, the subject matter of protection is only a database. A database is defined as "*a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means*" (DD Art. 1 (2)). Although this could seem like a technical definition with quite a narrow field of application, the said definition is very wide and encompasses such different subject matter as directories, dictionaries, newspapers, journals, the content of a website and so forth. It is obvious that raw data for the models often, if not most, come from databases.

If the subject matter is qualified as a database, then, according to the Database Directive, it can be protected by copyright or by the database right. Therefore, there are four possibilities: a database can be protected by copyright, by database right, by both of them or by none of them. Since copyright protection was discussed above, we do not address copyright-protected databases further. In all that was said before, copyright equally applies to a copyright-protected database.

The database right should be differentiated from copyright. The requirement for the database protection is not based on originality, but instead, the following three cumulative conditions should be met: 1) there should be an investment, 2) an investment should be "*qualitatively and/or quantitatively substantial"*, 3) an investment should be in the obtaining, verification or presentation of the contents (DD Art. 7 (1)).

The existing court practice (both from ECJ and national courts) shows that these formal requirements can be met quite easily. The concept of investment is interpreted broadly (it includes financial resources, time, energy, labour, time, and so forth) and the threshold of "*qualitatively and/or quantitatively substantial*" is not very high. As a result, a considerable number of datasets, used for language technology purposes, are covered by the database right. In such a case, the maker of a database has the right to prevent extraction and/or re-utilization of the whole or a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database (DD Art. 7 (1)). An act of extraction is defined as "*the permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form*", whereas 're-utilization' means "*any form of making available to the public all or a substantial part of the contents of a database by the distribution of copies, by renting, by on-line or other forms of transmission*" (DD Art. 7 (2)). It should be stressed, that contrary to copyright protection, the database right protects not the original selection or arrangement, but the content of a database itself (or, more precisely, the whole or a substantial part, evaluated qualitatively and/or quantitatively). According to the case law "*[t]he purpose of the protection by the sui generis right provided for by the directive is to promote the establishment of storage and processing systems for existing information and not the creation of materials capable of being collected subsequently in a database*" (C-203/02 para. 31).

Translating these legal rules to the field of language technologies, it could be noted that the database right could have a direct impact on language technology and its results. Many sources, from which data is taken for the compilation of datasets, are protected by the database right. Further, the same two questions, which are relevant in the context of copyright, arise: first, if there is a legitimate ground for the use of a database, and, second, how a model can be affected by the database right.

Answering the first question, there can be no doubt that normally a collection of raw data from a database (see chapter 2) constitutes an act of extraction. While the Database Directive restricts the scope of the said right to the instances, when the whole or substantial part of the content of a database is copied (Art. 7 (1)), this is exactly the case in a typical raw data collection process. The problem is solved by the data mining exception introduced by the Copyright Directive (Art. 3 and 4), which expressly provides exceptions not only to copyright but also to the database right. Therefore, while normally, the collection of data falls into the sphere of the extraction right, it can still be legitimate if the requirements of the exception of data mining are met. As was said, the collection of data and development of the model is considered data mining. Therefore, additional permission from the database maker is not necessary.

The remaining question concerns models and their status from the database right perspective. As it was seen, models rely on datasets. Arguably, in some cases the interference with the re-utilization right is possible. As was explained, re-utilization means any form of making available to the public all or a substantial part of the contents of a database by the distribution of copies, by renting, by on-line or other forms of transmission. It should be borne in mind, that ECJ confirmed several times that re-utilization should be interpreted is a wide manner (see C-203/02 para. 51; C-173/11 para 20). In one of its latest decisions, ECJ has reiterated that 're-utilization' refers to *any* unauthorised act of distribution to the public of the contents of a protected database or a substantial part of such contents, while the nature and form of the process used are of no relevance in this respect (C-202/12 para 37). It can theoretically be argued that in a model, a certain amount of data (e.g., discrete words) is transmitted, that is re-utilized. Still, the right of re-utilization is infringed only if the whole or a substantial part, evaluated qualitatively and/or quantitatively, of a database is used. Lawful users of the database have a specific right to re-utilize insubstantial parts of its contents for any purposes whatsoever.

In a model, the whole or a substantial part, evaluated quantitatively, of the content of a database is rarely re-created, but the same cannot be said for a substantial part, evaluated qualitatively. The substantial part evaluated qualitatively is a nebulous concept, which, according to the ECJ, refers to the scale of the investment in the obtaining, verification or presentation of the contents of the subject of the act of re-utilization, regardless of whether that subject represents a quantitatively substantial part of the general content of the protected database (C-203/02 para. 71). Also, a quantitatively negligible part of the content of a database may represent, in terms of obtaining, verification or presentation, a significant human, technical or financial investment (C-203/02 para. 71). So, in principle, even small excerpts of the original data can represent a qualitatively substantial part of the content of the protected database. Furthermore, the repeated and systematic re-utilization of insubstantial parts of the content of the database implying acts which conflict with a normal exploitation of that database or which unreasonably prejudice the legitimate interests of the maker of the database are not permitted (DD Art. 7(5)). Again, this can potentially pose a problem for language technologies, because it can be argued that in models some insubstantial parts of the database are re-utilized repeatedly and systematically.

In conclusion, the database right causes certain uncertainties and risks when it comes to the creation of models.

5 Personal data protection perspective on the creation of language models used for commercial purposes

Since several issues relating to personal data in language research have been previously addressed (see Kelli et al. 2019; Lindén et al. 2019; Kelli et al. 2018b) then only personal data aspects relevant for models are addressed in this article.

The General Data Protection Regulation (GDPR) defines personal data as "*any information relating to an identified or identifiable natural person*" (Art. 4 (1)).

It is possible that models contain personal data such as a name or an e-mail address. As a result, GDPR becomes applicable. To avoid legal restrictions stemming from the General Data Protection Regulation, it is advisable to anonymise data (for further discussion, see WP29 2014). The GDPR does not apply to anonymous information (Recital 26 of the GDPR).

However, it should be kept in mind that for personal data, there is no minimum segment in the audio synthesis. Even if the voice is synthesised using neural networks without any remnants of the person's original voice recording, having trained the network for research purposes using a publicly available radio transmission as training data, one is still using the personal data of that person when the person can be identified based on the synthesised output although there is no single bit in the network which could be attributed to the person's voice.

The main issue here is how to substantiate the processing[12] of personal data contained in a model. Generally speaking, the compilation of datasets containing personal data used to create models can be based on the consent, public interest research and legitimate interest (see, GDPR Art. 6 (1) a), e), f)). In case there is consent to process data for research purposes, or processing relies on public interest and the resulting model is used for research purposes as well, then there is no problem. There is also no problem if consent covers commercial use and public dissemination.

However, the situation becomes complicated when a dataset containing personal data is processed based on consent asked for research or on the public interest research exception, but the resulting model (where the personal data remains) is planned to be used for commercial purposes or be made publicly available. If the personal data is in the form of speech, then anonymization is rather difficult. In the described case, there are the following scenarios:

1. Argue that voice without any identifying information is not personal data (it is anonymous data). The key here is how to interpret the concept of an identifiable natural person (see Art. 4 (1); WP29 2007);
2. Ask for consent for commercial use (see WP29 2018);
3. Argue that the use of voice in the model is based on the legitimate interest (for further discussion, see WP29 2014a). Especially bearing in mind that the identification is impossible or almost impossible and the voice does not contain any data which would affect the data subject negatively;
4. Technically modulate the voice data so that it no longer resembles the original speaker without destroying the properties of the speech signal essential for the intended application.

The first and third options are somewhat uncertain and pose legal risks.

6 The commercialisation of language models

6.1 The general framework for commercialisation

The goal of commercialisation is to derive profit from adding value to some raw material. The raw material can be any input that is further processed before it is commercialised. A minimal act of further processing is copying. In this work, we explore the conditions for commercialising a language model. In the following, we will compare some practices in different countries for further processing language data to produce language models for commercialisation.

6.2 National approaches to commercial exploitation of language models

Czechia. At LINDAT/CLARIAH-CZ[13], hosted at Charles University, Prague, the choice made was rather cautious. If a source of data contains a non-commercial clause (such as –NC in CC licences), it is considered binding for any derivative work and also for models which are not deemed copyrightable. This is equally applied to data and models Charles University produces itself. That allows us to keep releasing our annotated corpora under CC licenses and yet try to protect their commercial use. That is

[12] The GDPR defines processing as "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction" (Art. 4 (2)).

[13] https://lindat.cz

why we release all UD models (parsing and morphology from >100 treebanks) under -NC, regardless of the original CC clauses attached to the dataset CC licence, and at the same time we cooperate with a commercial NLP company on licensing for the commercial sector under the following procedure:

(i) We say that for the models, we charge only for the convenience of the user, so that s/he does not have to go through the training process, usually just some small money (any commercial user can train the model themselves if they possess the expertise). However, this makes commercial use legal only for models from corpora that are free of further restrictions (e.g. CC-BY and similar).

(ii) For legal use of the models trained from datasets with restrictions (CC-BY-NC(-SA), but also one GPL, etc., see https://lindat.mff.cuni.cz/repository/xmlui/page/licence-UD-2.5#) the commercial NLP company contacts the dataset creators one by one and tries to make a deal for the commercial use of the dataset. Based on these deals they offer to their commercial client to clear the models for these datasets by getting them commercial licenses for the underlying datasets. That gets the clients into the situation when they have legal access to the datasets for commercial use, and they have paid us the fee for the convenience of training the models for them.

LINDAT/CLARIAH-CZ itself is one of those dataset creators, so if the commercial use case involves Czech, we get something for training the models and something for the commercial license of the annotated dataset(s).

Estonia. Ideally, all the groundwork from raw data collection to annotated datasets as described in section 2 (steps 1-3) would be done by researchers who have the benefits of using data for research purposes. The result of the first steps would remain a derivative or original work (including *sui generis* databases), strictly managed by the research organisation, subject to copyright and subject to conditions of use determined by the research organisation in accordance with the input data rights.

The model building phase (step 4) could also involve collaboration between researchers and companies, where companies outsource the expertise of researchers in data processing and researchers process or determine, the appropriate parameters in the (automatic) model creation process that meet the business model development goals.

Cases where researchers work (in part, also in their start-up) in a business enterprise or a start-up that has grown out of a research institution are more complicated.

The commercial use of language resources, including the creation of models, should, in any case, remain a tailor-made activity and the value-added for business purposes should be left to the researchers. In doing so, researchers can ensure that the model does not include raw data that conflicts with the conditions of use of the first steps (1-3).

For example, there are three different morphosyntactic analysis models available for the development of language technology in Estonia that can be used (and have been used) under the CC-BY-SA license. Researchers of the University of Tartu create these models by training on a variety of text materials with different legal regimes, but in the resulting model, none of the text used as reference data is in a recognisable form. The same is true with the statistical machine translation model.

Finland. As soon as an organisation has lawful access to a dataset, it can create a model based on the dataset. Provided that the model does not contain substantial reproducible parts of the original copyright-protected data, the model is an independent work which can be given whatever license its creator chooses. However, if a dataset has additional restrictions, e.g. non-commercial use, this means that the organisation cannot engage in such activities for producing models with the dataset unless the restriction is lifted. This is similar to the situation described for Czechia.

Greece. To the best of our knowledge, there are very few models for modern Greek publicly available[14], and these are distributed under a CC-BY-NC-SA licence, most probably because they have been trained on the relevant UD treebank licensed with the same terms. It also seems that there are not yet any discussions on the commercial use of models. As regards language resources of modern Greek, the interest lies mainly on annotated datasets and relevant tools and services.

[14] https://ufal.mff.cuni.cz/udpipe/models and https://spacy.io/models/el

Lithuania. There is no reliable data concerning the commercial exploitation of language models in Lithuania. Models are mainly prepared by educational and scientific institutions, and they tend to make models either publicly open[15], or make them available on request. It is important to note that models are shared only for non-commercial purposes. Requests for commercial use are declined. If models are shared, the specific public licence is applied. It has to be noted, that language resources used in the framework of the CLARIN-LT, are not licensed by one of the Creative Commons licences, but the specific public licence was prepared by Lithuanian team of lawyers.

7 Conclusion

The creation of language models relies on the use of language data. Language data could contain copyright-protected works, objects of related rights (performances, recordings, databases) and personal data. Therefore, its use is restricted by copyright and personal data protection laws. The process of creating a model can involve text and data mining (TDM). From the copyright perspective, TDM in itself is not an activity requiring a legal basis (consent or exception). However, to conduct TDM, there is a need to copy language data (copyright-protected works, objects of related rights) which must have a legal basis. The Digital Copyright Directive introduced a mandatory exception for TDM, which allows making reproductions for TDM.

The creation of a language model involves several complex human intellectual activities, such as choosing and annotating datasets as well as choosing the software and tweaking its parameters. The outcome of the preparatory software activities is applied to a prepared dataset to compile a language model.

If the created language model contains copyrighted language data used to develop the model, then the model is subject to the same copyright restrictions as the data. However, this depends on the model type. In most cases, models do not contain copyright-protected content.

From the perspective of database right, the use of models in certain cases theoretically could infringe re-utilization right, although the legal practice has not settled this issue yet.

There is also the question of whether they contain enough material to breach personal data regulations. For instance, models containing speech need to address personal data issues.

The authors' main conclusion is that language models usually do not have the same legal restrictions as language data used to create them.

The authors' key findings can be visualised with the following graph:

[15] Some examples can be found here: http://mwe.lt/, https://www.semantika.lt/, etc.

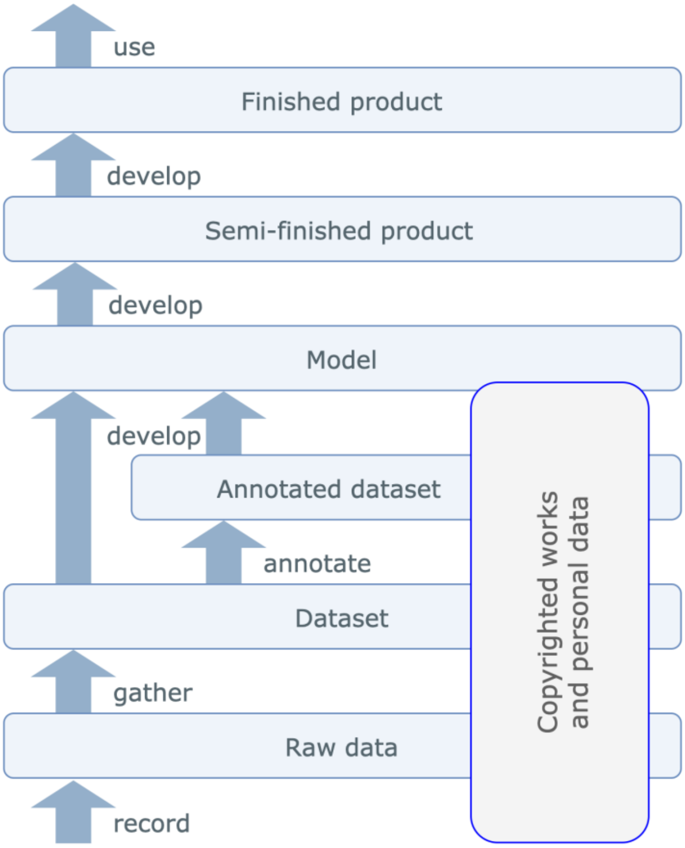


Figure 1: Process of developing language technology

Acknowledgements

Research is supported by the ERDF project "Federated Content Search for the Center of Estonian Language Resources (2014-2020.4.01.16-0134) under the activity "Support for Research Infrastructures of National Importance, Roadmap" and by the Research and Development Program "Estonian Language Technology 2018-2027" of the Ministry of Education and Research.

References

[Birštonas and Usonienė 2013] Ramunas Birstonas, Jurate Usoniene. 2013. Derivative Works: Some Comparative Remarks from the European Copyright Law. *UWM Law Review*, Volume 5.

[Berne Convention] Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979). Available at https://wipolex.wipo.int/en/text/283698 (2.2.2020).

[CC BY-ND] Creative Commons Attribution-NoDerivatives 4.0 International Public License. Available at https://creativecommons.org/licenses/by-nd/4.0/legalcode (2.2.2020).

[Czech Copyright Act] Copyright Act (121/2000). Available at https://www.wipo.int/edocs/lexdocs/laws/en/cz/cz043en.pdf (2.2.2020).

[C‑476/17] C‑476/17. Pelham GmbH, Moses Pelham, Martin Haas *vs* Ralf Hütter, Florian Schneider‑Esleben (29 July 2019). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1576587562212&uri=CELEX%3A62017CJ0476 (2.2.2020).

[C-5/08] Case C-5/08. Infopaq International A/S *vs* Danske Dagblades Forening (16 July 2009). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555243488182&uri=CELEX:62008CJ0005 (14.4.2019).

[C-203/02] Case C-203/02. The British Horseracing Board *vs* William Hill Organization (9 November 2004). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1580123225598&uri=CELEX:62002CJ0203 (27.1.2020).

[C-173/11] Case C-173/11. Football Dataco *vs* Sportradar (18 October 2012). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1580123660404&uri=CELEX:62011CJ0173 (27.1.2020).

[C-202/12] Case C-202/12. Innoweb *vs* Wegener ICT Media (19 December 2013). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1580124636913&uri=CELEX:62012CJ0202 (27.1.2020).

[Database Directive = DD] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. OJ L 77/20, 27.3.1996, pp. 20–28. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1580123910536&uri=CELEX:31996L0009 (27.1.2020).

[Digital Copyright Directive = DCD] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. OJ L 130, 17.5.2019, pp. 92-125. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1572352552633&uri=CELEX:32019L0790 (26.1.2020).

[Devlin et al. 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [Cs].

[Eckart de Castilho et al. 2018] Eckart de Castilho, R., Dore, G., Margoni, T., Labropoulou, P. & Gurevych, I. 2018. A legal perspective on training models for Natural Language Processing. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, ELRA. Available at http://www.lrec-conf.org/proceedings/lrec2018/pdf/1006.pdf (17.4.2019).

[Estonian Copyright Act] Copyright Act (12.12.1992). Available at https://www.riigiteataja.ee/en/eli/504042019001/consolide (2.2.2020).

[Finnish Copyright Act] Copyright Act (404/1961). Available at https://www.finlex.fi/en/laki/kaannokset/1961/en19610404 (2.2.2020).

[Grave et al. 2018] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, Tomas Mikolov. 2018. Learning word vectors for 157 languages. ArXiv Preprint ArXiv:1802.06893.

[GDPR] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1-88. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555312258399&uri=CELEX:32016R0679 (2.2.2020).

[Greek Copyright Law] Greek Copyright, Related Rights and Cultural Matters (Law 2121/1993 amended by Law 4531/2018). Available at https://www.opi.gr/en/library/law-2121-1993 (5.2.2020).

[Ilin and Kelli 2019] The Use of Human Voice and Speech in Language Technologies: The EU and Russian Intellectual Property Law Perspectives. Juridica International 28, 17-27. Available at https://www.juridicainternational.eu/public/pdf/ji_2019_1_17.pdf (2.2.2020).

[InfoSoc Directive] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. *Official Journal L 167*, 22/06/2001 P. 0010 – 0019. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555254956114&uri=CELEX:32001L0029 (14.4.2019).

[Kamocki et al. 2019] Pawel Kamocki, Erik Ketzan, Julia Wildgans, Andreas Witt. 2019. New exceptions for Text and Data Mining and their possible impact on the CLARIN. In: Inguna Skadina, Maria Eskevich (Ed.). Selected papers from the CLARIN Annual Conference 2018. Linköping University Electronic Press, 66-71. Available at http://www.ep.liu.se/ecp/article.asp?issue=159&article=007&volume= (2.2.2020).

[Kelli et al. 2019] Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramunas Birštonas, Silvia Calamai, Penny Labrpolou, Maria Gavrilidou, Pavel Straňák. 2019. Processing personal data without the consent of the data subject for the development and use of language resources. In: Inguna Skadina, Maria Eskevich (Ed.). Selected papers from the CLARIN Annual Conference 2018. Linköping University Electronic Press, 72-82. Available at http://www.ep.liu.se/ecp/159/008/ecp18159008.pdf (29.1.2020).

[Kelli et al. 2018a] Aleksei Kelli, Tõnis Mets, Lars Jonsson, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Age Värv. 2018. Challenges of Transformation of Research Data into Open Data: the Perspective of Social Sciences and Humanities. International Journal of Technology Management & Sustainable Development, 17 (3), 227-251.

[Kelli et al. 2018b] Aleksei Kelli, Krister Lindén, Kadri Vider, Penny Labropoulou, Erik Ketzan, Pawel Kamocki, Pavel Straňák. 2018. Implementation of an Open Science Policy in the context of management of CLARIN language resources: a need for changes? In: Maciej Piasecki (Ed.). Selected papers from the CLARIN Annual Conference 2017. Linköping University Electronic Press, 102-111. Available at http://www.ep.liu.se/ecp/147/009/ecp17147009.pdf (29.1.2020).

[Kelli et al. 2016] Aleksei Kelli, Kadri Vider, Krister Lindén. 2016. The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. In: Koenraad De Smedt (Ed.). Selected Papers from the CLARIN Annual Conference 2015. Linköping University Electronic Press, 13-24. Available at http://www.ep.liu.se/ecp/article.asp?issue=123&article=002 (29.1.2020).

[Kelli et al. 2012] Aleksei Kelli, Arvi Tavast, Heiki Pisuke. 2012. Copyright and Constitutional Aspects of Digital Language Resources: The Estonian Approach. Juridica International, XIX, 40-48. Available at https://www.juridicainternational.eu/public/pdf/ji_2012_1_40.pdf (3.2.2020).

[Klavan et al. 2018] Jane Klavan, Arvi Tavast, Aleksei Kelli. 2018. The Legal Aspects of Using Data from Linguistic Experiments for Creating Language Resources. Frontiers in Artificial Intelligence and Applications, 307, 71-78. Available at http://ebooks.iospress.nl/volumearticle/50306 (29.1.2020).

[Lindén et al. 2019] Krister Lindén, Aleksei Kelli, Alexandros Nousias. 2019. To Ask or not to Ask: Informed Consent to Participate and Using Data in the Public Interest. Proceedings of CLARIN Annual Conference 2019: CLARIN Annual Conference, Leipzig, Germany, 30 September – 2 October 2019. Ed. K. Simov and M. Eskevich. CLARIN, 56-60. Available at https://office.clarin.eu/v/CE-2019-1512_CLARIN2019_ConferenceProceedings.pdf (3.2.2020).

[Lithuanian Copyright Act] Law on copyright and related rights (18.5.1999). Available at https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/5f13b560b2b511e59010bea026bdb259 (2.2.2020).

[Marinos 2018] Michael-Theodore Marinos. 2018. The infringement of a copyrighted work with its "al-teration" – the delimitation between free and prohibited usage. Greek Law 1/2018 (59), p. 1-10. (in Greek). Available at https://mklpartners.gr/wp-content/uploads/meletes/metatropi_ergou.pdf (5.2.2020).

[Pennington et al. 2014] Jeffrey Pennington, Richard Socher, Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. Available: https://nlp.stanford.edu/projects/glove/ (3.11.2019).

[T‑19/07] Case T‑19/07. Systran SA, Systran Luxembourg SA *vs* European Commission (16 December 2010). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1580669711496&uri=CELEX:62007TJ0019 (2.2.2020).

[Yang et al. 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le. 2019. XLNet: Gen-eralized Autoregressive Pretraining for Language Understanding. ArXiv:1906.08237 [Cs]. Avaialble at: http://arxiv.org/abs/1906.08237 (3.11.2019).

[WP29 2018] Article 29 Working Party (WP29). Guidelines on consent under Regulation 2016/679. Adopted on 28 November 2017. As last Revised and Adopted on 10 April 2018. Available at https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051 (3.2.2020).

[WP29 2014] Article 29 Working Party (WP29). Opinion 05/2014 on Anonymisation Techniques. Available at https://iapp.org/media/pdf/resource_center/wp216_Anonymisation-Techniques_04-2014.pdf (3.2.2020).

[WP29 2014a] Article 29 Working Party (WP29). Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC. Available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf (3.2.2020).

[WP29 2007] Article 29 Working Party (WP29). Opinion 4/2007 on the concept of personal data. Adopted on 20th June. Available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf (2.2.2020).

Technical Solutions for Reproducible Research

Alexander König
Eurac Research, Italy /
CLARIN ERIC, the Netherlands
alex@clarin.eu

Egon W. Stemle
Eurac Research, Italy
Egon.Stemle@eurac.edu

André Moreira
CLARIN ERIC, the Netherlands
andre@clarin.eu

Willem Elbers
CLARIN ERIC, the Netherlands
willem@clarin.eu

Abstract

In recent years, the reproducibility of scientific research has increasingly come into focus, both by external stakeholders (e.g. funders) and by the research communities themselves. Corpus linguistics, with its methods for creating, processing and analysing corpora, is an integral part of many other disciplines that work with language data and therefore plays a special role. Moreover, language corpora are often living objects that are regularly improved and revised. At the same time, tools for the automatic processing of human language are also being developed further, which can lead to different results with the same processing steps and the same data. This article argues that modern software technologies, such as version control and containerisation, can mitigate the following problems: Software packaging, installation and execution and, equally important, the tracking of corpus modifications throughout its life-cycle. All in all, this leads to transparency of changes to raw data and software tools and thereby enhanced reproducibility.

1 Introduction

While reproducibility has always been one of the main pillars of scientific research, within the last ten years, this has come even more into focus for the social sciences and humanities. Prominent cases of scientific fraud, for example, the case of Diederik Stapel in the Netherlands (Levelt et al., 2012), have brought problems about the reproducibility of scientific research into focus. In this article, we discuss possible techniques to handle this problem by using standard tools from the realm of software development. We propose to use versioning software to ensure the persistence of data (see section 2) and containerisation to ensure the same for natural language processing (NLP) tool-chains (see section 3), both concepts are illustrated with case studies (see section 2.2 and section 3.2) and finally, we highlight some challenges we encountered along the way (see section 4).

2 Ensuring persistence of data

2.1 Introduction

After the initial data collection, a corpus usually keeps evolving while initial analyses are already being carried out. It is also likely that while working on the corpus and analysing the data, errors in the transcriptions or annotations are discovered, which need to be corrected. And with a rich annotation scheme that is constantly being re-evaluated and refined this is usually all the more true. Likewise, there are usually changes to the tools and the processing pipelines used for the processing of the corpus and involved in its analysis. More generally, the different research strategies of the method-oriented computational linguistics, and the hermeneutic tradition of the social sciences and humanities (SSH) may contribute to systemic iterations of the tools and their processing pipelines, and the corpus data and its assessment strategies (Kuhn, 2019).

While these kind of changes are unproblematic as long as the corpus is still in its ”building phase”, as soon as the first analyses have been made public, any change to the data will endanger the possibility of

Alexander König, Egon W. Stemle, André Moreira and Willem Elbers 2020. Technical Solutions for Reproducible Research. *Selected papers from the CLARIN Annual Conference 2019.* Linköping Electronic Conference Proceedings 172: 172 66–74.

reproducing these analyses. Therefore, the researchers have to preserve a version of the corpus exactly as it was when a specific analysis was made, and a snapshot of all tools that were used to process the corpus and were involved in its analyses must be available in order to ensure *reproducibility*.

Following Cohen et al. (2018), we assume that such availability of tools and data will help to achieve two of the three dimensions of *reproducibility* they mention. They introduce 'Three Dimensions of Reproducibility in Natural Language Processing': 1. 'Reproducibility of a conclusion', 2. 'Reproducibility of a finding', and 3. 'Reproducibility of a value'. By *conclusion* they mean "a broad induction that is made based on the results of the reported research, by *finding* they mean "a relationship between the values for some reported figure of merit with respect to two or more dependent variables", and by *value* they mean "a number, whether measured (e.g. a count of false positives) or calculated (e.g. a standard deviation)". Broadly speaking, availability of the original data gives access to the core values (numbers) of corpora, and access to the state of tools enables the processing and comparing of numbers to discover the same relations between the numbers. The more abstract conclusions are not affected.

Using version control systems (VCSs) to provide "a lightweight yet robust framework that is ideal for managing the full suite of research outputs such as datasets, statistical code, figures, lab notes, and manuscripts" has already been suggested by Ram (2013) for bio medicine; Stodden et al. (2018) (first printed edition from 2014) have a very comprehensive overview from computational science; Nüst et al. (2018) give an overview of reproducibility in the field of geographic information science (GIS) and are careful to include information on their paper's reproducibility, for which they publish data, code, and a description of the runtime environment[1]; and Brinckman et al. (2019) introduce the Whole Tale project that connects "computational, data-intensive research efforts with the larger research process – transforming the knowledge discovery and dissemination process into one where data products are united with research articles to create *living publications* or tales" with application in material science, astronomy, and archaeology. Within the SSHs, for example, the Center for Reflected Text Analytics (CRETA)[2] "focuses on the development of technical tools and a general workflow methodology for text analysis within Digital Humanities. Of particular importance is the transparency of tools and traceability of results, such that they can be employed in a critically-reflected way." For NLP, projects like LaMachine[3] bundle numerous open-source NLP tools, programming libraries, web-services, and web-applications in a single Virtual Research Environment with the possibility to install explicitly defined versions for all software to support scientific reproducibility. In the following, we will make our proposal to further stimulate discussion in the community – inspired by the CLARIN infrastructure, but keeping in mind a general validity beyond it.

If the corpus in question is a text corpus (probably being stored in some kind of XML format like e.g. TEI[4]), an obvious solution to these problems is to use existing VCSs like subversion[5] or git[6] to keep track of all changes within the corpus.

This is also possible for corpora that are not mainly text-based, for example, multimodal corpora. First, they often have a text-component in their annotations (which could have been done, for example in ELAN[7] or EXMARaLDa[8], both of which store their data in XML format) and it is usually this part that will change the most within the lifetime of such a corpus, while the primary audio/video data often remains untouched. Secondly, although the problem of storing large files in versioning tools is grave, there will likely be solutions in the foreseeable future. Using such an existing versioning software solution, all changes throughout the life cycle of a corpus can be tracked, and through the use of code hosting platforms like Github[9] or GitLab[10] all changes can be made visually appealing to the research community.

[1] https://doi.org/10.5281/zenodo.1227260

[2] https://www.creta.uni-stuttgart.de/en/

[3] https://proycon.github.io/LaMachine/

[4] https://tei-c.org/

[5] https://subversion.apache.org/

[6] https://git-scm.com/

[7] https://tla.mpi.nl/tools/tla-tools/elan/

[8] https://exmaralda.org/en/

[9] https://github.com/

[10] https://www.gitlab.com

Having the corpus available on an online repository platform, while having the advantage of being very transparent about all changes made to the data, might not be the ideal way of providing the data to other researchers. Therefore traditional data repositories like CLARIN Centres, META-SHARE[11], and zenodo[12] will still play a role in making the data available to the users and especially in providing findability (through participation in search interfaces like the VLO[13] or the OLAC catalogue[14]) and issuing persistent identifiers to specific versions.

2.2 Case Study: The DiDi Corpus

The Institute for Applied Linguistics (IAL) at Eurac Research is currently investigating how it can move towards such a setup for more reproducibility in research as outlined in the previous section. One of the first corpora that was transformed into such a strictly versioned environment is the DiDi corpus (Frey et al., 2016). The corpus is available under an academic non-commercial (ACA-NC) license from an on-premise GitLab installation[15]. The whole corpus data is divided into multiple parts for the different file formats in which the corpus is available and is accompanied by extensive documentation. The different versions of the corpus are realised as tags in GitLab, which are references to the full state of the data repository at different times during its development. Additionally, these tagged versions are also uploaded into the Eurac Research CLARIN Centre (ERCC), the CLARIN DSpace repository hosted by the IAL, so they can be easily downloaded by less tech-savvy users[16]. Another advantage is, of course, that this integration of the data into a CLARIN Centre will make the metadata available to various search engines (e.g. the VLO or the OLAC search) and it can therefore be discovered more easily. All the data for a tagged version is available both at the ERCC and on GitLab with each of these hosting platforms cross-referencing the other. At both places, all versions are accompanied by a changelog that explains the changes between versions. On GitLab, the interested user can also make use of the integrated version diff to get more fine-grained information on the changes between versions.

However, in the attempt to implement the paradigm for explicit versioning of corpora, the DiDi corpus has a feature that requires special attention: the corpus contains personal information for which the corpus creators have asked the users for their consent to share the data, and this consent was explicitly requested for re-use in academic contexts.

More generally, linguistic corpora often consist of personal data produced by individuals where both privacy and IPR concerns need to be considered. And if not all of the data can be made publicly available, there has to be additional access protection both on the side of the DSpace repository and on the side of GitLab. While it is easy to have some data require a login with an academic account (for example, by using the CLARIN federated login[17]) in DSpace, the GitLab repository should ideally not be made completely password protected, but have at least an openly available landing page that describes the corpus. At the IAL this has been implemented for the DiDi corpus using git submodules where the main repository with the documentation and the overview of the various data formats is publicly accessible and the actual data is in sub repositories that require a login[18]. Still, all license information and documentation is available *without* login (see Figure 1). It is likely that more complex access scenarios will prove even more difficult to map to a code hosting platform.

3 Methods and tools and their impact on reproducibility

3.1 Introduction

In linguistic research – especially in the sub-fields of corpus linguistics and natural language processing – data is often processed with the use of quite intricate software tool-chains. Ranging from more simple

[11] `http://www.meta-share.org`

[12] `https://zenodo.org`

[13] `https://vlo.clarin.eu/`

[14] `http://search.language-archives.org`

[15] `https://gitlab.inf.unibz.it/commul/didi`

[16] `https://hdl.handle.net/20.500.12124/7`

[17] `https://www.clarin.eu/content/federated-identity`

[18] `https://gitlab.inf.unibz.it/commul/didi/data-bundle`

```
data-bundle/
├── CHANGELOG.html
├── CHANGELOG.md
├── data-annis/    @ecfce535
│   (git submodule with restricted access)
├── data-didijson/ @e9077a17
│   (git submodule with restricted access)
├── data-didixml/  @f7eb581d
│   (git submodule with restricted access)
├── data-docs/     @47299aef
│   ├── CHANGELOG.md
│   ├── DiDi_annotation_layers_DE.pdf
│   ├── DiDi_annotation_layers_EN.pdf
│   ├── DiDi_anonymisation_DE.pdf
│   ├── DiDi_anonymisation_EN.pdf
│   ├── DiDi_cmc_annotations_DE.pdf
│   ├── DiDi_cmc_annotations_EN.pdf
│   ├── DiDi_metadata_DE.pdf
│   ├── DiDi_metadata_EN.pdf
│   ├── LICENSE
│   └── README.md
├── EULA-CLARIN-ACA-BY-NC-NORED.md
├── EULA-CLARIN-ACA-BY-NC-NORED.pdf
├── EULA-CLARIN-ACA-BY-NC-NORED.txt
├── LICENSE
├── Makefile
├── README_gen.sh
├── README.html
└── README.md
```

Figure 1: Directory structure of the DiDi repository on GitLab

tasks like tokenisation or lemmatisation to more complicated ones like fine-grained syntactic parsing. The unification of all the necessary tools for the automatic processing of human language into an integrated processing framework is more the exception than the norm. This inevitably leads to a wide variety of individual solutions each with their own installation procedures, development life cycles with maintenance and update schedules, etc. (Wieling et al., 2018). Furthermore, linguistic models that are often at the heart of such tools are also subject to change, and this change need not necessarily be synchronised with the tool itself, spanning an even wider range of possible combinations (see e.g. (Nothman et al., 2018)).

This short overview already shows the difficulty for other researchers to exactly recreate a certain toolchain to verify research results. It can only be ensured if the original researchers document their setup carefully, noting down exactly which version of a certain tool was used and how exactly the various tools were combined. An additional problem is that some software manufacturers do not make older versions of their products easily available, so even if the version is known it is not certain that it can be obtained when necessary. For this reason it has been discussed whether scientific software should be archived in research repositories alongside the data, but so far, while some CLARIN repositories do also host linguistic tools, little progress has been made in this regard.

The recent trend in software deployment and administration towards containerisation of services seems to us to be a promising solution to the aforementioned problems regarding the reproducibility of data processing in linguistic research.

Containerisation means that programs are not installed on a real computer or in a full-blown virtualised environment like a virtual machine, but instead in a very reduced environment that leaves out everything that is not vital for the program in question to work. The idea is to minimise both the amount of memory and processor time needed for such a containerised service and especially the possibility for unwanted side effects. With Docker[19] this way of packaging programs and services has been widely adopted within the last years and the additional possibility to orchestrate the deployment of such minimal containers using a platform like Kubernetes[20] makes using existing containers ”off the shelf” quite easy, especially because a lot of the big infrastructure providers (e.g. Google[21], Microsoft[22] or Amazon[23]) offer ways to deploy containers on their infrastructure for a moderate price and there seems to be a trend to make this kind of deployment as easy as possible[24].

As a researcher, building the tool-chain for a new project can be done directly in Docker. There are already a variety of places where the resulting docker images (from which various container instances can be created) can be stored for re-use by others. For example, GitLab offers such a Docker image registry. GitLab is also a place where the data can be stored (see section 2), and both the data and the tool-chain used could thus be stored in one place, making it much easier for researchers planning to recreate an experiment to get both in exactly the same versions that had been used originally. The wide availability of container hosting (see above) also means that it will be quite easy to simply take such a container with the whole tool-chain setup and use it to verify the results or look for something else in another set of data while ensuring that the same methodology is used as in the original research.

3.2 Case Study: Gitlab and Docker - A framework for reproducible research

CLARIN ERIC has been working on a technical solution to manage their infrastructure based on docker images and a set of scripts to build, test and run these images (Elbers et al., 2019). It turns out this approach is also quite useful to tackle a number of the issues with respect to reproducibility of research. The core concepts of the approach are (1) manage the container image definition (Dockerfile when using docker) in a git repository, (2) use tags and a continuous integration (CI) solution to build and publish the resulting container image, preferably into a publicly accessible container registry. These initial requirements have since been extended, to accommodate the reproducibility of research, with (3) a clear container image template as shown in listing 1, (4) a clearly specified input format which should be verified before running the analysis and (5) a predefined command and set of arguments to run the image as a container and collect its output as shown in listing 2. Altogether this ensures a workflow where the same set of tools can be run, in an environment that is guaranteed to be stable, on different inputs (as long as the input specification is followed) to produce comparable results. This has been documented on GitLab [25].

```
#Base this image on Alpine Linux to ensure a small footprint to
#start with.
FROM alpine:3.11

#Install or compile any required packages and/or scripts here

#Define input directory, this will be populated with data via a host
#mount (see listing 2 Docker run command)
VOLUME "/input"
#Define output directories, this is where the /run.sh should produce
#all results and this data is exported to the host system via a host
#mount (see listing 2 Docker run command)
VOLUME "/output/datasets"
VOLUME "/output/tables_and_plots"

#Add and set an executable entrypoint script. This script _must_ be
#implemented by the researcher, can call other scripts and is
```

[19] https://www.docker.com/

[20] https://kubernetes.io/

[21] https://cloud.google.com/kubernetes-engine/

[22] https://azure.microsoft.com/en-us/services/kubernetes-service/

[23] https://aws.amazon.com/containers/

[24] https://cloud.google.com/blog/products/serverless/introducing-cloud-run-button-click-to-deploy-your-git-repos-to-google-cloud

[25] https://gitlab.com/CLARIN-ERIC/reprolang

```
#responsible for:
#1. Ensure input data is available in /input
#2. (Optional) Verify input
#3. Process all input data so it is ready for analysis (e.g. extract
#   tarball)
#4. Run analysis on the input data
#5. Produce final result in /output/datasets and
#   /output/tables_and_plots
ADD run.sh /run.sh
RUN chmod u+x /run.sh
ENTRYPOINT /run.sh
```

Listing 1: Dockerfile template example

```
 1  docker run \
 2      -ti --rm --name=${PROJECT_NAME} \
 3      -v ${PWD}/input:/input \
 4      -v ${PWD}/output/datasets:/output/datasets \
 5      -v ${PWD}/output/tables_and_plots:/output/tables_and_plots \
 6      <image name>:<image tag>
 7
 8  #Explanation:
 9  #Line 1 instructs docker to start a container
10  #Line 2 sets an interactive TTY (-ti), will remove the container
11  #       when done (--rm) and give the container a name (--name)
12  #Line 3 configures what directory on the host to mount into /input
13  #       inside the container.
14  #Line 4 and 5 configure what directories on the host to mount into
15  #       /output/datasets and /output/tables_and_plots
16  #Line 6 specifies which image and which tag of that image to start
```

Listing 2: Docker run command

Based on the experience in using this approach to run the CLARIN ERIC production environment for more than two years, the following best practices are advised: (1) always use version pinning for any packages installed into the container image. This will ensure the image build will break if there is an issue with any of the packages, providing a clear signal something will be different in the environment compared to earlier runs. The integration with a CI pipeline, and specifically the GitLab CI pipeline as described in (Elbers et al., 2019), ensures that an image is created as soon as a new version (git tag) is released. This image provides the exact, tagged and verifiable (commit hash), run time environment to repeat the experiment as long as the image is available, even if the image build is failing because of package version, or other, issues. In addition to pinning versions in the container image, including scripts and dependencies at run time (via mounts) should also be avoided since it directly violates the requirement to embed all tools and packages into the container image.

Some other more generic best practices with respect to container images: try to reduce the resulting image size, e.g. when using Docker base your image on the Alpine Linux [26] image (alpine:3.11 [27] equals approximately 5.59MB) rather than the Ubuntu [28] image (ubuntu:18.04 [29] equals approximately 64.2MB), try to reduce the number of layers by grouping related commands, e.g. using one `RUN` statement and chain the commands using the `&&` operator.

Using the framework as described so far also requires the researchers to adopt a new workflow to produce results. Building this container image is not a step one has to perform after the fact, just before publishing the results. Rather this framework and the introduced best practices should be included into the research life cycle from the beginning and all results published should be obtained by running a properly tagged and published container image, run exactly according to the input specification. This guarantees that as long as the container image is available from the public container image registry and the dataset is available, the research can be reproduced. As long as the container image is available, the whole tool

[26] https://alpinelinux.org/

[27] https://hub.docker.com/_/alpine/

[28] https://ubuntu.com/

[29] https://hub.docker.com/_/ubuntu/

chain can also be used against different inputs without any additional effort. If, for whatever reason, the container image is not available anymore, the Dockerfile can be used to rebuild the container image. If any of the required dependencies are no longer available, the Dockerfile will have documented all required dependencies with explicit versions, commands and command line arguments. We advocate to take over all data management best practices that have been developed in the area of long term preservation for research data for container images as well.

4 Challenges and Pitfalls

Some of the possible problems that can be encountered were already addressed within the case studies (see section 2.2 and section 3.2), but we want to highlight them here again.

On the data side the most important problem that we encountered is that linguistic data is, by its very nature, personal data, which means that there will always be privacy concerns when sharing language corpora with the wider research community. Protecting data from some people while giving access to others is relatively easy on the side of a research data repository - after all that is one of their key functions - but it proves to be a bit more difficult on the side of a code hosting repository, where there normally is no need for fine-grained access permissions. The use of git submodules as described in section 2.2 seems to be an efficient solution to this problem, but it remains to be seen how well this works for more complicated authorisation scenarios.

There is also, as always when using external services for sensitive data, the consideration whether one should store data with a commercial provider, especially one based in another country and jurisdiction. Apart from privacy concerns, there is also the possibility that a commercial provider suddenly goes out of business or decides to move away from this type of service. One way to avoid these problems is the GitLab “community edition”[30] that is avaible as open source and can be installed on local infrastructure, meaning that the researcher/the institute will be able to keep full control of the data and container hosting.

On the technological side, there are also a few problems that need to be addressed. For example, `git` has just arrived at its third iteration of dealing with large files: First, `git annex`, then `git lfs`, and now `git` partial `clone`[31]. This means that the technological barrier to deal with large files in one of the more common VSCs has – at best – just been made a tad smaller, but whether this latest iteration will finally solve the problem remains to be seen. Of course, there exist more mature solutions to this problem, like Data Science Version Control (DVC)[32] that extends `git`’s functionality with explicit support for large files, but this has the disadvantage of introducing another git-like-tool into the mix. Albeit `git`’s popularity, if even ”more experienced Git users requested that someone else perform an operation” because they were scared (Church et al., 2014), it is certainly no exaggeration to say that one of the very important reasons for `git`’s success is ”[i]n a word, GitHub” [33] or, more generally, the online platforms. Overall, the online repository hosting platforms make many tasks easier while dealing with the intricacies of `git`, and for any additional tool, especially one that bears similarities with `git`, there is a need for such aid.

Another possible pitfall lies in the use of Dockerfiles to create a persistent setup of a tool-chain. Unfortunately, when writing a Dockerfile there is currently no enforcement of explicit versioning of the docker software used with the Dockerfile. This means that two containers built using the same Dockerfile at two different points in time can contain two slightly different versions of the software which may result in different behaviour. The software specifications within the Dockerfiles rely on versioned resources from the software builder. However, these are not always available. This is even true for the base images (e.g. Ubuntu) hosted at Docker Hub. They are constantly updated with the latest patches which might result in different behaviour. Accessing older versions of e.g. the Ubuntu image on Docker Hub is very difficult to impossible. A similar problem exists for the container image registry, where there is not necessarily an explicit connection between the image and the Dockerfile (version) that has been used to create it. All

[30] `https://about.gitlab.com/install/ce-or-ee/`

[31] `https://about.gitlab.com/blog/2020/03/13/partial-clone-for-massive-repositories/`

[32] `https://dvc.org/`

[33] `http://blogs.nature.com/naturejobs/2018/06/11/git-the-reproducibility-tool-scientists-love-to-hate/`

images exist more or less on their own and there is no mechanism with which one can mark an image as a successor or predecessor of another image.

The framework for reproducible research discussed in section 3.2 provides a solution to this challenge of making sure tool versions do not differ, as long as proper data management is applied to the container images. While the container image is accessible, guarantees on stability of the environment can be made and there should be no differences in versions of software. Only if the container image is lost or no longer accessible, these guarantees can no longer be made.

The biggest challenge to us, however, seems to be to change the workflow of researchers to incorporate this way of doing research into their projects from the beginning. Only if both data and processing software are properly managed from the very start of a new project, can it be assured that the research that is conducted within the project will be reproducible by other researchers later on. Especially the need for very explicit templates and usage guidelines as described in section 3.2 needs to be well-thought-out and then followed precisely.

5 Conclusion and Outlook

Reproducibility of corpus-linguistic research is a central problem within the linguistic community (Wieling et al., 2018) which is currently not well addressed in a large number of projects. In this paper, we have presented a promising approach to tackle this problem using existing tools from the software development world. We have showed some case studies where this approach has been implemented, but already encountered a number of potential problems of which we highlighted some. Nevertheless, this way of ensuring that results in corpus-based linguistic research can easily be reproduced by fellow researchers seems like an idea that is worth pursuing in the future.

CLARIN ERIC is already working on establishing best practices and a framework that can be used to improve the reproducibility of research in the way described in this paper. At the time of writing, a pilot test of this framework is ongoing in the scope of the Language Resources and Evaluation Conference (LREC) 2020[34], where it provides the technical basis for the REPROLANG 2020[35] shared task. Thereafter, CLARIN ERIC will publish a follow up article providing a thorough analysis of the pilot results and suggesting further improvements based on this concrete experience. The proposed framework and best practices are all based on open source technologies and public service offerings for publicly available data. Proper exit strategies can be put in place to deal with worst case scenarios where some of these services might disappear or become unusable for whatever reason. We are confident that these exit strategies can be implemented for all services used in the described approach.

References

Adam Brinckman, Kyle Chard, Niall Gaffney, Mihael Hategan, Matthew B. Jones, Kacper Kowalik, Sivakumar Kulasekaran, Bertram Ludäscher, Bryce D. Mecum, Jarek Nabrzyski, Victoria Stodden, Ian J. Taylor, Matthew J. Turk, and Kandace Turner. 2019. Computing environments for reproducibility: Capturing the “Whole Tale”. *Future Generation Computer Systems*, 94:854–867, May.

Luke Church, Emma Söderberg, and Elayabharath Elango. 2014. A case of computational thinking: The subtle effect of hidden dependencies on the user experience of version control. In Benedict du Boulay and Judith Good, editors, *Proceedings of the 25th Annual Workshop of the Psychology of Programming Interest Group, PPIG 2014, Brighton, UK, June 25-27, 2014*, page 16. Psychology of Programming Interest Group.

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany J. Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. Three Dimensions of Reproducibility in Natural Language Processing. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 156–165, May.

Willem Elbers, Egon W. Stemle, André Moreira, Alexander König, Luca Cattani, and Martin Palma. 2019. The clarin eric deployment infrastructure and its applicability to reproducible research.

[34] https://lrec2020.lrec-conf.org/

[35] https://lrec2020.lrec-conf.org/en/reprolang2020/

Jennifer-Carmen Frey, Aivars Glaznieks, and Egon W. Stemle. 2016. The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro, and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, Napoli, Italy, December.

Jonas Kuhn. 2019. Computational text analysis within the Humanities: How to combine working practices from the contributing fields? *Language Resources and Evaluation*, 53(4):565–602, December.

Willem JM Levelt, PJD Drenth, and E Noort. 2012. Flawed science: The fraudulent research practices of social psychologist diederik stapel.

Joel Nothman, Hanmin Qin, and Roman Yurchak. 2018. Stop Word Lists in Free Open-source Software Packages. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 7–12, Melbourne, AU, July. Association for Computational Linguistics.

Daniel Nüst, Carlos Granell, Barbara Hofer, Markus Konkol, Frank O. Ostermann, Rusne Sileryte, and Valentina Cerutti. 2018. Reproducible research and GIScience: an evaluation using AGILE conference papers. *PeerJ*, 6:e5072, July.

Karthik Ram. 2013. Git can facilitate greater reproducibility and increased transparency in science. *Source Code for Biology and Medicine*, 8(1):7, February.

Victoria Stodden, Friedrich Leisch, and Roger D. Peng, editors. 2018. *Implementing Reproducible Research*. Chapman and Hall/CRC, 1 edition, December.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Reproducibility in Computational Linguistics: Are We Willing to Share? *Computational Linguistics*, 44(4):641–649, December.

A CLARIN Contractual Framework for Sharing Personal Data for Scientific Research

Krister Lindén
University of Helsinki
Finland
krister.linden@helsinki.fi

Aleksei Kelli
University of Tartu
Estonia
aleksei.kelli@ut.ee

Alexandros Nousias
Nousias/Linardos Business & Legal Consultants
alexandros.nousias@gmail.com

Abstract

The development and use of language resources often involve the processing of personal data. Processing has to have a legal ground. The General Data Protection Regulation (GDPR) provides several legal grounds. In the context of scientific research, *Consent*, *Public interest* and *Legitimate interest* are relevant. The main question is when researchers should rely on Consent and when on Public or Legitimate interest to conduct research. All three grounds have their advantages and challenges. For comparing Consent and Public interest, the Clinical Trial Regulation is used as an example. In addition, we study how this has been implemented in Finland based on the guidelines for research data from the Data Protection Ombudsman and suggest an update of the CLARIN Deposition License Agreement templates to accommodate data sets with personal data.

1 Introduction

Language resources (LRs) contain material subject to various legal regimes. For instance, they may contain copyright protected works, objects of related rights (performances) and personal data. This affects the way language resources are collected and used. Intellectual property issues relating to language resources have previously been addressed by Kelli et al. (2016). The general approach to dealing with personal data[1] in research is outlined by Kelli et al. (2019), where they discuss how processing[2] personal data without consent (see also Klavan et al. 2018) as the legal basis is possible in various EU countries.

Personal data issues are relevant for language resources, given that they potentially contain oral speech or written text, which relate to a natural person.[3] According to the Charter of Fundamental Rights of the European Union (Charter 2012) "*Everyone has the right to the protection of personal data concerning him or her*" (Art. 8 (1)). The general framework for personal data protection is provided in the General Data Protection Regulation[4] (GDPR). This paper primarily outlines the regulatory framework for processing personal data for research purposes. Special attention is given to the legal grounds for processing for research purposes.

[1] The General Data Protection Regulation defines personal data as "*any information relating to an identified or identifiable natural person ('data subject')*" (Art. 4 (1)).

[2] Processing is defined as "*any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction*" (Art. 4 (2)).

[3] It should be emphasised that in case data is anonymised then data protection laws do not apply (see recital 26 of GDPR). For further discussion on anonymisation techniques, see WP29 2014.

[4] The GDPR is applicable in all EU Member States from 25 May 2018. It replaces the (Data Protection Directive).

Krister Lindén, Aleksei Kelli and Alexandros Nousias 2020. A CLARIN Contractual Framework for Sharing Personal Data for Scientific Research. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 75–84.

The Charter of Fundamental Rights of the European Union foresees that personal data "*must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law*" (Art. 8 (2)). In addition to consent, the GDPR provides other legal grounds for processing personal data as well. We focus on two legal grounds relevant for processing personal data for research purposes, i.e. consent and the use of data for public interest research while also alluding to some conditions for using legitimate interest.

The choice of legal ground, however, does not affect the other obligations of the data controller[5] under the GDPR. For instance, when using data for research purposes in the public interest, we also need to consider whether and how it is feasible to inform the data subjects.

If we get data directly from the data subjects, the GDPR requires that the data subjects be informed (GDPR, Art. 13), i.e. we need to document that the data subjects have been aware of our activity. If on the other hand, we reuse personal data from large databases or publicly available sources, it is not always possible to inform the data subjects in person.

The focus of this study is to show how one can collect personal data directly from the data subjects while still using research in the public interest as a legal ground. We call this model confirmation to participate in research of public interest, and show that this is not just a hypothetical model but it is already in use in the Clinical Trial Regulation (CTR 2014). Framed in the terminology of the GDPR, the CTR model is based on processing data for a task carried out in the public interest. The data subject nevertheless has to be informed about the processing and must consent to participating in the trial, i.e., the data subject gives informed consent to participate in the research and may end participation at any time. However, according to the CTR, the data subject does not give specific consent to process personal data in the way in which consent is defined in the GDPR. The CTR consent has therefore also been labled "broad consent"[6], "ethical consent" or "consent to participate", which has important consequences for the right of the data subject to limit processing of the data (e.g., the right to withdraw consent to process the data at any time).

Other types of research than clinical trials can also be carried out in the public interest using the same model for processing personal data, i.e., by *confirmation to participate in research carried out in the public interest*. Since key concepts of the data protection framework (personal data, data subject, etc.) are addressed in previous CLARIN publications (Kelli et al. 2016, 2019), they are not repeated here.

We first explore the consequences of GDPR enabling research in the public interset as a legal ground for doing scientific research vs. using GDPR consent as the legal ground in Sections 2-4. We then proceed to outline the documents that are needed to implement both of them in practice in Sections 5-6.

2 Processing personal data for research purposes according to the GDPR

Processing of personal data for research purposes can be illustrated with the graph in Fig. 1. The GDPR defines research broadly so that it covers "*technological development and demonstration, fundamental research, applied research and privately funded research*" (Recital 159). The GDPR provides the following requirements for processing data for research purposes (Art. 89):

1) processing for research purposes is subject to appropriate safeguards. The safeguards ensure that technical and organisational measures are in place in particular to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner;
2) the Member States may limit the following data subject's rights for research purposes (optional limitations):
    a) the right of access by the data subject (Art. 15);
    b) the right to rectification (Art. 16);

[5] According to the GDPR the controller is defined as "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data" (Art. 4 (7)).

[6] Chassang (2017) discusses the difference between the "broad consent" in CTR and the specific consent in GDPR (as well as in the preceding directive from 1994) where he regards CTR as *lex specialis*. However, CTR predates GDPR, which now accommodates the CTR-style consent under research in the public interest allowing us to apply it also to other areas of research with public interest as the legal ground.

c) the right to the restriction of processing (Art. 18);
d) the right to object (Art. 21);

The implementation of the optional limitations varies by country and they are exemplified and discussed in Kelli et al. (2019). However, there are two mandatory limitations[7] to the rights of data subjects with regard to research data: 1) the right to be forgotten[8] and 2) the right to be informed about the processing.[9]

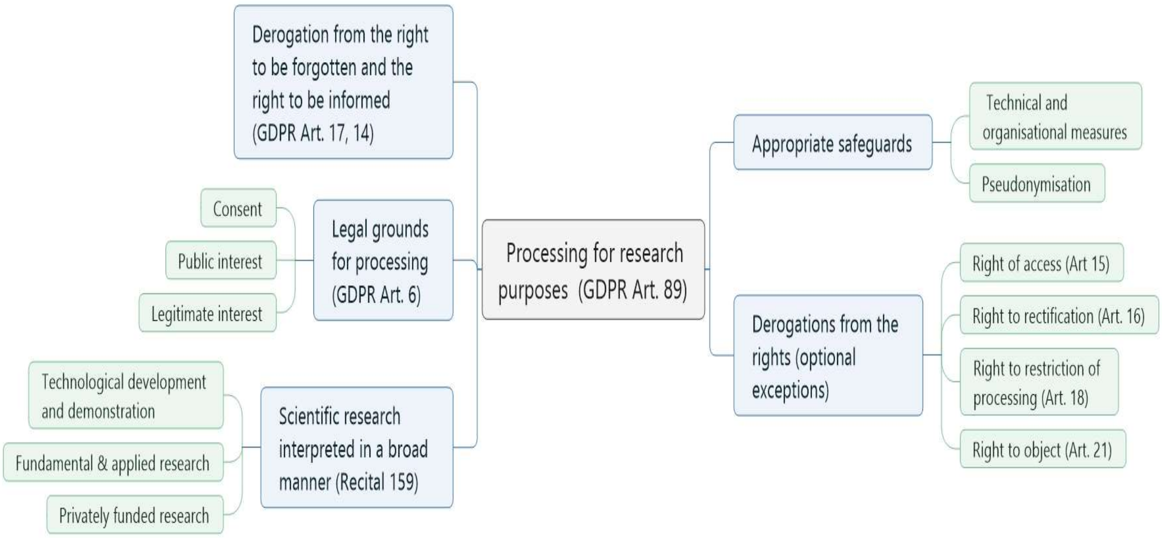

Derogation from the right to be forgotten and the right to be informed (GDPR Art. 17, 14)
Consent
Public interest
Legitimate interest
Legal grounds for processing (GDPR Art. 6)
Technological development and demonstration
Fundamental & applied research
Privately funded research
Scientific research interpreted in a broad manner (Recital 159)
Processing for research purposes (GDPR Art. 89)
Appropriate safeguards
Technical and organisational measures
Pseudonymisation
Derogations from the rights (optional exceptions)
Right of access (Art 15)
Right to rectification (Art. 16)
Right to restriction of processing (Art. 18)
Right to object (Art. 21)

Figure 1. Processing personal data for research

When processing data for scientific research "*further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes*" (GDPR, Art. 5b). The European Data Protection Supervisor gave an opinion on how to interpret this in January 2020 (EDPS 2020): "*The presumption is not a general authorisation to further process data in all cases for historical, statistical or scientific purposes. Each case must be considered on its own merits and circumstances. But in principle personal data collected in the commercial or healthcare context, for example, may be further used for scientific research purposes, by the original or a new controller, if appropriate safeguards are in place.*"

However, there is an additional aspect, which needs to be kept in mind before we focus on the separate legal grounds for processing personal data. Namely, it is essential that the legal ground for processing is chosen before the processing starts and its subsequent replacement could be problematic. According to the guidelines on consent "*the controller cannot swap from consent to other lawful bases. For example, it is not allowed retrospectively to utilise the legitimate interest basis in order to justify processing, where problems have been encountered with the validity of consent. Because of the requirement to disclose the lawful basis which the controller is relying upon at the time of collection of personal data, controllers must have decided in advance of collection what the applicable lawful basis is*" (WP29 2018: 23). It therefore makes sense to use a legal ground, which is as amenable to scientific research as possible when collecting new data in order to avoid unnecessary difficulties for further processing.

The GDPR provides six legal grounds for processing personal data: 1) consent; 2) performance of a contract; 3) compliance with a legal obligation; 4) protection of the vital interests; 5) the public interest or in the exercise of official authority; 6) legitimate interest (Art. 6). The processing for research purposes is not an individual legal ground. The processing for research purposes can rely on consent, the

[7] Mandatory limitations are directly applicable. They do not need to be incorporated into the national laws.
[8] GDPR Art. 17 (3) d.
[9] GDPR Art. 14 (5) b.

performance of a task carried out in the public interest, or legitimate interest[10]. For comparison, we briefly outline the three legal grounds that can be used in scientific research: consent, public interest and legitimate interest.

2.1 Consent

The processing based on the data subject's consent (Art. 6 (1) a) offers the highest possible protection for the data subject through the following mechanism:

1) the consent has to be freely given, specific, informed and unambiguous (Art. 4 (11));
2) the data subject can withdraw the consent without any detriment (Art. 7 (3));
3) the burden of proof lies with the controller (Art. 7 (1))

The Article 29 Working Party[11] (WP29) explains that consent "*focuses on the self-determination of the data subject as a ground for legitimacy. All other grounds, in contrast, allow processing – subject to safeguards and measures – in situations where, irrespective of consent, it is appropriate and necessary to process the data within a certain context in pursuit of a specific legitimate interest*" (2014a: 13). In case the acquisition of consent is complicated or administratively burdensome (e.g., anonymous web posts, legacy resources, public videos and so forth), some other legal ground than consent is clearly needed.

2.2 Public interest

The GDPR names the performance of a task carried out in the public interest as a legal bases for processing personal data (Art. 6 (1) e)). According to WP29, the performance of a task carried out in the public interest is also a ground for processing personal data in the research context (2014a: 22).

The concept of research in the public interest can usually be invoked by research projects affiliated with universities or research institutions having a legal mandate to do research in the public interest, but it also allows for companies acting in the public interest on behalf of a Member State, e.g., to ascertain the safety and/or efficacy of a procedure performed in addition to normal clinical practice as outlined in the Clinical Trial Regulation (CTR 2014).

Since language data often contains both personal data and copyrighted works, within the framework of development of language technologies it is also relevant to take into consideration the concept of research as defined in the field of copyright law. The Digital Copyright Directive (DCD) provides a special regulation on the text and data mining (TDM) for the purposes of scientific research (Art. 3). According to the Digital Copyright Directive the concept of scientific research covers "*natural sciences and the human sciences*" (Recital 12). The TDM exception allows public-private partnerships (Recital 11). Research organisations are the main beneficiary of the TDM exception (DCD Art. 3). The Digital Copyright Directive defines them as follows (Art. 2 (1)): "a *'research organisation' means a university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research:*

(a) on a not-for-profit basis or by reinvesting all the profits in its scientific research; or

(b) pursuant to a public interest mission recognised by a Member State;

in such a way that the access to the results generated by such scientific research cannot be enjoyed on a preferential basis by an undertaking that exercises a decisive influence upon such organisation". After the harmonisation of the Digital Copyright Directive with national laws, the definition most likely becomes relevant for personal data processing as well since research organisations as such are defined generally and not only for copyright purposes.[12]

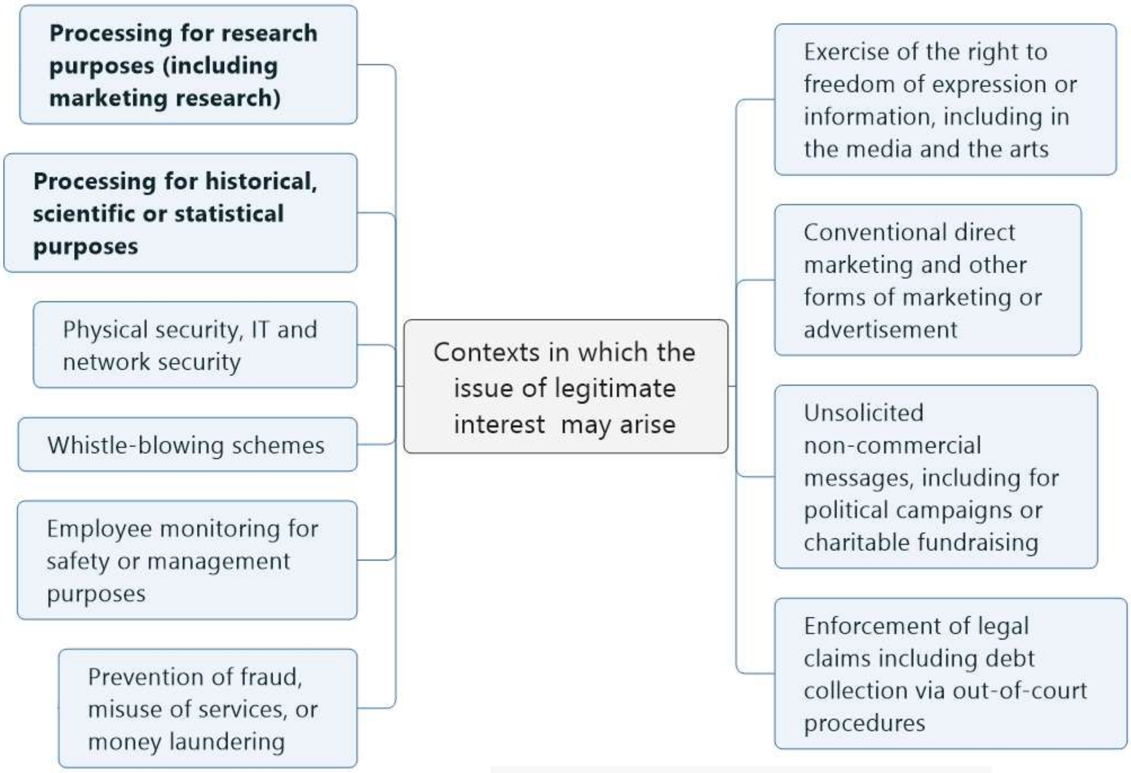
[10] Note that legitimate interest as a legal ground for research needs to be argued when one cannot claim to be acting with permission or in the interest of the data subject (legal grounds 1, 2 or 4 in Art. 6) or in the interest of the state (legal grounds 3 or 5 in Art. 6).

[11] The Article 29 Working Party (Art. 29 WP) is the independent European working party that dealt with issues relating to the protection of privacy and personal data until 25 May 2018, at which point it was succeeded by the European Data Protection Board.

[12] During the discussion concerning the implementation of the Digital Copyright Directive in the Ministry of Justice of Estonia (23-24 September 2019), it was indicated that the Estonian Organisation of Research and Develop-

2.3 Legitimate interest

Pursuant to the GDPR, the processing is lawful if it is "*necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data*" (Art. 6 (1) clause f). The guidelines provide several contexts where the legitimate interest could potentially serve as a legal ground for processing personal data. Such contexts are, e.g. "*conventional direct marketing and other forms of marketing or advertisement; unsolicited non-commercial messages, including for political campaigns or charitable fundraising; prevention of fraud, misuse of services, or money laundering; employee monitoring for safety or management purposes; whistle-blowing schemes; physical security, IT and network security, etc.*" (WP29 2014a: 25). As can be inferred from the examples, the legitimate interest covers purposes (including research purposes) that are typical for organisations (including commercial organisations) for which the interests of the organisation need to be explicitly argued to override the interests of the data subject. Fig. 2 illustrates the context for using legitimate interest.


Contexts in which the issue of legitimate interest may arise
Processing for research purposes (including marketing research)
Processing for historical, scientific or statistical purposes
Physical security, IT and network security
Whistle-blowing schemes
Employee monitoring for safety or management purposes
Prevention of fraud, misuse of services, or money laundering
Exercise of the right to freedom of expression or information, including in the media and the arts
Conventional direct marketing and other forms of marketing or advertisement
Unsolicited non-commercial messages, including for political campaigns or charitable fundraising
Enforcement of legal claims including debt collection via out-of-court procedures

WP29. Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC

Figure 2. Legitimate interest contexts

3 Data Processing according to the Clinical Trial Regulation

The Clinical Trial Regulation (CTR) entered into force on 16 June 2014, but the timing of its application depends on the development of a fully functional EU clinical trials portal and database, which will be confirmed by an independent audit. The Regulation becomes applicable six months after the European Commission publishes a notice of this confirmation, which is likely to occur in 2020. The GDPR refers to CTR for special requirements for the consent to participate in scientific research activities in

ment Act is analysed and amended (if needed) to make it compatible with the Digital Copyright Directive. According to the Estonian Organisation of Research and Development Act (ORDA) a research and development institution is a legal person or an institution in the case of which the principal activity is carrying out basic research, applied research or development, or several of the aforementioned activities (§ 3 (1) clause 1).

clinical trials (Recital 161). The informed consent in CTR is often confused with GDPR Consent. In this paper, we try to highlight the difference between *GDPR Consent,* which is a legal ground, and *informed consent,* which is a protective measure when the legal ground is *GDPR Public interest*.

In CTR (Article 2(2)21), *'informed consent' means a subject's free and voluntary expression of his or her willingness to participate in a particular clinical trial, after having been informed of all aspects of the clinical trial that are relevant to the subject's decision to participate or, in case of minors and of incapacitated subjects, an authorisation or agreement from their legally designated representative to include them in the clinical trial*. This informed consent constitutes a confirmation to participate in a clinical trial. Special requirements for the informed consent are provided in Chapter V of the CTR.

About withdrawal of the informed consent, CTR says in its Recital (76) ..., *while safeguarding the robustness and reliability of data from clinical trials used for scientific purposes and the safety of subjects participating in clinical trials, it is appropriate to provide that, without prejudice to Directive 95/46/EC [now replaced by GDPR], the withdrawal of informed consent should not affect the results of activities already carried out, such as the storage and use of data obtained on the basis of informed consent before withdrawal.* This means that the withdrawal of the informed consent only implies that the data subject stops participating in the trial. Data collected during the participation can still be stored, e.g. for verifying the results of activities already carried out.[13]

The legal basis for a clinical trial is research on behalf of a Member State to ascertain the safety and efficacy of a procedure performed in addition to normal clinical practice, which is a prototypical case for research in the public interest. However, for clinical trials, the approval of an ethics committee is also needed, because a non-standard clinical procedure will be applied potentially affecting the well-being of the data subjects.

In GDPR terminology, the CTR policy can be restated as the data subject's informed consent to participate in research in the public interest. The requirement to provide information about the research is a protective measure so that the informant can decide whether he consents to participating in the research. He does not explicitly consent to any specific use of the provided data as the legal ground for collecting the data is the public interest in the potential outcome of the research, which has consequences for withdrawing provided data.[14] For further details on the interaction between GDPR and CTR when reusing data for scientific research, see Pormeister (2020).

4 Discussion of consequences for other types of research

Other types of research than clinical trials can also be carried out in the public interest by research projects affiliated with institutions having a legal mandate to do research in the public interest. They can use the same legal ground for processing personal data. If research is carried out in the public interest and data is reused from large databases or public sources, it is not always possible to inform all data subjects about the processing, in which case a public record of processing activities may be deemed sufficient.

However, if data is collected directly from the data subjects, and the legal basis according to GDPR is *research in the public interest*, it may be useful to avoid confusion among data subjects and researchers alike by naming the verification for having provided mandatory information "*confirmation to participate in research carried out in the public interest".* The confirmation relies on informed consent to participate, so the consent is given only with regard to the participation and not with regard to the processing of the personal data, for which the legal basis is research in the public interest.

Since research is conducted in the public interest, the data subject's right to be forgotten is mandatorily limited by (GDPR Art. 17 (3)d) to the extent that processing is necessary for research purposes in so far as the right to be forgotten is likely to render impossible or seriously impair the achievement of the objectives of that processing. In many cases, the right to be forgotten could impair the replicability of the research.

[13] The GDPR has a similar approach. It provides that "*The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal*" (Art. 7 (3)). However, GDPR also provides for storage extension for research (Art. 5) and a research purpose extension (Art. 5).

[14] As data is collected for research purposes, there is also a limited right to be forgotten (Art. 17).

It should be noted that data collected in this way can only be used for research purposes. According to the GDPR[15], such data can still be reused for other research purposes (GDPR Art. 5), but the transfer of the data to another research organisation needs to be protected to make sure that the data is processed only for research purposes[16]. If one wishes to make such personal data publicly available, e.g. as an illustrating example or a video clip potentially identifying the data subject on the internet or at a conference, *GDPR Consent* must be acquired from the data subject.

The legal framework for processing personal data for research purposes is based on the GDPR and national laws of the EU Member States. This means that in addition to the mandatory limitations of the rights of data subjects enforced by the GDPR, researchers that wish to develop language resources containing personal data may have further rights to maintain the research data integrity through nationally implemented limitations to the rights of data subjects.

5 Finland as a case study for research in the public interest as a legal ground

In Finland, the new Data Protection Act (FINLEX 1050/2018) entered into force on January 1, 2019. Collecting personal data to be used for scientific research purposes is currently shifting towards using research in the public interest as the predominant legal ground, while the preferred way to obtain personal data under the previous legislation was consent to use data. This is due to the fact that Finland was among the countries that took nearly full advantage of the leeway allowed in the GDPR to enact derogations to the rights of the data subjects when using data for scientific research in the public interest. As argued in the previous Sections of this paper, this is not something radically new introduced by the GDPR, it is merely extending a practice that has already been established in a clinical research setting to other domains of scientific research. To distinguish scientific research from other kinds of research, the Supreme Administrative Court (SAC) of Finland has established (FINLEX KHO:2013:181) that providing confidential information for scientific research requires: 1) an adequate research plan, 2) sufficient scientific qualifications of the participating researchers, 3) autonomous and public agency of the researchers and 4) scientific research objectives.

The Office of the Data Protection Ombudsman (DPO) of Finland published guidelines on January 30, 2020, for how to use personal data in research "The path to data protection in research" (Data Protection Ombudsman 2020), which for practical purposes are more detailed with regard to consent without prejudice to other available options such as scientific research in the public interest. The DPO points out that the GDPR takes a broad perspective on scientific research, i.e. scientific research purposes may include development and presentation of technology, basic research, applied research or privately funded research, but this does not extend the definition of scientific research beyond what has been prescribed by the SAC. In particular, the DPO emphasizes that scientific and historical research as well as statistical purposes entail increasing public knowledge in society, e.g. combining information from several data sources may provide valuable new information for disease prevention or treatment, which may serve as a basis for new policies improving the well-being of citizens and improve the efficacy of welfare services.

The guidelines recognize that the GDPR aims to facilitate access to data for scientific and historical research. As mentioned above, according to the GDPR all processing of personal data for scientific and historical research purposes is compatible with the original purpose if applying adequate protective and organisational measures, e.g. using data encryption with authorized access. However, further processing for scientific research purposes still requires taking all the other aspects of the data protection legislation into account, e.g. the data subjects need to be informed before the processing starts unless there are compelling and legitimate reasons. As it is not always possible to reach the data subjects personally, some other sufficiently public means may also be used to inform them about the further processing of the data. The DPO points out that a research purpose that is incompatible with the original purpose, for which the data was originally collected, is still possible but requires that the data subject be informed,

[15] GDPR is applicable only within the EU, but many countries have agreements with the EU, see e.g. https://gdpr-info.eu/issues/third-countries/

[16] The reuse still needs to answer to general GDPR requirements such as the data minimisation principle. According to Article 89(1) and Recital 156, for further processing, the controller should also assess the feasibility to fulfil the reuse purposes by processing anonymous or pseudonymous data.

and if the legal ground for collecting data was GDPR Consent, a new consent is needed for such incompatible further processing.[17] The fact that personal data collections can be reused if fulfilling the various criteria imposed by the GDPR is fundamental for a research infrastructure like CLARIN.

For a research infrastructure, the conditions on which the data can be communicated to others are key issues: the foremost among them are the purpose of use and the storage conditions of the personal data set. With regard to the definition of the original research purpose, the DPO guidelines are vague saying only that sweeping definitions such as "future research" or "general research purposes" are too broad and therefore inadequate. This still leaves us with more specific options, e.g. "scientific research for the purposes of humanities and social sciences" or "scientific research on language-aware technology" limiting the use to particular fields of scientific research. This may be problematic if we foresee multi-disciplinary research on the data. On the other hand, specifying one or more research questions such as "scientific research for enhancing the well-being of citizens" limits future research topics but allows multi-disciplinary research for such purposes.

Data should be stored only to the extent needed, i.e. neither the type of data nor the time-span should be excessive. However, it is usually necessary to keep data for verification of the research results and sometimes more data will be collected in longitudinal studies, in which case the original data needs to be kept for comparison. If an end-point is difficult to define, one may instead set up criteria for re-evaluating the need to keep the data at regular intervals. If the longitudinal study spans several decades, the original data subjects will eventually pass away, at which time the data is no longer personal.

6 CLARIN Deposition License Agreements

To facilitate the adoption of the requirements of the GDPR, the CLARIN License Agreements need to be updated with provisions for communicating data sets containing personal data. In general, this was the intention for the category of RES license agreements (for further discussion on categories, see Oksanen et al. 2010; Kelli et al. 2018), but very limited amounts of personal data were available for transfer and reuse at the time when the template was designed. With the GDPR, the definition of personal data has been considerably expanded and many more data sets may contain personal data so specific provisions for communicating such data needs to be adopted.

In addition, the GDPR assumes that an institution, which potentially has an appointed data protection officer, takes responsibility for the personal data set and that a record of processing activities is established by the organisation. An individual researcher may not have the authority to bind his organisation contractually to such activities, so for personal data sets the end-user is not primarily an individual researcher but a research organisation.

With an increased number of data sets identified as containing personal data, there is a rather motely practice for personal data set transfer within the EU. It may therefore be useful to agree on similar rules for personal data exchange at least within a national CLARIN consortium and further between organisations in the various CLARIN member states.

From a CLARIN perspective, the proposed agreement structure aims to establish a CLARIN B or C Centre as a Data Processor for the national CLARIN consortium with each of the consortium members, or some external party, as a Controller of its own personal data sets. To this end, we propose modifications that follow the suggestions proposed already by Kelli & al. (2016). The main agreement is renamed as the CLARIN Framework Deposition Agreement (FADA) with two appendices:

1) the Data Protection Agreement (DAPA), and
2) the Deposition License Agreements (DELA).

The CLARIN FADA establishes a framework of common deposition rules for data sets that can be communicated by a CLARIN Centre.

Individual data sets are added as attachments to the CLARIN FADA keeping only data set specific information in the DELA, which thereby reduces to a 1-page main document for each data set referring to the general conditions in the FADA and to four data set specific appendixes:

1) the data identification, description and citation texts,
2) the deposition license conditions with an end-user license agreement template,

[17] Note that further processing to verify previous research results is therefore allowed without renewal of consent, and such research may be varied and extensive as long as the final purpose is to verify previous results.

3) a list of third party copyrights or database rights, and
4) the personal data description and the purpose of use of the data set.
Appendixes 3 or 4 may explicitly be left empty if there are no third party rights or no personal data in the data set.

In CLARIN, the RES licensing scheme is suitable for communicating copies of data sets with personal data. In the suggestions for how to implement the ethical intent of the GDPR in a research setting, Pormeister (2020) recommends that the original controller stays informed about all further use of a personal data set in order to inform the data subjects about such further use when necessary. The CLARIN RES license requires that data sets not be communicated to a third party by the end-user, because new end-users can obtain a copy directly from CLARIN. As CLARIN remains a mere processor of personal data for the purpose of communicating such data to research organisations, the original controller stays informed about all requests for further use of a data set. If there is a request for using a data set for a research purpose which is not sufficiently compatible with the original purpose of use, the data subjects need to be informed. From a CLARIN perspective, it is mostly a practical question whether the CLARIN Centre as a processor is commissioned to inform the data subjects or the original controller informs them, and how one goes about informing them, i.e. is personal communication possible or is a public announcement sufficient.

7 Conclusion

The development and use of language resources often involves the processing of personal data. To process data for research purposes according to the GDPR, it is possible to invoke research in the public interest as the legal basis for publicly funded research projects carried out at research institutions with a legal mandate to do research in the public interest. If data is collected directly from data subjects, they have to be informed about the data processing so that they can opt-in by confirming their willingness to participate. If they no longer wish to participate, they have a right to stop the processing of their data, but they do not automatically have a right to be forgotten and their personal data may be reused for other research purposes. This model for using and reusing personal data is already established in the CTR, but it can be extended to other domains of scientific research as well through the provisions of the GDPR.

In the paper, we discuss how to implement this in practice through a CLARIN Framework Deposition Agreement containing the general provisions of the CLARIN Deposition License Agreement, with an appendix containing general data protection conditions and another appendix containing data set specific information.

References

[Charter 2012] Charter of Fundamental Rights of the European Union OJ C 326, 26.10.2012, p. 391-407. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT (6.2.2020).

[Chassang 2017] Gauthier Chassang (2017). The impact of the EU general data protection regulation on scientific research. *Ecancermedicalscience*, *11*, 709. doi:10.3332/ecancer.2017.709. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5243137/ (6.2.2020).

[CTR 2014] Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use. Avialable at https://ec.europa.eu/health/human-use/clinical-trials/regulation (31.1.2020).

[Data Protection Directive] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L 281, 23/11/1995 p. 0031 – 0050. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31995L0046&qid=1522340616101&from=EN (29.3.2020).

[Data Protection Ombudsman 2020] Guidelines on Scientific Research published in Finnish on January 30, 2020 [Tieteellinen tutkimus] Available in Finnish at: https://tietosuoja.fi/tieteellinen-tutkimus (31.1.2020).

[Digital Copyright Directive = DCD] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. OJ L 130, 17.5.2019, pp. 92-125. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1572352552633&uri=CELEX:32019L0790 (26.1.2020).

[GDPR] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1-88. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1515793631105&uri=CELEX:32016R0679 (5.2.2020).

[EDPS 2020] A Preliminary Opinion on data protection and scientific research by the European Data Protection Supervisor. 6.1.2020. Available at: https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf (31.1.2020).

[FINLEX KHO:2013:181] Decisions by the Supreme Adminstrative Court in Finland concerning "Asiakirjajulkisuus - Tutkimuslupa - Tieteellinen tutkimus - Tutkimuksen tieteellisyys - Reseptitiedosto - Lääkeyritys - Kansaneläkelaitos" [Public access to documents – Research permit – Scientific Research – Conditions for scientific research – Medical prescription data files – Medical company – the Social Insurance Institution of Finland]. Available in Finnish at: https://www.finlex.fi/fi/oikeus/kho/vuosikirjat/2013/201303651 (31.1.2020).

[FINLEX 1050/2018] Data Protection Act enacted by the Parliament of Finland (English translation). Available at: https://www.finlex.fi/en/laki/kaannokset/2018/en20181050 (31.1.2020).

[Kelli et al. 2019] Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramunas Birštonas, Silvia Calamai, Penny Labrpolou, Maria Gavrilidou, Pavel Straňák (2019). Processing personal data without the consent of the data subject for the development and use of language resources. In: Inguna Skadina, Maria Eskevich (Ed.). Selected papers from the CLARIN Annual Conference 2018. Linköping University Electronic Press, 72-82. Available at http://www.ep.liu.se/ecp/159/008/ecp18159008.pdf (29.1.2020).

[Kelli et al. 2018] Aleksei Kelli, Krister Lindén, Kadri Vider, Penny Labropoulou, Erik Ketzan, Pawel Kamocki, Pavel Straňák (2018). Implementation of an Open Science Policy in the context of management of CLARIN language resources: a need for changes? In: Maciej Piasecki (Ed.). Selected papers from the CLARIN Annual Conference 2017. Linköping University Electronic Press, 102-111. Available at http://www.ep.liu.se/ecp/147/009/ecp17147009.pdf (29.1.2020).

[Kelli et al. 2016] Aleksei Kelli, Kadri Vider, Krister Lindén (2016). The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. In: Koenraad De Smedt (Ed.). Selected Papers from the CLARIN Annual Conference 2015. Linköping University Electronic Press, 13-24. Available at http://www.ep.liu.se/ecp/article.asp?issue=123&article=002 (29.1.2020).

[Klavan et al. 2018] Jane Klavan, Arvi Tavast, Aleksei Kelli (2018). The Legal Aspects of Using Data from Linguistic Experiments for Creating Language Resources. Frontiers in Artificial Intelligence and Applications, 307, 71-78. Available at http://ebooks.iospress.nl/volumearticle/50306 (29.1.2020).

[Oksanen et al. 2010] Ville Oksanen, Krister Lindén, Hanna Westerlund (2010). Laundry Symbols and License Management: Practical Considerations for the Distribution of LRs based on experiences from CLARIN' in Proceedings of LREC 2010: Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management. Available at https://helda.helsinki.fi/handle/10138/29359 (5.2.2020). [ORDA] Organisation of Research and Development Act. Entry into force 2.05.1997. English translation available at https://www.riigiteataja.ee/en/eli/513042015012/consolide (21.1.2019)

[Pormeister 2020] Kärt Pormeister (2020). Transparency in relation to the data subject in genetic research – an analysis on the example of Estonia. Doctoral dissertation 76, School of Law, University of Tartu. p. 189. January 13, 2020. Available at: https://dspace.ut.ee/handle/10062/66697 (31.1.2020)

[VLO] CLARIN Virtual Language Observatory. Available at https://vlo.clarin.eu/ (6.2.2020).

[WP29 2018] Article 29 Working Party (WP29). Guidelines on consent under Regulation 2016/679. Adopted on 28 November 2017. As last Revised and Adopted on 10 April 2018. Available at https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051 (3.2.2020).

[WP29 2014] Article 29 Working Party (WP29). Opinion 05/2014 on Anonymisation Techniques. Available at https://iapp.org/media/pdf/resource_center/wp216_Anonymisation-Techniques_04-2014.pdf (3.2.2020).

[WP29 2014a] Article 29 Working Party (WP29). Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC. Available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf (3.2.2020).

[WP29 2007] Article 29 Working Party (WP29). Opinion 4/2007 on the concept of personal data. Adopted on 20th June. Available at http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf (6.2.2020).

CLARIN-IT and the Definition of a Digital Critical Edition for Ancient Greek Poetry

A New Project for Ancient Fragmentary Texts with a Complex Tradition

Anika Nicolosi
Department DUSIC
University of Parma, Italy
anika.nicolosi@unipr.it

Monica Monachini
ILC - CNR
Pisa, Italy
monica.monachini@ilc.cnr.it

Beatrice Nava
Alma Mater Studiorum
University of Bologna, Italy
beatrice.nava2@unibo.it

Abstract

Ancient Greek studies, and Classics in general, is a perfect field of investigation in Digital Humanities. Indeed, DH approaches could become a means of building models for complex realities, analyzing them with computational methods and sharing the results with a broader public. Ancient texts have a complex tradition, which includes many witnesses (texts that handed down other texts) and different typologies of supports (papyri, manuscripts, and epigraphs). These texts are the basis of all European Literatures and it is crucial to spread their knowledge, in a reliable and easy way. Our project on ancient Greek fragmentary poetry (DEA - *Digital Edition of Archilochus: New models and tools for authoring, editing and indexing an ancient Greek fragmentary author*), growing out of the existing experience, tries to define a TEI-based digital critical edition combined with NLP techniques and semantic web technologies. Our goal is to provide a complete and reliable tool for scholars, suitable for critical studies in Classics, and a user-friendly environment also for non-specialist users. The project represents one of the attempts within the context of CLARIN-IT to contribute to the wider impact of CLARIN on the specific Italian community interested in Digital Classics. It is intended to improve services in fostering new knowledge in SSH digital research and sustaining the existing one.

1 State of the art: Classical Studies in the Digital Era

The era of data outbreak and digital consumption has significantly changed and keeps changing it with increased emphasis on interdisciplinary studies and new technologies. In this context, even Classical Philology needs to provide new methods and concepts capable of both seizing opportunities and addressing the challenges that technologies can offer to the discipline, in particular to the development of a scholarly, born-digital, edition of the ancient texts. Conversely, in the field of language technologies, the needs for a wider audience, with diverse profiles, opens up new challenges: easily usable tools, adaptable to different types of content, become crucial. Texts in the humanities' fields can often be heterogeneous (by genre, by era, by type) and the quality of resources, in particular the quality of digital editions of texts, is receiving increasing attention.

Classics, in general, is a perfect field to demonstrate how Digital Humanities could become the humanist way of building complex models of complex realities, analyzing them with computational methods and communicating the results to a broader public. Digital Classics have undergone a great development, starting from the last ten years of the 20th century, and today we have many sources available and refined tools,[1] *e.g.* the main tools in the field as the Thesaurus Linguae Graecae (TLG)[2] and the

[1] For the state of the art of digital philology with a focus on Ancient Greek and Latin, see Berti (2019).

[2] It offers texts and grammatical analysis of the texts, linking to LSJ and other sources (http://stephanus.tlg.uci.edu/).

Anika Nicolosi, Monica Monachini and Beatrice Nava 2020. CLARIN-IT and the Definition of a Digital Critical Edition for Ancient Greek Poetry: a New Project for Ancient Fragmentary Texts with a Complex Tradition. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 85–93.

Perseus Digital Library (PDL).[3] There are also other very important projects, e.g. Digital Corpus of Literary Papyri (DCLP)[4] and Trismegistos[5] in general, Centre de Documentation de Papyrologie Littéraire (MP^3 - CeDoPaL),[6] The Leipzig Fragmentary Texts Open Series (LOFT),[7] Musisquedeoque,[8] Pinakes,[9] and bibliographical tools as Annè Philologique (APh).[10] Quite often these initiatives do not interact with each other, and they are only known by specialist users. At this stage, they are very important tools, but they cannot (or do not want to) replace paper editions.

We also have to consider that the treatment of literary texts currently does not correspond to scholars' expectations. These texts have a complex tradition, which includes many witnesses (texts that handed down other texts) and different typologies of supports (papyri, manuscripts and also inscriptions). To accomplish a complete understanding of certain topic, it is not enough to provide a traditional paper text; much more information is needed that scholars usually obtain by comparing several paper editions, lexica and/or more digital tools and imagines. We have now the opportunity to change this scenario, since digital technology allows us to manage much more data and adopt new approaches to a traditional discipline. Fostering the development of new learning habits, we can improve more modern research methodologies and new practices in didactics (always based on the good practices inherited from the previous tradition).

2 Ancient Greek Poetry and Digital Edition: texts with a complex tradition

DH may help to realize something new in the field. It is not enough today to only describe materials, and/or give information, and/or make available texts and/or analyze them, etc. It is necessary to integrate all these functions in a unique workbench where the researcher can find all (s)he needs. And to achieve this it is essential to focus on a specific case study. We think that the ancient Greek literature can be the right test, thanks to its potentialities and peculiarities.

Thanks to modern science we can preserve and hand down ancient Greek texts to next generations (not only among specialists in the field), exactly what ancient scholars did many centuries ago at the Alexandrian Library (3rd c. BCE). Fragmentary ancient Greek poetry is very different from other literary texts. In fact, its tradition is more complex since it has different kinds of sources (manuscripts, papyrus, epigraphs) with variants and *lacunae*. It is not enough to provide a single text randomly chosen; it is necessary to carry out a complete revision of the texts, according to their updated critical edition, and take into account all the textual proposals made by scholars. In this way, studying previous editions and the secondary literature, it is possible to enrich the new edition with new hypotheses.[11]

As noted by Pierazzo (2014), scholars agree that a digital edition, in addition to allowing direct or mediated access to a collection of information larger than the one currently available on paper, must be designed in such a way that ensures the greater freedom allowed by the tablet or the personal computer support. This digital versatility lacks in the paper edition. The digital medium, as we know, allows us to manage much more data than we can do in a paper edition, in which the apparatus must often be proportionate to the size of the page and therefore often provides only the main proposals. We can easily manage and store in a single place all the hypotheses of previous editors and the additional useful information linked to the edition; therefore, we can facilitate a new philological approach offering all the

[3] It offers texts, translations and grammatical analysis of the texts, linking to LSJ. Unfortunately, it offers miscellaneous text and presents old (out of copy-right) editions. For example, Archilochus, as a single author is not available; there are only texts and translations, with notes, of all iambographers, ed. by Edmonds in 1931 (http://www.perseus.tufts.edu/hopper/).

[4] The project, by Heidelberg and New York University, is an enlargement for literary papyrus' texts of Papyri.info (editing tool for documentaries papyrus' texts). DLCP and Papyri.info integrate tools for philological research, aggregating fundamental international resources for the study of antiquity; they provide resources to easily edit papyrus' texts, using Leiden+ mark-up language from EpiDOC, TEI/XML based, specific to ancient texts (http://litpap.info/).

[5] As, for example, Leuven Database of Ancient Books (LDAB) https://www.trismegistos.org/ldab/graphs_help.php

[6] http://web.philo.ulg.ac.be/cedopal/fr/

[7] For literary fragmentary ancient prose texts (http://www.dh.uni-leipzig.de/wo/lofts/) developed as the digital recognition of a traditional book about Istro's prose fragments (3rd c. BCE), linked to PDL, see Almas and Berti (2013).

[8] http://mizar.unive.it/mqdq/public/ricerca/avanzata

[9] https://pinakes.irht.cnrs.fr/

[10] http://www.brepols.net/Pages/BrowseBySeries.aspx?TreeSeries=APH-O

[11] See Nicolosi (2015).

interpretations of previous editors through a single resource, with the addition of new scholars' hypotheses.

At the same time, applying Natural Language Processing (NLP) techniques to a collection of fragmented texts can be particularly challenging because of the multiple options for text reconstruction. Automatic linguistic analyses of the whole *corpus* of fragments can indeed not only support new readings and interpretations but also lead us to a greater certainty as regards text corrections, integrations and authorship attribution. Moreover, by enriching the fragments with a set of specific lexica in appropriate format we facilitate data searchability and interoperability.

2.1 DEA Project: A new Digital and Critical Edition

DEA, which stands for *Digital Edition of Archilochus: New models and tools for authoring, editing and indexing an ancient Greek fragmentary author*, is a project lead by Anika Nicolosi (University of Parma), principal investigator, in collaboration with ILC-CNR of Pisa and CLARIN-IT. The project brings different expertise and skilful researchers together, combining philological and computational linguistic skills;[12] it takes advantage from philological studies about ancient Greek literature and from recent methodological approaches to language resources developments. Our project grows out of existing experiences and tries to define a new and complete digital edition of Archilochus' fragments, which includes the use of Linked Open Data (LOD).[13]

We have around 300 fragmentary poems by Archilochus, who lived in the 7th century BCE and who was closely related to Homer. This poet is crucial for Western literary tradition because Greek and Latin authors often mention him,[14] and many of his themes and motives (war, myth, love) are then echoed in modern western literature. Some fragments have only recently been published, however, what is currently lacking is a complete on-line critical edition of his works.[15]

The project plans to develop an innovative digital portal, to become a reference for the studies on the Ancient Greek author(s), DEA will not develop a tool that replicates or integrates the traditional ones; the new platform aims to become an essential resource for the study of an ancient author, scientifically updated and reliable. The main objective of the project is, hence, to provide scientifically reliable texts, with critical apparatuses, and translations, and to make available an online and easily accessible augmented corpus of ancient Greek fragmentary literature. For this reason, it is important to find an easy and effective way of managing all these data, also according to the dictates and the needs of the philological tradition.

The DEA project also aims at defining a methodology for a digital critical edition. Moreover, one of the DEA's main goal is to develop resources, materials and tools in order to improve research in Greek Classical philology, integrating the digital resources already available with new ones. The DEA project involves the digitization, with critical and philological control, of the whole *corpus* of Archilochus' fragments, supplemented with a complete apparatus. The result will be a new typology of a complete, augmented, scholarly, born-digital, edition (annotated and interoperable according to current standards, enriched with information from lexical and geographic knowledge) with particular attention to users' needs and requirements, and to usability and portability. DEA is well placed to become a model for the digitalization of authors with a complex text tradition, in particular, fragmentary ancient authors.

The result will be an innovative tool that will become fundamental to share research and data, by combining the accuracy of traditional philology with new and more intuitive use and access. The objective is to manage the critical information and support:

[12] It arises from the experience in the field of Prof. Nicolosi (University of Parma) who has several international studies in the field (journals and Meetings' Proceedings), and 3 monographs about Archilochus (2007, 2013, 2017), and the resources and tools developed by the Pisa group and hosted by the Italian CLARIN-IT national data centre.

[13] See Monachini, Khan, Frontini and Nicolosi (2018) and Brando, Frontini and Ganascia (2016).

[14] We can mention, among others: later Greek authors, as Pindar, Critias, Aristophanes, Old Comedy in general, Herodotus, Plato, Aristoteles, New comedy, Callimachus, Theocritus, Apollonius Rhodes, Epigrammatic poets (as Dioscorides, Meleager, etc.); among the Roman poets, authors as Lucillius, Horace, Catullus, Quintilian; Greek authors of Imperial period, as Ps.-Longinus, Emperor Hadrian, Dion of Prusa, Plutarch, Sextus Empiricus, Emperor Julian, Patristic authors (as Clement of Alexandria, Origen and Synesius); Menander Rhetor.

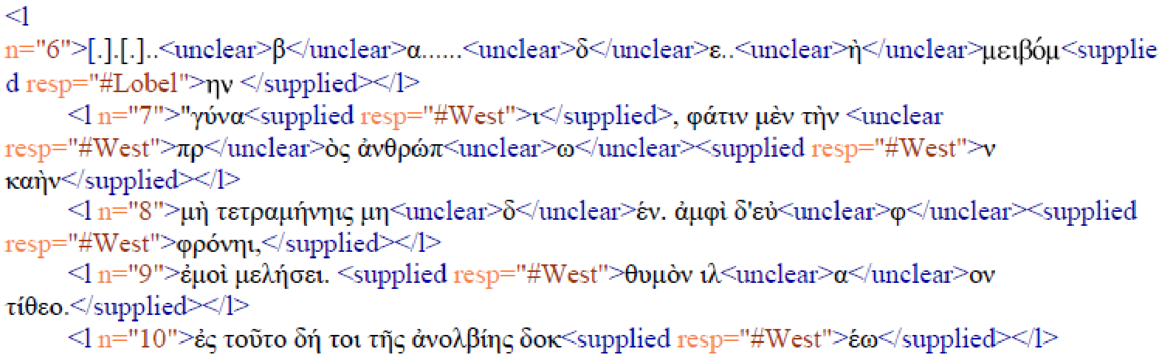
[15] The work starts from Nicolosi (2013) and Nicolosi (2017).

- new readings and interpretations;
- stronger certainty of corrections, integrations;
- authorship attribution;
- creation of a *corpus* of fragmentary ancient texts;
- searchable and interoperable data;
- enrichment of data with lexical datasets in LOD and other existing resources;
- a complete edition that is useful not only for scholars interested in Classical and Ancient Studies but also for non-specialists.

The online platform aims to become an interdisciplinary, multi-purpose research tool, enabling the final user to browse, explore and study digital editions by means of a user-friendly interface. Scholars and students will be able to easily access and search the complete corpus of the ancient author, accessing reliable critical and scientific information related to the field of study.

To reach these objectives and to set up the project correctly, we have tested the solutions to make all data searchable and easily interoperable. The first step is the TEI transcription and the annotation of the text. We have developed a case-study on a sample of TEI Archilochus' fragments with text, apparatus and witness, linguistic analysis, translations and commentary, with reference to Nicolosi (2013). This constitutes a pilot set of Archilochus' papyrus texts with metadata, bibliography, translation, and apparatus.

We are testing some functional solutions with the TEI encoding, both for managing the critical apparatus and witnesses and for representing textual phenomena of classical philology (*lacunae*, uncertain readings, omissions, corrections). The aim is to find a measured balance between the accuracy that philological study requires, and the synthesis provided by digital tools, with the purpose of making information accessible in a user-friendly, yet controlled and organized manner. It is necessary a more detailed annotation level that configures the text in its complexity and makes visible gaps, supplements, and doubtful readings. For this purpose we are using the TEI guidelines[16] (mostly the tagset of the module 11: Representation of Primary Sources. See Figure 1).[17]



```
<l
n="6">[.].[.]..<unclear>β</unclear>α......<unclear>δ</unclear>ε..<unclear>ὴ</unclear>μειβόμ<supplie
d resp="#Lobel">ην </supplied></l>
        <l n="7">"γύνα<supplied resp="#West">ι</supplied>, φάτιν μὲν τὴν <unclear
resp="#West">πρ</unclear>ὸς ἀνθρώπ<unclear>ω</unclear><supplied resp="#West">ν
καὴν</supplied></l>
        <l n="8">μὴ τετραμήνῃς μη<unclear>δ</unclear>έν. ἀμφὶ δ'εὺ<unclear>φ</unclear><supplied
resp="#West">φρόνηι,</supplied></l>
        <l n="9">ἐμοὶ μελήσει. <supplied resp="#West">θυμὸν ἰλ<unclear>α</unclear>ον
τίθεο.</supplied></l>
        <l n="10">ἐς τοῦτο δή τοι τῆς ἀνολβίης δοκ<supplied resp="#West">έω</supplied></l>
```

Figure 1

Starting from the module Critical Apparatus of the TEI guidelines, we are encoding three different annotation levels, divided along different typologies of witnesses and different text's references.

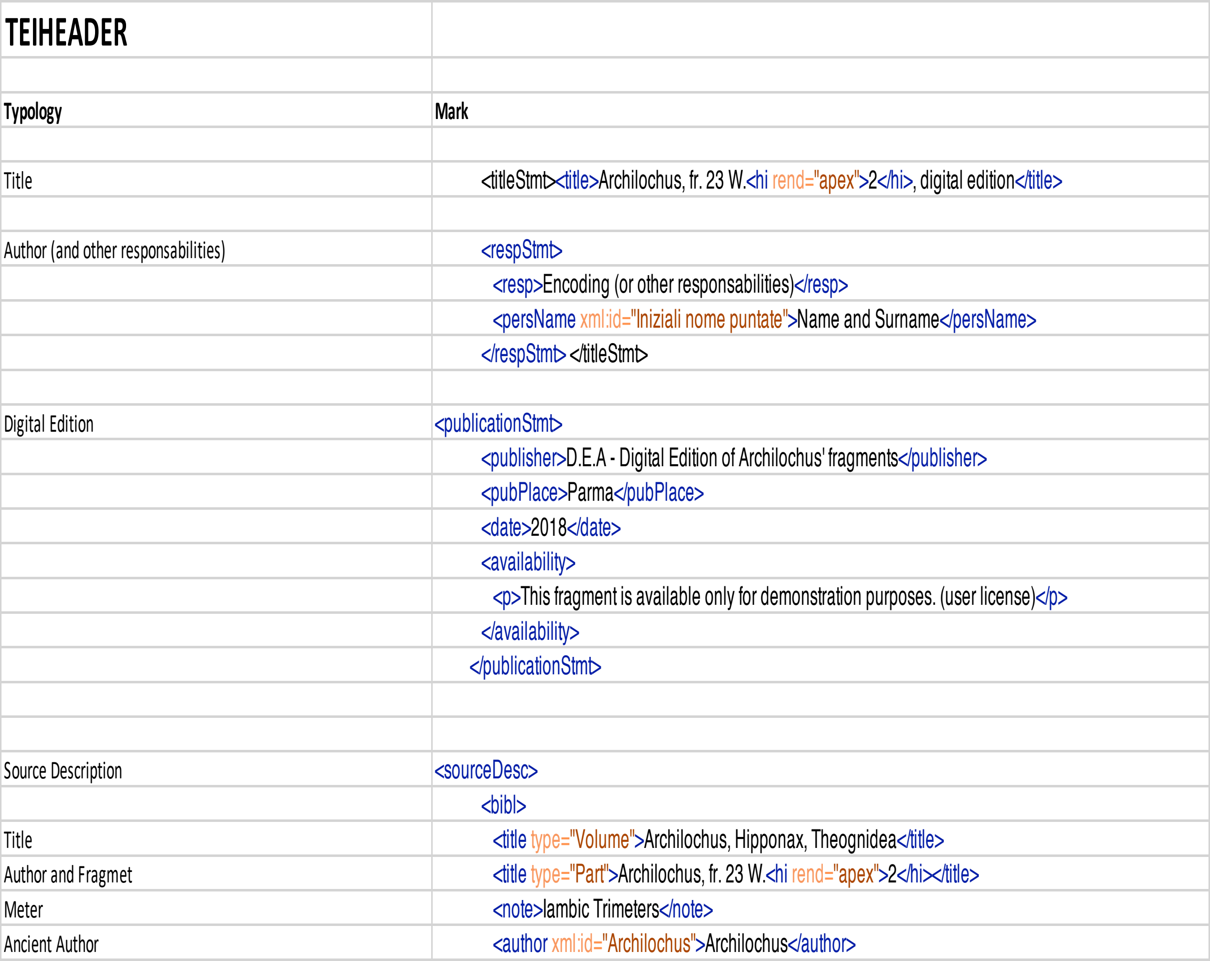
1. The tag <listWit> includes:

a) ancient witness, associated with an xml: id and some essential information (<msDesc>);
b) modern editions, each associated with an xml: id and the complete bibliographic reference (<bibl>).

[16] The current version of the guidelines is the P5. See https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.
[17] The text's reference for these exempla is Archil. fr. 23 W.^{2}, studied by Beatrice Nava during her BA level thesis in Classics, a.a. 2013/2014, Tutor Prof.ssa A. Nicolosi, at the University of Parma.

2. The tag <listBibl> includes bibliographic references, associated with an xml: id of:

a) secondary literature;
b) ancient texts (not indirect witnesses, but used for conjecture).



TEIHEADER	
Typology	**Mark**
Title	<titleStmt><title>Archilochus, fr. 23 W.<hi rend="apex">2</hi>, digital edition</title>
Author (and other responsabilities)	<respStmt>
	<resp>Encoding (or other responsabilities)</resp>
	<persName xml:id="Iniziali nome puntate">Name and Surname</persName>
	</respStmt> </titleStmt>
Digital Edition	<publicationStmt>
	<publisher>D.E.A - Digital Edition of Archilochus' fragments</publisher>
	<pubPlace>Parma</pubPlace>
	<date>2018</date>
	<availability>
	<p>This fragment is available only for demonstration purposes. (user license)</p>
	</availability>
	</publicationStmt>
Source Description	<sourceDesc>
	<bibl>
Title	<title type="Volume">Archilochus, Hipponax, Theognidea</title>
Author and Fragmet	<title type="Part">Archilochus, fr. 23 W.<hi rend="apex">2</hi></title>
Meter	<note>Iambic Trimeters</note>
Ancient Author	<author xml:id="Archilochus">Archilochus</author>

Figure 2

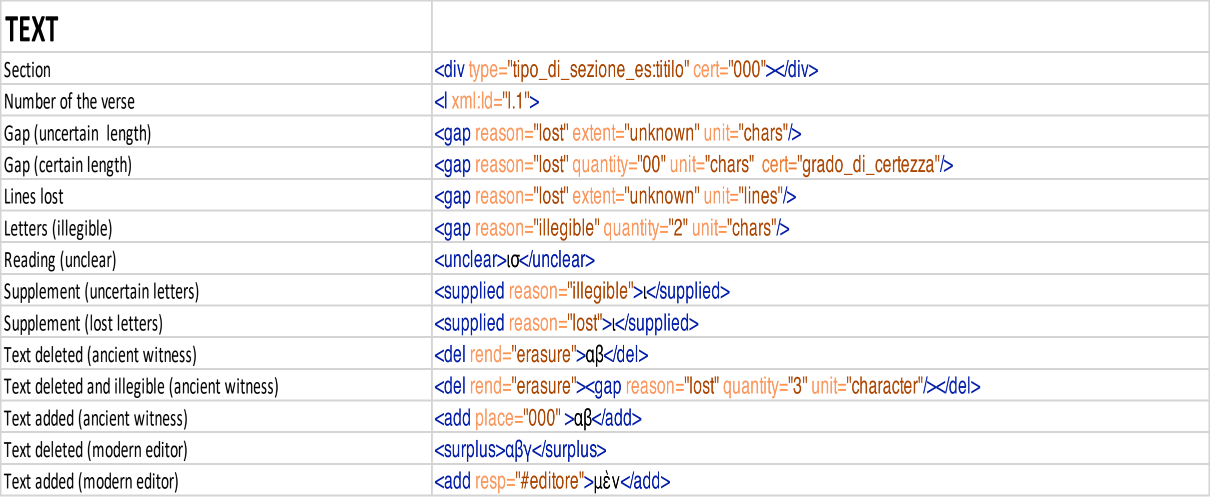
The coding of the apparatus is made with the parallel segmentation method. In the apparatus (<app>) we can differentiate between, for example:

a) Lesson of the reference text (e.g. Nicolosi), for example with the tag <lem>ἴλ<unclear>α</unclear>ον</lem>

b) Reading proposed in a modern edition (related to listWit), for example <rdg wit="#Latte_1955">ἴλ<unclear>ε</unclear>ον</rdg>

c) Reading from secondary literature (reference refers to listBibl), for example <rdg wit="#Bossi_1990">example</rdg>

d) Hypothesis of a modern editor supported by an ancient text, indicated with @source; here the reference refers to listBibl, for example <rdg wit="#Adrados_1990" source="#Exemplum">example</rdg>



TEXT	
Section	<div type="tipo_di_sezione_es:titilo" cert="000"></div>
Number of the verse	<l xml:ld="l.1">
Gap (uncertain length)	<gap reason="lost" extent="unknown" unit="chars"/>
Gap (certain length)	<gap reason="lost" quantity="00" unit="chars" cert="grado_di_certezza"/>
Lines lost	<gap reason="lost" extent="unknown" unit="lines"/>
Letters (illegible)	<gap reason="illegible" quantity="2" unit="chars"/>
Reading (unclear)	<unclear>ισ</unclear>
Supplement (uncertain letters)	<supplied reason="illegible">ι</supplied>
Supplement (lost letters)	<supplied reason="lost">ι</supplied>
Text deleted (ancient witness)	<del rend="erasure">αβ</del>
Text deleted and illegible (ancient witness)	<del rend="erasure"><gap reason="lost" quantity="3" unit="character"/></del>
Text added (ancient witness)	<add place="000" >αβ</add>
Text deleted (modern editor)	<surplus>αβγ</surplus>
Text added (modern editor)	<add resp="#editore">μὲν</add>

Figure 3

To standardize the model, we created a TEIheader that contain a structured detailed description of the content of ancient texts (Figures 2 and 3). In this model, that can be applied to both manuscript texts and papyrus (or epigraphy) texts, the metadata include general information such as bibliographical sources, manuscript or papyrus (or epigraphy) description, available open data (as imagines); in the <body>, we encode gaps, lacunae, different readings, corrections, symbols and then apparatuses with scholars' readings and hypotheses. The idea is to test, at this stage, the usability of what the TEI guidelines suggest. These guidelines proved to be suitable for our purposes, since they cover all our needs. Even if the coding is often time consuming, we think it would be important to keep using XML/TEI directly, without any TEI encoding tool. This allows us to have the control on the information we are encoding and, at the same time, to re-think our model, thus making it more and more accurate for our research goals. In this sense, we are not doing something new: we are not suggesting new TEI elements or new encoding models; we are simply facing some of the encoding problems as other projects.[18]

However, we are doing something that has never been done before. Indeed, by looking at the complete and up-to-date catalogue of digital edition compiled by Franzini[19] we find that, out of a total amount of 309 editions, there are only 18 scholarly editions based on antique sources and using an XML/TEI transcription. Five of them are a reproduction of a printed edition, but none of them concerns ancient Greek poetry and provides a critical apparatus. Moreover, only 3 of these editions provide a full text transcription, but, again, without a clear critical apparatus. Another aspect of novelty is the use of standards and best practices from the Semantic Web which ensures interoperability with resources available as Linked Open Data (LOD). We are able to augment our digital edition by linking the information contained in the fragments to the structured knowledge available in language resources published as LOD, thus allowing much more complete analyses.

It is the case of e.g. the Lexicon LSJ-LOD which allows us to deepen the linguistic analysis, or the Pleiades Gazetteer of ancient places which allows the enrichment of our edition with geospatial references. We have already linked some of the geographical references contained in the fragments to the Pleiades and obtained a nice improved version of the text itself, thus exploiting the potential of resources made available to scholars.

The next step foresees to develop a treebank: Archilochus' fragments can be collected in a *corpus* of syntactically annotated fragments to be used, as a gold standard to train systems aimed at performing automatic syntactic analysis of ancient Greek fragmentary texts. The treebank will be provided with textual analysis coupled with translations to be used in a didactic context to foster studying, teaching, and learning of ancient Greek language and literature.

[18] For example, regarding the coding of primary sources transcriptions we are using the same XML/TEI model used by other editions based on text of different periods, e.g. Medieval Nordic Text Archive (Menota), Jane Austen's Fiction Manuscripts Digital Edition, the DCLP project, mentioned above etc.

[19] The catalogue is available online at https://dig-ed-cat.acdh.oeaw.ac.at.

Finally, we can use the results obtained in this small but complex field of study to create a replicable model for other fields of literary studies. The investigation of technological aspects should hopefully act as an important test for future projects in the field of classical studies.

2.2 CLARIN-IT and DEA project: data and metadata

Research Infrastructures (RI) are a key element that can provide significant developments as they offer opportunities to store, develop and share data and tools; they are the perfect framework where to spread knowledge about good practices related to a discipline. CLARIN-ERIC (www.clarin.eu), the Common Language Resource Infrastructure for SSH, was born to promote the development of technological solutions aimed at making language resources available to scholars, researchers, students through a unified and standardized mode of access to data and computational tools. This involves making available digital repositories where data, tools and language services are catalogued, stored, retrieved and used in a simple and intuitive way by users. It represents the perfect solution to bring together producers and developers of language technology with its users.

The digitization of Archilochus' texts allows the creation of a philologically and critically controlled product and is aimed to develop crucial resources, materials, and tools for study and research in the field. The project represents one of the attempts within the context of CLARIN-IT[20] to contribute to the wider impact of CLARIN on the specific Italian community interested in Digital Classics. It may help to develop services aimed to foster new knowledge in SSH digital research,[21] and to sustain the existing one. The case study dealt with in CLARIN is meant to offer the opportunity to develop a new type of approach to the study of Classics, opening new horizons for the storage, analysis and study of data of the discipline.

As Nava (2019) points out: "DEA can be regarded as a case study in the framework of CLARIN-IT and its interests and specialization towards the Digital Classics. The ILC4CLARIN repository (https://ilc4clarin.ilc.cnr.it/) offers the corpus, along with other existing digitized resources for Ancient Greek provided as. LOD.[22] This allows us to enrich our corpus with LOD lexical datasets and to integrate our data with other existing resources, with the final aim of obtaining a complete edition that is useful not only for scholars interested in Classical and Ancient Studies but also for non-specialist users."

Moreover, as she said: "linguistic annotation allows the development of new teaching methods of Ancient Greek that are aimed at encouraging beginners and the creation of new resources to be integrated in frameworks, such as TüNDRA (https://weblicht.sfs.uni-tuebingen.de/Tundra/), specifically implemented for Classics. More generally, the annotations enable an interactive approach to texts that is more inviting and immediate for the students. […] In our case, having a corpus of fragmentary texts allows us to develop linguistic services for teaching (*e.g.* Hyper-Text Archilochus)".[23] To sum up, improving the currently existing tools for Ancient Greek with regards to their performance on fragmentary texts would offer very important upgrades for the study and teaching of Ancient Greek.

At this stage, we only tested some cases (encoding text and metadata) and we stored them in the CLARIN-IT repository. We profited from a Clarin Mobility Grant to advance our studies and our knowledge;[24] it was a clear example of how CLARIN can profitably contribute to training activities.[25] In addition, we can consider that an experience like the Tour de CLARIN (in Italy from 1st February to 31th March 2019, see http://www.ilc.cnr.it/it/content/tour-de-clarin-it) can spread knowledge and can be fundamental for the dissemination of research projects.

[20] See Monachini and Frontini (2016).

[21] See Monachini et al. (2018).

[22] Version of the TEI-dict Perseus Liddell-Scott Jones *Greek-English dictionary* (http://lari-datasets.ilc.cnr.it/ml/).

[23] It is only a preliminary study to test a sample of a prototype that provides the learner with a set of resources and tools that ease a critical assessment of ancient texts (http://hdl.handle.net/20.500.11752/OPEN-83).

[24] See CLARIN blog (https://www.clarin.eu/blog/tei-and-ancient-greek-fragmentary-poetry).

[25] A seminar was given also in Parma (https://www.clarin.eu/blog/clarin-it-presents-their-roadshow-seminars).

3 Conclusions: addressing a specific research challenge

The DEA project is currently under development and we hope that its implementation will foster the training of new professional figures specialized in a specific research field of the Digital Classics. The project is intended to deal with a well-defined number of fragments of a single author that will be investigated and made available in a complete and reliable way, but the final aim is to provide a replicable model for studies of wider interest related to Ancient Poetry, as Ancient Greek Song and/or Hellenistic Poetry.

DEA aims at bridging the digital gap between traditional studies and the growing world of data. It is focused on applying Digital Humanities methods – such as XML-TEI encoding or NLP technologies – to poetry analysis, classification, and publication, in order to ensure accessibility, interoperability, sharing, enrichment. These objectives are ensured by using semantic web technologies to link and publish literary datasets in a structured way in the Linked Data cloud, thus ensuring the interaction between texts and terminologies, ontologies, lexicons available as LOD.

The final aim of the project is to make available and easily accessible - for scholars, students, and common people - scientifically reliable texts of the ancient literary, that is crucial for our Cultural Heritage. This allows the development of new tools in the form of a collaborative workbench for the creation of digital edition which goes beyond the limits of paper editions. Moreover, the creation of a corpus and a treebank can help to deepen the study of ancient Greek structures and may have strong impact on the study and learning of the ancient Greek and the ancient literature in general, thus allowing to develop a portfolio of competences for users at different stages of learning. From a long-term perspective, collecting trees in a bank may pave the way to allow the development of appropriate tools towards the analysis of ancient Greek. The project is aimed to integrate the available digital resources, implementing and enriching what already exists and to develop crucial resources, materials and tools for study and research. All in all, we aim to develop a new type of approach to the study of Classics, opening new horizons for the study of the discipline. The project is expected to have a crucial impact in the field.

Research infrastructures seem to be the perfect place to make the results obtained concrete and visible. This newly created resource will be integrated into the existing CLARIN repository and provided with appropriate metadata. Finally, we can make ours the wishes of Nava (2019). It would be useful that the set of CLARIN-IT services were enriched with an integrated, specialized online platform aimed to support Digital and Classical Phylogists to proof-read, encode and enrich classical texts. In this respect, it would be helpful for CLARIN-IT to promote the development of tools tailored to researchers with no computational skills to help them in performing linguistic and textual annotations (morpho-syntactic, semantic, etc.) in a user-friendly and intuitive way.

References

Bridget Almas and Monica Berti. 2013. Perseids Collaborative Platform for Annotating Text Re-Uses of Fragmentary Authors. In *DH-Case 2013. Proceedings of the 1st International Workshop on Collaborative Annotations in a Shared Environment: metadata, vocabularies and techniques in the Digital Humanities*, 1-4. ACM New York, NY, USA (doi 10.1145/2517978.2517986)

Monica Berti. 2019. *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*. De Gruyter, Berlin-Boston (ISBN 9783110599572).

Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. 2016. REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets. *Complex Systems Informatics and Modeling Quarterly*, 0 (7), 60–80.

Monica Monachini and Francesca Frontini. 2016. CLARIN, l'infrastruttura Europea Delle Risorse Linguistiche per Le Scienze Umane e Sociali e Il Suo Network Italiano CLARIN-IT. *IJCoL - Italian Journal of Computational Linguistics, Special Issue on NLP and Digital Humanities*, 2 (2), 11–30.

Monica Monachini, Anas Fahad Khan, Francesca Frontini and Anika Nicolosi. 2018. Linked Open Data and the Enrichment of Digital Editions: The Contribution of CLARIN to the Digital Classics. In *Proceedings of the CLARIN Annual Conference 2018*, edited by Inguna Skadina and Maria Eskevich, 159-162. Pisa, Italy. https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf.

Monica Monachini, Anika Nicolosi et al. 2018. Digital Classics and CLARIN-IT: What Italian Scholars of Ancient Greek Expect from Digital Resources and Technology. In *Selected Papers from the CLARIN Annual Conference*

2017, Budapest, 18–20 September 2017, edited by M. Piasecki*,* 61–74. *Linköping University Electronic Press*, Linköpings universitet.

Beatrice Nava. 2019. *Tour de CLARIN: Interview with*, https://www.clarin.eu/blog/tour-de-clarin-interview-beatrice-nava

Anika Nicolosi. 2013. *Archiloco: elegie*. Pàtron Editore, Bologna, Italy (ISBN 9788855532365). Google-Books-ID: 9uj5oAEACAAJ.

Anika Nicolosi. 2015. *Analisi testuale e linguistica di Lirici Arcaici e Adespoti Giambici ed Elegiaci: Ipotesi di ricerca di applicazione della Filologia Computazionale al Greco Antico*. ILC-CNR, Pisa, Italy, November 6. http://www.ilc.cnr.it/sites/default/files/presentations/ILC-Thematic-Seminar_11.06.2015_Presentation.pdf

Anika Nicolosi. 2017. *Archiloco. Testimonianze e frammenti*. Aracne Editrice, Roma, Italy (ISBN 9788825508550).

Elena Pierazzo. 2014. *Digital Scholarly Editing: Theories, Models and Methods*. HAL Id: hal-01182162

CLARIN-supported Research on Modification Potential in Dutch First Language Acquisition

Jan Odijk
Utrecht University, The Netherlands
j.odijk@uu.nl

Abstract

This paper analyses data to address a specific linguistic problem, i.e. the acquisition of the modification potential of the three more or less synonymous Dutch degree modifiers *heel*, *erg* and *zeer*, all meaning 'very', which show syntactic differences in modification potential. It continues the research reported on in (Odijk, 2016). The analysis makes crucial use of linguistic applications developed in the CLARIN infrastructure, in particular the treebank search applications *PaQu* (Parse and Query) and *GrETEL* Version 4.00. The analysis benefits from the use of parsed corpora (treebanks) in combination with the search and analysis options offered by PaQu and GrETEL. Earlier work showed that despite little data for *zeer* modifying adpositional phrases adult speakers end up with a generalised modification potential for this word. In this paper, I extend the dataset considered, and find more (but still little) data for this phenomenon. However, I also find a similar amount of data that form counterexamples to the non-generalisation of the modification potential of *heel*. I argue that the examples with *heel* concern constructions with idiosyncratic semantics and therefore are not counted as evidence for the general rule of modification. I suggest a simple statistical analysis to account for the fact that children 'learn' that *heel* cannot modify verbs or adpositions though there is no explicit evidence for this and they are not explicitly taught so.

1 Introduction

In this paper I analyse data to address a specific linguistic problem, i.e. the acquisition of the modification potential of the three more or less synonymous Dutch degree modifiers *heel*, *erg* and *zeer*, all meaning 'very'. It continues the research reported on in (Odijk, 2016). The analysis makes crucial use of linguistic applications developed in the CLARIN infrastructure, in particular the treebank search applications *PaQu* (Parse and Query (Odijk et al., 2017)) and *GrETEL* Version 4.00 (Odijk et al., 2018), both of which make use of the Dutch syntactic parser Alpino (Bouma et al., 2001). The words that are being investigated are highly ambiguous. Most of the ambiguity is resolved by considering the syntactic context they occur in. Therefore, the analysis benefits from the use of parsed corpora (treebanks). Though the automatically created parses contain errors and require manual verification, the data analysis process is considerably speeded up and facilitated by these parses in combination with the search and analysis options offered by PaQu and GrETEL.

This paper is organised as follows: I introduce the linguistic problem in section 2. Section 3 introduces the treebank search applications used. Section 4 describes earlier work done on this type of problem and on the specific problem itself. This earlier work was carried out on relatively small corpora. Section 5 describes the complexity of first language acquisition and the simplifications and idealisations I assume to address the problem. In section 6 I describe which corpora I used in the research and report on the treebank query results found. Section 7 proposes considerations that may lead to an analysis of the problem. Section 8 summarises the main findings of this paper and suggests future research.

Jan Odijk 2020. CLARIN-supported Research on Modification Potential in Dutch First Language Acquisition. *Selected papers from the CLARIN Annual Conference 2019.* Linköping Electronic Conference Proceedings 172: 172 94–107.

2 The Problem

The three Dutch words *heel*, *erg* and *zeer* are (near-)synonyms meaning 'very', i.e. (stated informally) they modify a word or phrase that expresses a (gradable) property or state and specify that its modifiee has the property or state it expresses to a high degree. Of these, *heel* can modify adjectival (A) phrases only, while *erg* and *zeer* can modify not only adjectival, but also verbal (V) and adpositional (P) phrases. This is illustrated in example (1).[1]

(1) a. Hij is daar heel / erg / zeer blij over
he is there very / very / very glad about
'He is very happy about that'

b. Hij is daar *heel / erg / zeer in zijn sas mee
he is there very / very / very in his lock with
'He is very happy about that'

c. Dat verbaast mij *heel / erg / zeer
That surprises me very / very / very
'That surprises me very much'

In (1a) the adjectival phrase *blij* 'glad' can be modified by each of the three words. In (1b) the (idiomatic) adpositional phrase (PP) *in zijn sas* can be modified by *zeer* and *erg* but not by *heel*. The same holds in (1c) for the verbal phrase *verbaast*.[2] In English, the same holds for the word *very*: it can only modify adjectives.[3] For verbs and prepositional phrases one cannot use *very* but one can use the expression *very much* instead:

(2) a. He is very happy about it
b. He is *very / very much in love with her
c. It surprised me *very / very much

The distinctions illustrated in the preceding section are purely syntactic in nature. The words *heel*, *zeer* and *erg* are synonyms or near-synonyms, and the expressions *blij* and *in zijn sas* are near-synonyms as well, which makes it unlikely that the differences can be derived from semantic properties. It is also not in any way obvious how the differences could follow from universal principles of language or language acquisition.

There are other differences among the words *heel*, *erg* and *zeer*. If any of these differences is somehow related to the difference under investigation then it must be a difference in which *heel* opposes the other two words *erg* and *zeer*. However, this is not the case (Odijk, 2015).

3 The Treebank Search Applications PaQu and GrETEL 4.0

It is important to investigate the use of these words in their syntactic context, because they are (as many words in natural language) highly ambiguous. Odijk (2016) shows that *heel* is 6-fold ambiguous, *erg* is 4-fold ambiguous, and *zeer* is 3-fold ambiguous, but he also shows that the ambiguity is largely resolved by considering the syntactic context. For this reason I address the problem using the treebank search applications *Parse and Query* (*PaQu*) (Odijk et al., 2017) and *GrETEL* Version 4.00 (Odijk et al., 2018).

Both applications make existing manually verified treebanks for Dutch such as *LASSY-Small* for written Dutch (van Noord et al., 2013) and the *Spoken Dutch Corpus* (Oostdijk et al., 2002) available for search. They also enable a researcher to upload a text corpus and associated metadata, and have it automatically parsed by the Alpino parser (Bouma et al., 2001), after which the resulting treebank is made available for search.

[1] An asterisk is used to mark ill-formed expressions.

[2] or maybe the whole VP *verbaast mij*.

[3] and certain adverbs. I assume that words traditionally assigned the part of speech 'adverb' are either adjectives or (intransitive) adpositions.

The syntactic structures inside the treebanks are encoded in XML. Both applications offer XPath to search in these syntactic structures for words, grammatical properties and constructions. Each of them also offers additional search options: PaQu offers a very easy way to query for grammatical dependency relations between words, and GrETEL offers query by example facilities (Augustinus et al., 2012). Both treebank applications also offer various ways of analysing the search results, for data and metadata combined.

4 Earlier work

The type of problem dealt with here has, at least for English phenomena, figured prominently in the language acquisition literature (Baker, 1979; Berwick, 1985; Pinker, 1989; Yang, 2016), e.g. for accounting for the acquisition of adjectives that can be used predicatively but not attributively, and for accounting for dative constructions, in which some but not all verbs allow the double object construction in addition to the *to*-dative construction. This paper will not propose a general new solution to this problem, but has the more modest aim of analysing the relevant Dutch data for the problem at hand.

Odijk (2015) analyses the Dutch CHILDES corpora (MacWhinney, 2000) for the words *heel*, *erg* and *zeer*. These corpora contain transcriptions of adult-child interaction with (monthly) sessions recorded between the children's ages of approximately 1 year and 8 months and 6 years.[4] Since the children have to acquire the lexical properties of these words from the input provided by the adults (and other participants), this work focuses on the child-directed speech. The findings, together with findings in additional corpora, will be summarized in section 6.

5 First Language Acquisition

First language acquisition is extremely complex: the input is speech, which has to be turned into a sequence of phonetic symbols by the child while it has to build up the phone(me) inventory of the language it is acquiring at the same time. The speech is spontaneous, and therefore contains phenomena that are typical for spontaneous speech such as:[5]

Hesitations and filled pauses e.g. *en ehm (.) gaan we nog ehm (.)+* (*and hmm go we still hmm*)

Repetitions *een molen [/] molen* (*a mill mill*)

False starts and retracing <geef jij> [//] kom jij op mijn verjaardag ? (*give you come you on my birthday ?*)

Unfinished utterances (see example under hesitations)

Of course, the speech signal does not contain word boundaries, so the child has to find out somehow where the word boundaries are so that the input sequence of phone(me)s can be tokenized into a sequence of word tokens.

For the phenomenon under investigation here, the child must 'know' or find out that a categorisation of words into parts of speech is relevant, find out what the part of speech tags for its language are, and find out for each word what its part of speech tag is. In addition, each of the three words under investigation here is multiply ambiguous, and many of the candidate modifiees are ambiguous.

For these reasons only a few aspects of first language acquisition are considered here and various simplifications and idealisations are assumed. For example, the analysis starts from an orthographic transcription, enriched with annotations for hesitations, filled pauses, retracings, etc.[6] The ambiguity of the words cannot be avoided, but the focus here is on only one meaning of the words under investigation, viz. the meaning *very*.

[4] The version of the corpus in PaQu contains approximately 1.9 million tokens.

[5] The annotations in the examples are CHAT-annotations as used in CHILDES corpora.

[6] Though these annotations are not always correct and surely not complete in the actual CHILDES corpora.

It is also assumed that children 'know' or have somehow found out that they should be 'looking for' grammatical dependencies, e.g. head-complement relations, modifier-modifiee relations etc., and that they are able to do so (though it is not obvious how they achieve this).

Since this paper investigates modifier-modifiee relations, I specifically make a number of assumptions on *modification*. Modification has two aspects: a *syntactic* aspect, and a *semantic* aspect. Syntactic modification specifies the syntactic structure(s) in which modifiers and modifiees can occur. I assume that there is an operation of modification *M* that applies to two elements X and Y and creates a syntactic modification structure. I assume that it yields a single configuration, formulated here in terms of the structures assumed in the treebanks used here: X syntactically modifies Y by the operation M(X,Y) = [mod/X, hd/Y] (order irrelevant), i.e. a node X with grammatical relation *mod* (modifier) modifies a node Y with grammatical relations *hd* (head) if they are siblings of the same node.[7]

A syntactic modification relation has a semantic pendant. The study of the semantics of modification is of course a research field in itself. However, for the purposes of this paper minimal assumptions suffice: if X is a syntactic modifier of Y, the corresponding semantic modification is built up compositionally on the basis of the meaning of X ($[\![X]\!]$), the meaning of Y ($[\![Y]\!]$) and the meaning of the syntactic modification operation ($[\![M]\!]$), i.e. $[\![M]\!]([\![X]\!], [\![Y]\!])$. This assumption will play a crucial role in section 7.

I will also assume that children 'know' or find out that syntactic selection restrictions of a modifier on a modifiee are specified in terms of syntactic category. The notation *mod A*, *mod V*, and *mod N* specifies the property of a word or phrase that it can modify an A, V or N, resp.

6 Treebank Query Results

In this section, the main results for the queries for the three words *heel*, *erg* and *zeer* as modifiers are presented as reported by (Odijk, 2015), as well as for the results of these queries in the Basilex corpus and the Lassy-Large Wikipedia part.

The corpora in which the queries have been carried out are characterised in Table 1.

Corpus	**#utts (k)**	**#tokens (m)**	**modality**	**spontaneity**	**formality**
LASSY-Small	65	1	written	prepared	formal
CGN	130	1	spoken	mixed	mixed
VanKampenJAC	61	0.3	spoken	spontaneous	informal
VanKampenCHI	47	0.15	spoken	spontaneous	informal
CHILDES Dutch	545	1.9	spoken	spontaneous	informal
Basilex		13.5	written	prepared	formal
Wikipedia	8707	145	written	prepared	very formal

Table 1: Corpora analysed in this study and their characteristics.

Each corpus has been characterised in terms of its size, i.e. its number of utterances (where available) and its number of tokens, its modality (written language or (transcripts of) spoken language), the spontaneity of its content and an indication of its formality. LASSY-Small is a treebank for written Dutch of app. one million tokens. Its written text is explicitly prepared and rather formal (e.g. it does not contain social media and usenet data). The Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) treebank contains app. one million tokens for spoken Dutch. It has several subcomponents, differing in spontaneity and formality (e.g., it contains prepared read speeches but also spontaneous conversations). The Van Kampen corpus is one of the corpora in the CHILDES collection for Dutch. The child-directed utterances (VanKampenJAC) were investigated separately from the utterance of the target children (VanKampenCHI). The Van Kampen corpus contains transcriptions of the natural interaction between parents and children. I also investigated the whole CHILDES collection for Dutch.

The BasiLex corpus consists of 13.5 million tokens[8] of texts written for children in primary education.

[7] In order to properly work on the flat structures in the treebank which allow more than 2 siblings, the formulation should be generalised somewhat, but this is not essential for the purposes of this paper.

[8] 11.5 million if interpunction symbols are ignored.

It contains various genres, with 40% of the tokens coming from educative materials, 40% from child literature, and 20% from media (newsfeeds, subtitles, etc.). The time coverage is 1976-2013. At the time of these investigations, it was not possible yet to host a full treebank for the Basilex corpus. For that reason, a subcorpus was created by selecting all sentences containing any word form of the lemmas *heel*, *erg* or *zeer* and the resulting corpus was uploaded in PaQu. It is known there as the corpus HEZ-Basilex-JO, shared with everyone who is logged in.[9] This subcorpus contains 26,239 sentences. PaQu parsed these sentences using Alpino. Several of the sentences come from educational material, which often contains words in alternative spellings (e.g. *ver-schrik-ke-lijk* instead of *verschrikkelijk* 'horrible', *SLAAAP* instead of *slaap* 'sleep'), and exercises with incomplete words or lists of alternative words. In such cases, the automatic parses are often wrong:

(3) Zijn kinderen hebben 'm erg gemi...
his children have him very mis...
'His children mis... him very much'

(4) erg blij / bijt / bij
very glad / bite / by
'very glad / bite / by'

Finally, the Wikipedia part of Lassy-Large contains 145 million utterances of carefully prepared written language and it is, as an encyclopedia, very formal in nature. It is part of the (550 million token) SoNaR Corpus. Querying the whole treebank for the SoNaR Corpus with the treebank search applications is, despite the development of special techniques to speed up querying (Vandeghinste and Augustinus, 2014; Vanroy et al., 2017), unfortunately not yet possible for us, though the Institute for the Dutch Language recently made a version of GrETEL 4.0 available in which each of SoNaR's components can be searched separately.[10]

6.1 Mapping D-COI Part of Speech Tags

The treebanks consulted use the *de facto* standard for part of speech tagging of Modern Dutch words, so-called D-COI tags (Van Eynde, 2005). This tag set makes distinctions that differ somewhat from what is needed here. I describe here how I reclassified D-COI tags to the distinctions I want to make: Some tags map directly on tags I use, e.g. *adj* = adjective maps to *A*, *ww* = verb maps to *V*, *n* = noun maps to *N*, *vz* = adposition maps to *P*. For other tags the mappings are slightly more complex:

- *vnw* = pronoun. The pronominal nature of a word is an important morpho-syntactic distinction, but in my view it is independent of part of speech assignment. Words with the D-COI tag *vnw* were automatically mapped to *A* (e.g. for *veel* 'many', *weinig* 'few' and their comparative and superlative forms) or to *N* (e.g. for *wat* 'a few').

- *bw* = adverb. Words with D-COI tag *bw* are manually mapped to *A* or *P*, depending on the word.

- *mwu* = multiword unit. The characterisation of a word combination as a multiword unit is an important distinction, but in my view it is independent of part of speech assignment. Word combinations labeled with the D-COI tag *mwu* were manually mapped to *A*, *N*, *V* or *P* depending on the specific word combination.

- *tw* = numeral. Words with the D-COI tag *tw* are mapped to *N*.

The queries search for the words *heel*, *erg* or *zeer* when occurring as a modifier (grammatical relation *mod*). I specifically also searched for sentences that contain *heel*, *erg* or *zeer* and a predicative (*predc*) or locative (*ld*) complement, because such sentences are likely to contain incorrectly analysed examples of modification of adpositions. I also searched for uses of these words with a different grammatical relation: these should be irrelevant if the parse is correct but might contain misparsed examples.

[9] Everybody can log in by just using one's e-mail address.

[10] `https://portal.clarin.inl.nl/chn-gretel/ng/home`.

Corpus / m tokens	mod A	mod V	mod P
LASSY-Small	295.6	0.0	0.0
CGN	2899.4	0.0	7.9
VanKampenJAC	2191.1	3.3	3.3
VanKampenCHI	1616.9	0.0	6.5
CHILDES Dutch	2512.4	3.2	8.5
Basilex	172.0	0.0	1.7
Wikipedia	90.5	0.0	0.3

Table 2: Results of queries for *heel* as a modifier in a variety of corpora (relative frequency per million tokens).

6.2 Main Results for *heel*

Table 2 summarises the query results for *heel* as a modifier.

Some remarks on these figures are required. I will discuss some cases where *heel* appears to modify a verb (6.2.1) or an adposition (6.2.2).

6.2.1 *heel* Modifying Verbs

First, I discuss some special cases of *heel* modifying a verb. In the query results for Lassy-Small one does find the part of speech code for verb in the treebanks ('ww') as being modified by *heel*, but these are artifacts of the structure of the treebank, in which adjectives derived from participial verbs are categorised as verbs, as in (5):

(5) Examples of adjectives derived from participles, which are categorised as *ww* (verb) in the treebank:

a. heel gecompliceerd ('very complicated')

b. heel overtuigend ('very convincing')

c. heel vervelend ('very boring/unpleasant')

Second, under the substantivised use of infinitives the word is also characterised as *ww*, though it has actually converted to a noun. The modifier *heel* only has the interpretations it has as a modifier of a noun ('whole') in such constructions. See (6):

(6) Examples of substantivised verbs that are categorised as *ww* (verb) in the treebank:

a. het hele ... gebeuren ('the whole ... happening')

b. hun hele hebben en houden
their whole have and hold
'all their possessions'

In the Spoken Dutch Corpus, there are some examples of *heel* modifying verbs, but they are ill-formed for me, and are used almost exclusively by Flemish speakers in informal registers. I found similar examples in the SoNaR corpus. I suspect that people who use this can use *heel* in the sense of *geheel* 'completely' (and this is how I glossed them in (7)). This surely requires further investigation, but I will not deal with these examples here. Some examples:

(7) *heel* modifying verbs by Flemish speakers in informal registers:

a. ...heel te verdwalen... ('to get completely lost')

b. ...heel omgebouwd... ('completely rebuilt')

In VanKampenJAC also one example occurs (session laura030.cha, speaker JAC):

(8) Ik kijk heel uit
I look very out
'I am very cautious'

The example is ill-formed for me, and in this particular case I could check the example with the speaker. She confirmed that the sentence is ill-formed for her too, and that it must have been a performance or transcription error. Such an example, and several other examples, do show that ill-formed input is offered to children, who must thus be robust against such ill-formed input.

In the Dutch CHILDES as a whole more examples of *heel* modifying a verb occur, but these are all utterances by children, who apparently did not get the rules yet. Researching them is outside the scope of this paper.

6.2.2 *heel* Modifying Adpositions

There are also some cases where *heel* modifies or appears to modify an adposition. First, there are some examples in which *heel* modifies an adverbial PP:

(9) a. heel in de verte
very in the far-th
'at a very great distance'

b. heel in het begin
very in the beginning
'in the very beginning'

c. heel af en toe
very off and to
'very infrequently'

d. heel in de verte
very in the distance
'at a very great distance'

I found ten different cases (the four of (9) and *heel in het algemeen* lit. very in the general ('very generally'), *heel in het bijzonder* lit. very in the particular 'more particularly', *heel in het kort*, lit. very in the short 'very briefly', *heel op het laatst* lit. very at the last 'at the very end', *heel uit de verte* lit. very from the far-th 'from a very great distance', and *heel aan het eind* lit. very at the end 'at the very end'.) Such examples were found in most corpora (CGN, VanKampenJAC, CHILDES Dutch, Basilex and Wikipedia).

Furthermore, I found one additional example:

(10) ...'t heel voor de hand ligt...
...it very before the hand lies...
'...it is very obvious...'

Though the present participle form of this expression *voor de hand liggen**d*** is adjectival in nature and is often modified by *heel*, modification of the verbal form is ill-formed according to my judgement as a native speaker. I will assume it is a performance error.

Finally, I found one example (by a Flemish speaker) where *heel* modifies an adposition and where it probably means 'completely': *heel beneden* 'completely downstairs'.

In CHILDES, there are several examples of *heel* modifying an adposition in the children's speech but also one by an adult (which is ill-formed, according to my judgement as a native speaker):

(11) heel iets naar buiten BOU mat20501.429 (father)
very somewhat to outside
'a little bit to the outside (?)'

again showing that the language acquisition device must be robust against ill-formed input.

6.3 Main Results for *erg*

Table 3 shows the treebank query results for modification by *erg*.

There is one example in the children's speech in VanKampenCHI where *erg* appears to modify an adposition, but no other peculiarities.

Corpus / m tokens	mod A	mod V	mod P
LASSY-Small	156.0	13.7	5.5
CGN	324.6	78.1	13.2
VanKampenJAC	112.5	49.6	0.0
VanKampenCHI	77.6	6.5	6.5
CHILDES Dutch	189.7	44.1	2.1
Basilex	324.5	73.8	3.5
Wikipedia	128.3	12.2	2.1

Table 3: Results of queries for *erg* as a modifier in a variety of corpora (relative frequency per million tokens).

6.4 Main Results for *zeer*

Table 4 shows the treebank query results for modification by *zeer*.

Corpus / m tokens	mod A	mod V	mod P
LASSY-Small	307.4	7.3	2.7
CGN	207.0	7.9	1.8
VanKampenJAC	6.6	6.6	0.0
VanKampenCHI	6.5	0.0	0.0
CHILDES Dutch	6.4	2.7	1.6
Basilex	26.7	1.7	0.3
Wikipedia	342.0	18.3	1.9

Table 4: Results of queries for *zeer* as a modifier in a variety of corpora (relative frequency per million tokens).

There are a few examples that might involve modification of an adposition by *zeer*, but they might also involve modification of the verb or the whole verb phrase. It concerns modification of the expression *op prijs stellen* lit. *on price put* 'appreciate'[11] and of the expression *in de smaak vallen* lit. *in the taste fall* 'like (with arguments reversed)'[12], all by adults. They are analysed in the treebank as modifying the verb and that is surely defensible and actually most likely the correct analysis.

6.5 Summary of the Query Results

The results for all corpora except Basilex and Wikipedia were already reported in (Odijk, 2015) and (Odijk, 2016). His findings for the child-directed speech in these corpora can be summarised as follows:

- Of the three words *heel*, *erg* and *zeer*, *heel* occurs most frequently.
- There is an overwhelming number of cases where *heel* modifies an adjectival phrase (>92%).
- Modification of verbal phrases by *heel* does not occur.
- There are many examples where *erg* modifies an adjectival phrase, but also a significant number of cases where it modifies a verb phrase.
- There are very few examples of *zeer* modifying an adjectival phrase, and also very few in which it modifies a verb.
- There are no clear examples with *erg* or *zeer* modifying a PP.

For the problem under investigation, this means:

[11] Utterances jos20021.354 and tom20507.71 from the Groningen Corpus.

[12] Utterance iri30323.1283 from the Groningen corpus.

Corpus	***heel***	***erg***	***zeer***
Basilex	1.7	3.5	0.3
Wikipedia	0.3	2.1	1.9

Table 5: Relative frequency per million tokens of *heel*, *erg* and *zeer* modifying an adposition in Basilex and Wikipedia.

- The data seem appropriate for acquiring the property that *heel* modifies adjectival but no verbal phrases.
- It is less clear how modification of PPs can be excluded, since there are some examples where *heel* modifies PPs.
- The absence of data for *zeer* makes it difficult to state anything about the acquisition of its modification potential.

In order to address the latter two problems, more data are needed. Unfortunately, there are no other CHILDES data for Dutch that are relevant in this context. However, Odijk (2016) observes that *heel* occurs very early in the children's speech (1;11), with *erg* occurring only a year later (2;10), and *zeer* very late (4;8). He ascribes the late occurrence of *zeer* to its more formal character. A corpus of data typical for the input that children hear or read from the age of 5 years old would be ideal to address these problems. The BasiLex corpus (Tellings et al., 2014) is exactly such a corpus: it contains texts that are directed at children at primary school. In addition, it is significantly larger than the CHILDES corpora. Because of the late acquisition of *zeer*, BasiLex's focus on texts that are targeted at children between the ages of 6 and 12 appears to make it particularly appropriate for investigating the modification potential of *zeer*.

I used PaQu to investigate the properties of modifiees of *heel*, *erg* and *zeer*, respectively. A manual analysis of the query results was carried out in order to map the more refined distinctions made by PaQu onto the distinctions needed here, and to correct wrong parses by Alpino.

The crucial data are presented together in Table 5. Strikingly, examples with *heel* modifying a PP are more frequent (1.7 / million tokens) than *zeer* modifying a PP (0.3 / million tokens) in Basilex, but this does not have the effect that the adult grammar allows modification of PPs by *heel* in general. Conversely, despite their low frequency even in this larger corpus, the adult grammar allows modification of PPs by *zeer* generally. In addition, the frequency of *zeer* modifying PPs is so low, that one might wonder whether they are taken into account at all in the acquisition process. After all, utterances may be analysed incorrectly by the child, or might be misheard, or might be mispronounced by the speaker, so it seems reasonable to require a minimum number of occurrences of a phenomenon before it is taken into account in adapting lexical properties or grammar rules, at least in the case of unconscious acquisition, as is the case here. In any case, the language acquisition procedure must be robust against some noise ((Yang, 2016, 13) and references there).

Concluding, despite a larger and more representative corpus, the same questions still lie before us:

- Why does the presence of PPs modified by *heel* not lead to generalising the modification potential of *heel* to PPs generally?
- Why is the modification potential of *zeer* generalised to PPs generally despite its very low frequency?

Perhaps also the Basilex corpus is not big enough to get a representative overview. Therefore, an even larger corpus, the Wikipedia part of Lassy-Large (145 million tokens), has been investigated. Though this corpus is not representative for language acquisition at all, it might give us insight into the degree of representativity of the CHILDES corpora and the BASILEX corpora for the problem at hand.

It is clear that *zeer* occurs much more often here than in the earlier corpora as a modifier of adjectival, verbal and adpositional phrases. However, even here, in this large and very formal corpus, the frequency of *zeer* modifying a PP is extremely low (1.9 per million). Examples of *heel* modifying a PP are less frequent in this corpus, but I ascribe this to the rather formal nature of this corpus. I conclude that even this very large and very formal corpus does not provide an answer to the major questions that the data raise.

7 Towards Analysis of the Data

In this section, a tentative attempt to analyse the data is presented. I will first discuss the possibility of analysing these data using Yang's theory on the *Sufficiency Principle* in section 7.1. I argue that this theory does not contribute to explaining these data. In section 7.2 I argue that the combinations of *heel* modifying adpositions are idiosyncratic in nature and cannot provide evidence for a productive rule of modification. Finally, in section 7.3 I sketch my proposal for the acquisition of these constructions.

7.1 The Sufficiency Principle

Since the modifier *zeer* has properties based on very little data, it is natural to investigate whether it has these properties not from direct positive evidence but from a productive rule that applies to it. It seems to me that there is no productive rule in Dutch that determines the modification potential of degree modifiers, so if this assessment of the facts is correct, it is unlikely that any theory of productivity of rules will contribute to addressing this problem. If there would be a productive rule, it should be a rule that predicts that degree modifiers can modify adpositions if it is to account for the fact that *zeer* has the potential to modify adpositions despite very little positive evidence for this.

One approach that addresses the issue of the productivity of rules has been proposed by (Yang, 2016). He considers (inter alia) the exclusively predicative use of *A*-adjectives (e.g. *asleep*, *awake*, *alone*, *away*) and dative alternation in English. He claims that the relevant words belong to a class, and that the members in this class generalise their modification or complementation potential in accordance with what he calls the *Sufficiency Principle* (Yang, 2016, 177):[13]

(12) Let *R* be a generalisation over *N* items, of which *M* items are attested to follow *R*. *R* can be extended to all *N* items if and only if: $N - M < \theta_N$ where $\theta_N = N/\ln N$.

Applying this hypothesis to the problem of this paper requires first of all establishing a class that the degree modifiers belong to. They certainly do not have morphological properties in common, but one might consider them as members of the semantically defined class of degree modifiers. Yang defines the class of verbs to account for dative alternation phenomena also semantically (as 'verbs of caused possession that involve the transfer of objects, entities or abstract information' (Yang, 2016, 201)). Second, a rule that applies to this class must be postulated. Actually, the relevant rule should predict that *mod P* is a property of degree modifiers. However, if this rule is productive, it will be impossible to have exceptions to this rule that do not have *mod P* as a property (such as *heel*) if negative evidence plays no role in first language acquisition (as is generally assumed): we basically then have an instance of Baker's Paradox here (Baker, 1979). I inventoried around 145 words and expressions that can act as degree modifiers.[14] The largest subclass, members of which can have the property *mod A* | *mod V* | *mod P*, contains minimally 35 and maximally 83 elements.[15] Even if all 83 belong to this class, applying the *Sufficiency Principle* (12) yields the following result: N = 145, and, for the postulated rule, M = 83: $145 - 83 = 62$. This should be smaller than θ_{145} but it is larger than θ_{145}, where $\theta_{145} = 145/ln145 \approx 29$. I conclude that even the best candidate rule under these assumptions is predicted by the *Sufficiency Principle* not to be productive. I conclude that the *Sufficiency Principle* cannot account for the relevant facts.[16]

[13] Yang actually has 'if and only **iff**', which I assume is a typo; He also has := instead of the equal sign, of which I also assume it is a typo.

[14] I did so by crucially using a different application developed in the context of CLARIN: Cornetto, which offers a search interface to the Dutch WordNet (Vossen et al., 2013).

[15] I was not yet able to determine the property *mod P* for all these words: my intuitive judgements on theses examples are uncertain and I was not yet able to do corpus searches for all these words. Fortunately, this is not crucial here, as will become clear below.

[16] But of course, this does not mean that the *Sufficiency Principle* is wrong.

It might be investigated what predictions the *Sufficiency Principle* makes *during* first language acquisition, to model various stages of the language acquisition process, but I will leave that to future research.

7.2 The Idiosyncratic Nature of *heel* Modifying Adpositions

In this section, I argue that constructions in which *heel* modifies adpositions are idiosyncratic constructions that must be acquired one by one and that do not constitute evidence for a productive rule such as the modification rule.

The first consideration in this regard comes from a closer look at the PPs modified by *heel*, e.g. *heel in de verte*. I stated that *heel* syntactically modifies the adpositional phrase (PP) *in de verte*. However, this PP expresses a location, and locations cannot be semantically modified by degree modifiers such as *heel*, *erg* and *zeer*. This is clear from the examples in (13) and (14):

(13) hij staat (*erg) op het veld
he stands very on the field
'He is standing (*very much) on the field'

(14) a. Zij zit in de put
She sits in the well
'She is sitting in the well / She is depressed'

b. Zij zit erg in de put
She sits very in the well
'*She is sitting very much in the well / She is very depressed'

Modifying the location expressed by the PP *op het veld* 'on the field' by the degree modifier *erg* leads to ill-formedness (13). The phrase *in de put zitten* in (14) is ambiguous between a literal interpretation (with the PP *in de put* as a location 'in the well') and an idiomatic interpretation (in which *in de put* expresses a mental state 'depressed'). Modifying the PP by a degree modifier disambiguates the PP, which then only has the mental state interpretation.

If degree modifiers cannot semantically modify locations, then what does *heel* modify semantically in *heel in de verte*? A look at the gloss and the translation makes this clear. *Heel* in this expression semantically modifies the adjective *ver* which is part of a derived noun (*ver-te* i.e. *far-th*) inside a noun phrase contained in the PP syntactically modified by *heel*, cf. the translation *at a **very great** distance*. This meaning cannot arise from the normal rule of modification with compositional semantics (see section 5), which requires that the meaning of the full expression is derived from the meaning of *heel* and the meaning of the whole PP *in de verte*. I thus conclude that these constructions cannot be seen as special instances of the normal rule of modification. In fact, the semantic modification of the morpheme *ver-* by *heel* cannot be part of any productive linguistic rule, and the expression must thus be stored as an instance of an idiosyncratic mapping between form and meaning. This is confirmed by the fact that only a handful of different examples of this construction were found (in quite large corpora) and by the fact that no or only very limited variation is possible. For example, in example (9d), one cannot replace the noun by semantically related nouns (15):

(15) a. * heel in de nabijheid
very in the closeness
'at a very small distance'

b. * heel in de hoogte
very in the height
'at a very great height'

c. * heel in de diepte
very in the depth
'at a very great depth'

and for *de verte* only the prepositions *in* 'in' and *uit* 'out of' are possible, but e.g. *naar* 'towards' is not (16):

(16) * heel naar de verte
very to the distance
'towards the far distance'

Similar restrictions hold for the other examples. This thus disqualifies these examples as instances of regular modification of a PP by *heel*.

This analysis, however, does not apply to the expression *heel af en toe*. Here the part *af en toe* has both an unusual syntax (coordination of two adpositions) and an idiosyncratic meaning ('occasionally') but *heel* modifies the part *af en toe* as a whole, in accordance with the rule for modification. I still argue that the combination of *heel* and *af en toe* is idiosyncratic.[17] I observe that *af en toe* can**not** be modified by *erg* and *zeer*.[18] In addition, the expression *nu en dan* lit. now and then 'occasionally', which is a synonym or near-synonym of *af en toe*, cannot be modified by any of the words *heel*, *erg* or *zeer*.[19] I therefore conclude that the combination *heel af en toe* is also an idiosyncratic combination and not an instance of the regular modification rule.

7.3 Towards an Analysis

If the cases where *heel* modifies adpositions are idiosyncratic and do not provide evidence for general rules or principles, a simple statistical learning strategy can account for the data. I make several assumptions: (1) that the modification potential of words is acquired by positive evidence only; (2) that each property of a lexical item has an associated activation score, which increases each time there is evidence in the input for this property; (3) that the activation score must be higher than a threshold θ_{min}. This is necessary to be robust against ill-formed, misheard or mis-analysed input. It then follows that *heel* selects A, and only A (no positive evidence for mod V or mod P), while *erg* and *zeer* select not only A but also V and P.

The question remains what the value of θ_{min} is or how it is determined. This is a matter that has to be determined empirically by studying multiple cases. No firm conclusions can be drawn on the basis of the phenomenon studied here alone. It is possible that assuming a decay function, which lowers the activation score over time, might play a role here too.[20] Here I speculate that θ_{min} must be very low to account for *zeer* selecting P ($<$ 0.3 / million tokens), and it might also be a function of the number of relevant examples encountered, so that its value actually increases over time if there is sufficient input. Future research will have to clarify whether these speculations correspond to the facts.

8 Concluding Remarks and Future Work

This paper analysed data to address a specific linguistic problem, i.e. the acquisition of the modification potential of the three more or less synonymous Dutch degree modifiers *heel*, *erg* and *zeer*, all meaning 'very'. The analysis makes crucial use of linguistic applications developed in the CLARIN infrastructure, in particular treebank query applications. The use of treebanks was necessary because of the high ambiguity of the words. In addition, the use of the CLARIN applications made it possible to base the analysis provided in this paper on a far larger empirical base than would have been possible without these applications, and the applications enable the researcher to query the data efficiently.

[17] It was also suggested to me that *af en toe* might actually be an adjective instead of an adposition. It is not so easy to determine what category *af en toe* belongs to. Many standard tests are inconclusive. However, the so-called PP-over-V test (Broekhuis, 2013, 8) suggests that *af en toe* is adpositional, cf. the well-formedness of e.g. *je zou het bijna vergeten af en toe*, lit. one would it almost forget occasionally, 'one would occasionally almost forget it', with *af en toe* to the right of the infinitive *vergeten*. Whatever category it is, a different category assignment would not account for the idiosyncracies observed in the main text.

[18] A search in the 550 million token SoNaR Corpus, which contains 32,119 occurrences of *af en toe* yields exactly one result of *erg* modifying *af en toe*.

[19] The 550 million token SoNaR corpus contains 3,909 occurrences of *nu en dan*, and there are no occurrences of modification by *heel*, *erg* or *zeer*. There are 5 occurrences of the combination *zo heel* modifying *nu en dan*, but *zo* is obligatorily present in these constructions. I assume that *zo heel* is also an idiosyncratic combination.

[20] Such a decay function might provide an account of the phenomenon of language attrition.

References

Liesbeth Augustinus, Vincent Vandeghinste, and Frank Van Eynde. 2012. Example-based treebank querying. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

C.L. Baker. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry*, 10(4):533–581.

Robert Berwick. 1985. *The Acquisition of Syntactic Knowledge*. MIT Press, Cambridge, MA.

Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers*, 37(1):45–59.

Hans Broekhuis. 2013. *Syntax of Dutch: Adpositions and Adpositional Phrases*. Amsterdam University Press. http://www.oapen.org/record/462289.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3 edition.

Jan Odijk, Gertjan van Noord, Peter Kleiweg, and Erik Tjong Kim Sang. 2017. The parse and query (PaQu) application. In Jan Odijk and Arjan van Hessen, editors, *CLARIN in the Low Countries*, chapter 23, pages 281–297. Ubiquity, London, UK. DOI: http://dx.doi.org/10.5334/bbi.23. License: CC-BY 4.0.

Jan Odijk, Martijn van der Klis, and Sheean Spoel. 2018. Extensions to the GrETEL treebank query application. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 46–55, Prague, Czech Republic, January 23-24. http://aclweb.org/anthology/W/W17/W17-7608.pdf.

Jan Odijk. 2015. Linguistic research with PaQu. *Computational Linguistics in the Netherlands Journal*, 5:3–14, December.

Jan Odijk. 2016. A Use case for Linguistic Research on Dutch with CLARIN. In Koenraad De Smedt, editor, *Selected Papers from the CLARIN Annual Conference 2015, October 14-16, 2015, Wroclaw, Poland*, number 123 in Linköping Electronic Conference Proceedings, pages 45–61, Linköping, Sweden. CLARIN, Linköping University Electronic Press. http://www.ep.liu.se/ecp/article.asp?issue=123&article=004, http://dspace.library.uu.nl/handle/1874/339492.

Nelleke Oostdijk, Wim Goedertier, Frank Van Eynde, Lou Boves, Jean-Pierre Martens, Michael Moortgat, and Harald Baayen. 2002. Experiences from the Spoken Dutch Corpus project. In *Proceedings of the third International Conference on Language Resources and Evaluation (LREC-2002)*. ELRA.

Steven Pinker. 1989. *Learnability and Cognition*. MIT Press, Cambridge, MA.

Agnes Tellings, Micha Hulsbosch, Anne Vermeer, and Antal van den Bosch. 2014. BasiLex: an 11.5 million words corpus of Dutch texts written for children. *Computational Linguistics in the Netherlands Journal*, 4:191–208, 12/2014.

Frank Van Eynde. 2005. Part of speech tagging en lemmatisering van het D-COI corpus. CGN report, Centrum voor Computerlinguïstiek, KU Leuven, Leuven, Belgium, July. http://www.ccl.kuleuven.be/Papers/DCOIpos.pdf.

Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 147–164. Springer Berlin Heidelberg.

Vincent Vandeghinste and Liesbeth Augustinus. 2014. Making large treebanks searchable. The SoNaR case. In *Proceedings of the LREC 2014 2nd workshop on Challenges in the Management of Large Corpora (CMLC-2)*, pages 15–20, Reykjavik. http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-CMLC2%20Proceedings-rev2.pdf.

Bram Vanroy, Vincent Vandeghinste, and Liesbeth Augustinus. 2017. Querying large treebanks: Benchmarking GrETEL indexing. *Computational Linguistics in the Netherlands Journal*, 7:145–166, 12/2017.

Piek Vossen, Isa Maks, Roxanne Segers, Hennie van der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke. 2013. Cornetto: a lexical semantic database for Dutch. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch, Results by the STEVIN-programme*, Theory and Applications of Natural Language Processing, chapter 10, pages 165–184. Springer, Berlin Heidelberg.

Charles Yang. 2016. *The Price of Productivity: How Children Learn to Break the Rules of Language*. MIT Press, Cambridge, Mass.

Word Usage in German Texts on Women's Suffrage around 1900. Corpus Building, Lexical Documentation and the CLARIN-D Infrastructure

Anna Pfundt
Justus Liebig University
Giessen, Germany
anna.pfundt@germanistik.uni-giessen.de

Melanie Grumt Suárez
Eberhard Karls University
Tübingen, Germany
melanie.grumt-suarez@uni-tuebingen.de

Thomas Gloning
Justus Liebig University
Giessen, Germany
thomas.gloning@germanistik.uni-giessen.de

Abstract

The paper presents core aspects of a project that examines word usage in the German controversy on women's suffrage around 1900. The investigation is based on a variety of written texts (including journal articles, books, and pamphlets) that began to condense in the 1880s and developed into a complex thematic network until the introduction of women's suffrage in 1918. The focus of this paper is the presentation of two infrastructure aspects. First, the corpus compilation (ongoing and already published to some extent) is accomplished in the CLARIN-D infrastructure component German Text Archive ("Deutsches Textarchiv", hereafter DTA). This project exemplifies a basic user need: to make new corpus texts available from the very beginning of a project which can then be analyzed for the project's purposes with the powerful search tool architecture of the DTA. This kind of strategy is a win-win situation for the author, for the infrastructure, and for the whole research community. Second, the dissertation and the text corpus will be accompanied by a lexical documentation on word usage in the discourse domain. This component relies heavily on standards and best practices adopted in the CLARIN-D community and in the ZHistLex-project. The results are to be published and hosted by one of the German CLARIN centers. This project is a use case with solutions that can be seen as generic for similar kinds of projects, be it dissertations, term papers or larger research projects.

1 The First Women's Movement in Germany and its discourse cosmos[1]

The so-called First Women's Movement in Germany[2] started around the middle of the 19th century and ended in the late 1920s. It was not a unified movement but rather a complex and dynamic configuration of groups with different positions and different strategies. They shared, however, the aim to improve the situation of women, to enhance the spectrum of their political rights and the possibility to participate in public life, to study at universities, to earn a living on their own, to choose specific occupations. There were other topics and aims, e.g. the professional situation of prostitutes, sexual morals and the situation of working class women, which were the main topics of specific groups. The improvement of the situation of working class women, e.g., was one of the aims of the proletarian women's movement with Clara Zetkin (1857-1933) as a leading figure. Here, the fight for women's rights is deeply embedded in the overall struggle for the rights of the working class. The bourgeois wing of the First Women's Movement

[1] By "discourse cosmos" we mean the complex and dynamically evolving ensemble of communicative contributions that belong to a discourse topic and its sub-topics which often are organized in specific "threads".

[2] See Frevert (1995); Gerhard (1990); Gerhard (2009); Hervé (1988); Rosenbusch (1998); Schaser (2006); Twellmann (1972).

Anna Pfundt, Melanie Grumt Suárez and Thomas Gloning 2020. Word Usage in German Texts on Women's Suffrage around 1900. Corpus Building, Lexical Documentation and the CLARIN-D Infrastructure. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 108–118.

in Germany had its leading figures as well, e.g. Helene Lange (1848-1930) or Anita Augspurg (1857-1943). From 1865 onwards, these groups began to become organized, e.g. in the "Allgemeiner Deutscher Frauenverein" and in other bodies of a more local or general nature. All these developments were part of a wider debate that started with the French Revolution and its ideas of equality including the equality of women and men. The German debates were related to debates in other languages and in other countries, both in respect of central ideas and in respect of textual interlinking.

The First Women's Movement brought about a dynamic discourse cosmos, the structure of which can be analyzed along main topics, discourse threads, actors, and a specific configuration of media and text types. Among the first topics to establish specific discourse threads were the issues of equal or at least improved education for women and the right to work for a living. Only later, by the end of the 19th century, the topics of political participation and specifically the right to vote brought about new discourse threads. The protagonists used a wide range of media, text types and oral genres for their communicative purposes, e.g. petitions to parliaments, position papers in support of petitions, printed pamphlets, journal articles, personal letters, open letters, speeches, conference discussions, and others. One of the most important linguistic research tasks is to show, how all these communicative genres and the linguistic means used in these debates were used in specific ways to support the relevant political aims.

Apart from argumentative strategies, word usage is one of the most important aspects of language use in the contributions to the discourse cosmos at hand. Like in modern discourse threads, e.g. those on abortion, nuclear energy or immigration, an important part of word usage in the discourse cosmos of the First Women's Movement is closely connected to specific positions, to communicative tasks, and to specific strategies. Word formation, metaphor, the use of topic-specific vocabulary or semantic innovation are but a few of the aspects that belong to the research tasks for the project that we will now describe.

2 Word usage in the German controversies over women's suffrage 1870-1918: A research project and its infrastructure support

The subject of Anna Pfundt's dissertation project[3] is the role of word usage in the discussions about women's suffrage from the 1870s until 1918. This discourse was taking place in a vast spectrum of texts (including journal articles, books, and controversy texts) that began to condense in the 1880s and then developed a complex discourse network until the introduction of women's suffrage in Germany in 1918. The discussion of this specific topic is an important part of a broad discourse on a widespread spectrum of controversial points brought about by the First Women's Movement in Germany, especially the aspect of political participation. It is an important aspect of this research task to include quite different positions that were promoted in the fight for women's suffrage by women groups of different convictions and strata, e.g. the champions of the proletarian party vs. different shades of bourgeois positions.

Word usage plays a central role in the constitution of points of view, in the formulation of views and their justification, as has often been the case in discourses about alternative and competitive word usages (see Pfundt, 2017; Gloning, 2012). The use of words is characterised by specific thematic profiles, but also by the controversial nature of the object and by aspects of historical development over several decades. In addition to the specific uses ("senses") of words and phrases, forms of word formation, metaphors, ad hoc uses, foreign-language expressions and functionally oriented vocabulary are also part of the study. Another important aspect for the project is to examine the connection of the use of words specific to different "camps" and their perspectives and opinions.

On the one hand, the research strategy is methodologically based on the work of the Düsseldorf School of discourse studies around Dietrich Busse, Georg Stötzel and Martin Wengeler (e.g. "Kontroverse Begriffe", ed. by Stötzel and Wengeler, 1995). In line with this methodology, the research is characterised by a discourse historiography in which the thematic developments of a discourse are reconstructed in narrative form in connection with the description of the lexical means used. On the other hand, the project of Anna Pfundt goes beyond this methodology in two respects: First, within the project the corpus texts are made available publicly (an ongoing task), second, word usage will be documented in a specific lexicographic component, which will be closely connected to the dissertation and to the corpus texts by way of structural or digital interlinking.

[3] See Pfundt (2017).

Here is an example for this kind of interlinking of dissertation, text corpus and lexical documentation: In the investigation, the narrative presentation of the discussion and the use of words anchored in it is based on a corpus of German texts on early women's suffrage, the texts of which are successively fed into the German Text Archive (DTA). By way of illustration we present a short example from an early text[4] by Hedwig Dohm (1831-1919), one of the earliest champions of women's suffrage in Germany, where she comments on the assumption of different spheres of men and women in a very critical way:

> O über dieses Geschwätz von der **Sphäre** des Weibes, den Millionen Frauen gegenüber, die auf Feld und Wiese, in Fabriken, auf den Straßen und in Bergwerken, hinter Ladentischen und in Bureaus im Schweiße ihres Angesichts ihr Brot erwerben.
>
> Wenn die Männer vom weiblichen Geschlecht sprechen, so haben sie dabei nur eine ganz bestimmte Klasse von Frauen im Sinn: Die Dame. Wie nach dem bekannten Ausspruch jenes bekannten österreichischen Edelmannes der Mensch erst bei dem Baron anfängt, so fängt bei den Männern das weibliche Geschlecht erst da an, wo es Toilette und Conversation macht und Hang zu Liebesintriguen und Theaterlogen verräth.
>
> Geht auf die Felder und in die Fabriken und predigt eure **Sphärentheorie** den Weibern, die die Mistgabel führen und denen, deren Rücken sich gekrümmt hat unter der Wucht centnerschwerer Lasten! (Dohm, 1876, p. 126-127)

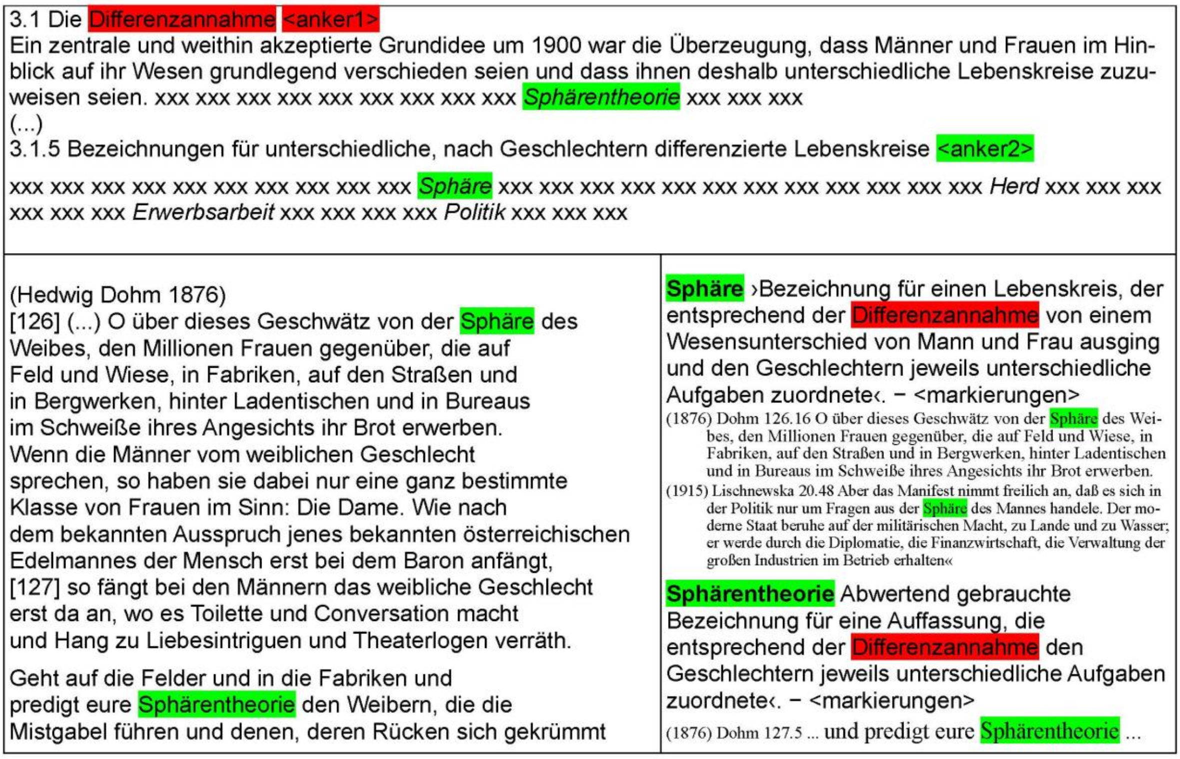
In this passage, Hedwig Dohm harshly criticizes the assumption of different spheres of men and women, an assumption which was common in 19th century thinking and which nowadays is called "difference hypothesis" (Differenzannahme, Differenzhypothese).

This passage can illustrate three integrated components of the dissertation. First, the whole text of Hedwig Dohm is now available in CLARIN-D's German Text Archive (DTA) at the Berlin-Brandenburg Academy CLARIN center in a digital format. As such it is not only available for download and Open Access in different formats, it can also be used with the powerful search facilities that are provided by the DTA. Second, this passage will be commented on and analyzed in the dissertation, especially in the chapter on the "difference hypothesis" and its role in the debates on women's suffrage around 1900. Third, the component "Lexical documentation" will contain dictionary-like entries with lemmata like "Sphäre", "Sphärentheorie", "Sphärenanbeter", where the peculiar aspects of the usage of these words in their discourse contexts are analyzed and documented. We call this documentation "dictionary-like", because the organization follows the patterns of article structuring that is well-known from traditional dictionary making. There are, however, components that go well beyond traditional dictionary making, e.g. in providing an underlying "discourse ontology" by a specific system of markers, which allows to search for lexical items that are related e.g. to the "difference hypothesis". This third component, the extensive discourse dictionary, currently contains almost 1100 entries (see section 4, below).

The three components (dissertation/investigation, digital text corpus, digital discourse dictionary) and their interrelations are shown schematically[5] in the following figure (Wolff/Geyken/Gloning, 2015 fig. 6). The field in the upper part of the figure represents a (future) passage of the investigation component on the "difference hypothesis", below on the left is the relevant passage from one of the corpus texts (there will be others), below on the right is an example that represents the component "lexical documentation". The colors indicate structural connections that will be digitally implemented in a later stage of the project.

[4] `http://www.deutschestextarchiv.de/book/show/dohm_frauenfrage_1876?p=134` (last access 2020-05-03).

[5] The figure is not a suggestion for screen design but rather a visualization of the structural connections between investigation, dictionary and corpus texts.



3.1 Die Differenzannahme <anker1>
Ein zentrale und weithin akzeptierte Grundidee um 1900 war die Überzeugung, dass Männer und Frauen im Hinblick auf ihr Wesen grundlegend verschieden seien und dass ihnen deshalb unterschiedliche Lebenskreise zuzuweisen seien. xxx xxx xxx xxx xxx xxx xxx xxx xxx *Sphärentheorie* xxx xxx xxx
(...)
3.1.5 Bezeichnungen für unterschiedliche, nach Geschlechtern differenzierte Lebenskreise <anker2>

xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx *Sphäre* xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx *Herd* xxx xxx xxx xxx xxx xxx *Erwerbsarbeit* xxx xxx xxx xxx *Politik* xxx xxx xxx

(Hedwig Dohm 1876)
[126] (...) O über dieses Geschwätz von der Sphäre des
Weibes, den Millionen Frauen gegenüber, die auf
Feld und Wiese, in Fabriken, auf den Straßen und
in Bergwerken, hinter Ladentischen und in Bureaus
im Schweiße ihres Angesichts ihr Brot erwerben.
Wenn die Männer vom weiblichen Geschlecht
sprechen, so haben sie dabei nur eine ganz bestimmte
Klasse von Frauen im Sinn: Die Dame. Wie nach
dem bekannten Ausspruch jenes bekannten österreichischen
Edelmannes der Mensch erst bei dem Baron anfängt,
[127] so fängt bei den Männern das weibliche Geschlecht
erst da an, wo es Toilette und Conversation macht
und Hang zu Liebesintriguen und Theaterlogen verräth.

Geht auf die Felder und in die Fabriken und
predigt eure Sphärentheorie den Weibern, die die
Mistgabel führen und denen, deren Rücken sich gekrümmt

Sphäre ›Bezeichnung für einen Lebenskreis, der entsprechend der Differenzannahme von einem Wesensunterschied von Mann und Frau ausging und den Geschlechtern jeweils unterschiedliche Aufgaben zuordnete‹. − <markierungen>

(1876) Dohm 126.16 O über dieses Geschwätz von der Sphäre des Weibes, den Millionen Frauen gegenüber, die auf Feld und Wiese, in Fabriken, auf den Straßen und in Bergwerken, hinter Ladentischen und in Bureaus im Schweiße ihres Angesichts ihr Brot erwerben.

(1915) Lischnewska 20.48 Aber das Manifest nimmt freilich an, daß es sich in der Politik nur um Fragen aus der Sphäre des Mannes handele. Der moderne Staat beruhe auf der militärischen Macht, zu Lande und zu Wasser; er werde durch die Diplomatie, die Finanzwirtschaft, die Verwaltung der großen Industrien im Betrieb erhalten«

Sphärentheorie Abwertend gebrauchte Bezeichnung für eine Auffassung, die entsprechend der Differenzannahme den Geschlechtern jeweils unterschiedliche Aufgaben zuordnete‹. − <markierungen>

(1876) Dohm 127.5 ... und predigt eure Sphärentheorie ...

Figure 1: Interrelations between investigation, text corpus and discourse dictionary.

With regard to the text corpus component, a win-win situation arises for the author of the dissertation and for the CLARIN-D infrastructure component DTA. Every new contribution increases the textual basis, which can be analyzed with the powerful search tools[6] of the DTA. For the DTA, each of these texts is a thematic enrichment and – with regard to texts written by women – also a contribution to increasing the proportion of female authors. This is what we will now outline in more detail in the following section 3. In the subsequent chapter 4 we will expand on the infrastructure aspect of the component "lexical documentation".

3 Infrastructure I: The TdeF-Corpus in CLARIN's German Text Archive

The "Deutsches Textarchiv"/German Text Archive, henceforth DTA, is one of the most important infrastructure components within CLARIN-D.[7] It is hosted and managed by the Berlin-Brandenburg Academy of Sciences and Humanities and basically serves all research communities that work with texts from the 17th to the 19th centuries. These communities include German studies, especially literary studies, historical linguistics and lexicography, the study of intellectual history, discourse history, the history of terminologies to name but a few. The core corpus is balanced according to groups of text types and to subject fields over the decades. The organization of the core corpus is systematically documented on the DTA website.[8]

With respect to specific user needs and projects, the DTA allows the integration of new texts and text collections in a section with additions: "Erweiterungen des Deutschen Textarchivs" (DTAE)/Additions to the German Text Archive. Users can submit their texts together with relevant metadata which have to conform to the criteria of DTA's basic format of text encoding and meta data organization. These textual additions are ascribed to their producers which is an important point in respect of the metrics of academic success. But more importantly, these additional texts can be used as sub-corpora by using a

[6] `http://www.deutschestextarchiv.de/doku/DDC-suche_hilfe` (last access 2020-05-03). `http://odo.dwds.de/~moocow/software/ddc/querydoc.html` (last access 2020-05-03).

[7] `http://www.deutschestextarchiv.de/` (last access 2020-05-03).

[8] `http://www.deutschestextarchiv.de/doku/textquellen#dta-kernkorpus` (last access 2020-05-03).

specific element of the query syntax of the DTA. This option is currently used for building up a special corpus with German texts from the First Women's Movement.

This sub-corpus has the title "Texte der ersten Frauenbewegung"/German texts of the First Women's Movement, hereafter TdeF.[9] At present, the focus lies on texts that deal with the topic of women's suffrage, since this is the subject of Anna Pfundt's dissertation. It is planned, however, to subsequently enrich the textual substance over the years with contributions on other topics as well, e.g. education of women, access to university studies or the situation of working class women. Hence, the current TdeF corpus has a double function: it serves as an empirical basis for the investigations on word usage in Anna Pfundt's research project, but it is also the nucleus for a fully-fledged corpus of First Women's Movement texts in German.

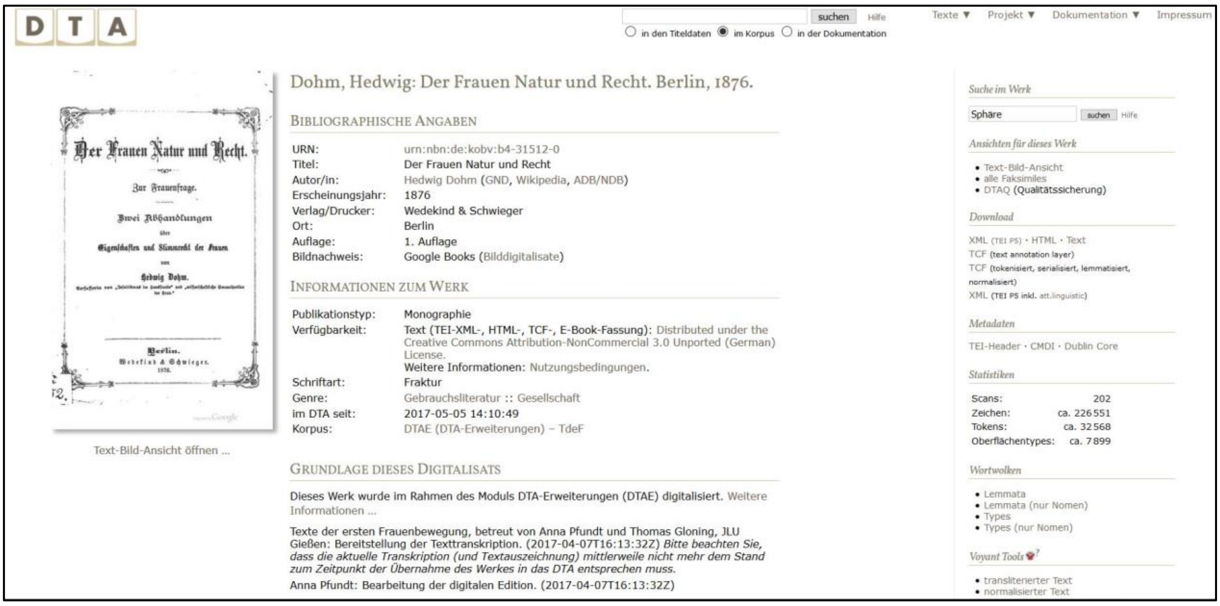
At present, it includes 70 texts with 2481 pages, that are already published. This may not constitute a huge corpus yet, but it *is* a valuable contribution of texts written by women and dealing with highly important topics from the debates on the emancipation of women. The corpus is focused on the period 1870-1919, since the first texts on the issue of women's suffrage appeared around 1870 and women were granted the right to vote in 1918. The central media of the time were books, magazines and newspapers; therefore, the corpus texts are independent works in the form of books, brochures or lectures as well as magazine articles from the press organs of the First Women's Movement and newspaper articles. The spectrum of authors is manifold: opponents and supporters of the women's movement, individual persons and representatives of organizations, women and men (e.g. August Bebel) of different ages, some authors remain anonymous.

In order to create and prepare the texts as searchable full texts in a sub-corpus and to make them publicly accessible, they are digitized according to the specifications of the DTA base format.[10] The DTA base format has become a quasi-standard for the encoding of texts from the 17th century onwards. For lexical analysis, the texts in the sub-corpus TdeF can be addressed as a separate text group `for` queries. In the first step, the texts are either produced with ABBYY Recognition Server which is able to deal with blackletter typefaces, or they are keyed in manually depending on the quality and complexity of the images. Still in other cases transcriptions on the web can be used as a starting point. In all cases meticulous proofreading is necessary. In the second step, the texts are structurally annotated according to the DTA principles stated in the base format. As the third step, the relevant metadata are provided, equally following the DTA guidelines.

The following figure shows the start screen with Hedwig Dohm's text "Der Frauen Natur und Recht" (1876; 'On the nature and the rights of women'). From here, different user options are possible, ranging from download in different formats, online reading in different modes or using the search facilities and the visualization tools of the DTA environment.

[9] `http://www.deutschestextarchiv.de/doku/textquellen#tdef` (last access 2020-05-03).

[10] `http://www.deutschestextarchiv.de/doku/basisformat` (last access 2020-05-03). The DTA base format is now widely used within the community. The use of oXygen for work on historical texts is supported by specific DTA tutorials and documentation.



D T A
suchen
Hilfe
○ in den Titeldaten ● im Korpus ○ in der Dokumentation
Texte ▼
Projekt ▼
Dokumentation ▼
Impressum
Der Frauen Natur und Recht.
Zur Frauenfrage.
Zwei Abhandlungen
über
Eigenſchaften und Stimmrecht der Frauen
von
Hedwig Dohm.
Berlin.
Wedekind & Schwieger.
1876.
Text-Bild-Ansicht öffnen ...

Dohm, Hedwig: Der Frauen Natur und Recht. Berlin, 1876.

Bibliographische Angaben

URN:	urn:nbn:de:kobv:b4-31512-0
Titel:	Der Frauen Natur und Recht
Autor/in:	Hedwig Dohm (GND, Wikipedia, ADB/NDB)
Erscheinungsjahr:	1876
Verlag/Drucker:	Wedekind & Schwieger
Ort:	Berlin
Auflage:	1. Auflage
Bildnachweis:	Google Books (Bilddigitalisate)

Informationen zum Werk

Publikationstyp:	Monographie
Verfügbarkeit:	Text (TEI-XML-, HTML-, TCF-, E-Book-Fassung): Distributed under the Creative Commons Attribution-NonCommercial 3.0 Unported (German) License. Weitere Informationen: Nutzungsbedingungen.
Schriftart:	Fraktur
Genre:	Gebrauchsliteratur :: Gesellschaft
im DTA seit:	2017-05-05 14:10:49
Korpus:	DTAE (DTA-Erweiterungen) – TdeF

Grundlage dieses Digitalisats

Dieses Werk wurde im Rahmen des Moduls DTA-Erweiterungen (DTAE) digitalisiert. Weitere Informationen ...

Texte der ersten Frauenbewegung, betreut von Anna Pfundt und Thomas Gloning, JLU Gießen: Bereitstellung der Texttranskription. (2017-04-07T16:13:32Z) *Bitte beachten Sie, dass die aktuelle Transkription (und Textauszeichnung) mittlerweile nicht mehr dem Stand zum Zeitpunkt der Übernahme des Werkes in das DTA entsprechen muss.*

Anna Pfundt: Bearbeitung der digitalen Edition. (2017-04-07T16:13:32Z)

Suche im Werk

Sphäre

Ansichten für dieses Werk

- Text-Bild-Ansicht
- alle Faksimiles
- DTAQ (Qualitätssicherung)

Download

XML (TEI P5) · HTML · Text
TCF (text annotation layer)
TCF (tokenisiert, serialisiert, lemmatisiert, normalisiert)
XML (TEI P5 inkl. att.linguistic)

Metadaten

TEI-Header · CMDI · Dublin Core

Statistiken

Scans:	202
Zeichen:	ca. 226551
Tokens:	ca. 32568
Oberflächentypes:	ca. 7899

Wortwolken

- Lemmata
- Lemmata (nur Nomen)
- Types
- Types (nur Nomen)

Voyant Tools

- transliterierter Text
- normalisierter Text

Figure 2: Hedwig Dohm: *Der Frauen Natur und Recht* (1876). DTAE start screen.
`http://www.deutschestextarchiv.de/book/show/dohm_frauenfrage_1876`

Once the texts are available in the DTA they can be addressed as a specific sub-corpus with a broad range of query options. The options that are most relevant for our research project on word usage include: the search for specific word-forms, the search for lemmatized words (which will provide all the relevant word-forms), the search for word-formations and their elements (e.g. word-formations with “Sphäre”), the search for derivational morphemes and the word-formations they produce, the search for different parts of speech. These options can be combined, e.g. adjectives formed with “-lich”, they can also be combined with metadata, e.g. “Adjectives in the TdeF sub-corpus before 1900”. An important feature is the opportunity to extract quotations with their metadata (author, work title, date), which are important components of the articles in the lexical documentation system.

As we have mentioned previously: providing source texts from a research project in an ongoing way for the DTA and thus building up a steadily growing project-specific sub-corpus *within* the DTA leads to a win/win/win-constellation for at least three parties: firstly, for the individual researcher and his or her project, who can take advantage of the powerful search functionalities of the DTA’s query machine; secondly, for the DTA, because it steadily increases its holdings and gets further texts from specific areas (e.g. discourse topics, author groups, text types); thirdly, other researchers with different research interest can take advantage of the newly available texts as well, e.g. a person studying word formation in German adjectives or someone doing research on the evolution of text types, independent research questions that are not related to the First Women’s Movement in Germany.

4 Infrastructure II: Lexical Documentation and CLARIN-D/ZHistLex

The digital lexical documentation system fulfils three specific functions within the overall structure of the project: single word documentation; providing an underlying discourse ontology across different dimensions; to allow for interlinking between the text of the investigation and the lexical documentation system.

Firstly, the lexical documentation system is supposed to document words and the usages of words that are relevant for the discourse topic under study (women’s suffrage). As mentioned before, words like “Sphäre” oder “Sphärentheorie” are used in specific ways and for specific purposes in the debates. For this purpose, the traditional form of representation is the word article with its structured positions (e.g. lemma, explanation of meaning, quotations) as we know it from dictionaries. Hence, the structural backbone of the lexical documentation system consists of articles which document the use of words and their specific contribution to the discourse on women’s suffrage. “To document” in this respect means to explain and to characterize, how the words are used, and to give quotations from the spectrum of texts. For this task, the TEI Lex-0[11] annotation scheme is used. It is a standard which is currently adopted both in CLARIN-D contexts and in the ZHistLex[12] project which was funded by Germany's Federal Ministry of Education and Research (BMBF).[13] The article components and information positions include: lemma, definitions or other kinds of semantic explanation, comments on specific functions in the discourse, references to and comparisons with the information in the standard historical dictionaries of German, if applicable information on word history and/or borrowing, cross-references, finally a block with quotations together with their metadata. If there are different usages (senses), these components are repeated for each sense. The arrangement of the articles is strictly alphabetic, but it is clear that electronic access in a digital system will include other forms of access as well.

Secondly, the lexical documentation makes use of an underlying discourse ontology across different dimensions. The basic idea here is, that historical vocabularies are organized along different criteria and that they build up multi-dimensional structures that evolve over time (cf. Gloning, 2003). A word like "Sphärenanbeter" ('person that worships the idea that men and women have different spheres, tasks, and rights in life'), e.g. may be characterized in respect of its part of speech, its pattern of word formation, its communicative functions (referring to a type of person; expressing a negative/critical assessment), its place in the system of ideas of the First Women's Movement (here the so-called *Differenzhypothese*) etc. If these aspects are marked up in a systematic and consistent way, the documentation allows for combinatorial search, e.g. the search for words with which persons are negatively evaluated, or the search for compounds of the type N+N that have something to do with the difference hypothesis. For this task, the parts of the ontology that are specific to the discourse topic at hand, must be the result of the analysis of the texts. There are no preexistent discourse ontologies.

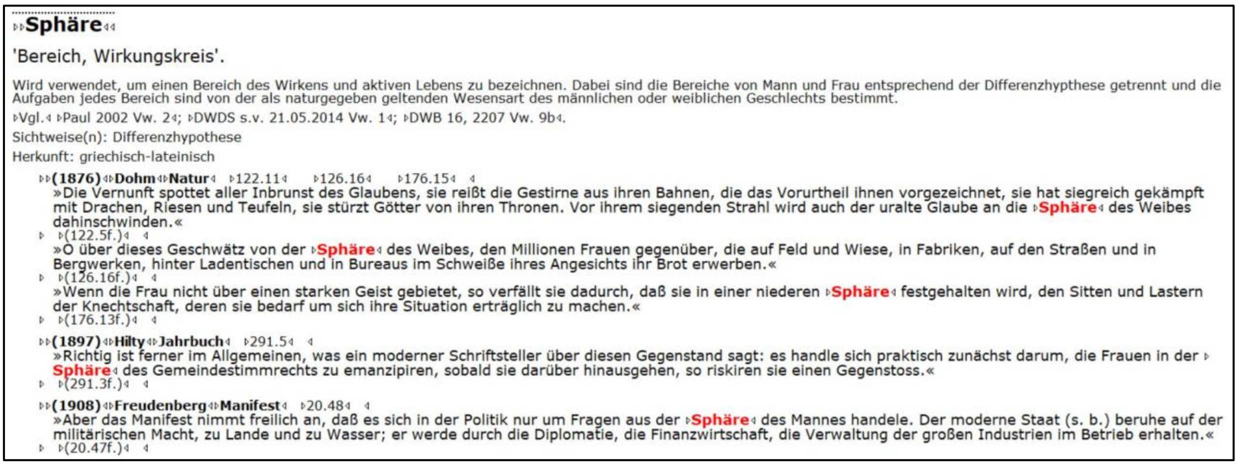
Thirdly, the components of the lexical documentation system are interrelated to specific passages of the investigation and its sub-chapters. Since it is the main objective of the research project to investigate different aspects of word usage in all its complexities, there will be specific chapters on these aspects of their relations, e.g. the role of word formation, the use of metaphor, the role of loanwords, the use of creative language use, the contribution of word usage to communicative aims of the authors, to name only the principal ones here. In the chapters of the dissertation it will not be possible to list all and to comment on all the words that are relevant for a point in question. For instance, in the chapter about the role of borrowing and loanwords it is not possible to list all the words and components from the Latin tradition. But it is possible in that chapter to provide the complex search string by which these words can be retrieved from the lexical documentation system.

Now we will give an example of the article structure and its TEI Lex-0 encoding. For this purpose, we present the articles "Sphäre" and "Sphärenanbeter", words that are used in the passage from Hedwig Dohm's book, quoted above. For the main part of the article the components are pretty straightforward and present no unexpected structures. The only peculiarity is the (mis-)use of the tag <usg> for implementing the discourse ontology and the analytical markup for different aspects of the word usages in question. We place this tag within the <sense> tag because the values of the analytical markup are quite often relative to a specific usage (sense) and not to the word with its multiple senses. Figure 3 below shows the article "Sphäre" in oXygen's author view:

[11] The last character is zero, which in the current font almost looks like an "o"; TEI Lex-0 is a new suggestion for a lexicographical baseline encoding format. See: https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html (last access 2020-05-03).

[12] https://zhistlex.de (last access 2020-05-03).

[13] https://www.bmbf.de/en/index.html (last access 2020-05-03).



Sphäre

'Bereich, Wirkungskreis'.

Wird verwendet, um einen Bereich des Wirkens und aktiven Lebens zu bezeichnen. Dabei sind die Bereiche von Mann und Frau entsprechend der Differenzhypthese getrennt und die Aufgaben jedes Bereich sind von der als naturgegeben geltenden Wesensart des männlichen oder weiblichen Geschlechts bestimmt.

Vgl. Paul 2002 Vw. 2; DWDS s.v. 21.05.2014 Vw. 1; DWB 16, 2207 Vw. 9b.

Sichtweise(n): Differenzhypothese

Herkunft: griechisch-lateinisch

(1876) Dohm Natur 122.11 126.16 176.15

»Die Vernunft spottet aller Inbrunst des Glaubens, sie reißt die Gestirne aus ihren Bahnen, die das Vorurtheil ihnen vorgezeichnet, sie hat siegreich gekämpft mit Drachen, Riesen und Teufeln, sie stürzt Götter von ihren Thronen. Vor ihrem siegenden Strahl wird auch der uralte Glaube an die **Sphäre** des Weibes dahinschwinden.«

(122.5f.)

»O über dieses Geschwätz von der **Sphäre** des Weibes, den Millionen Frauen gegenüber, die auf Feld und Wiese, in Fabriken, auf den Straßen und in Bergwerken, hinter Ladentischen und in Bureaus im Schweiße ihres Angesichts ihr Brot erwerben.«

(126.16f.)

»Wenn die Frau nicht über einen starken Geist gebietet, so verfällt sie dadurch, daß sie in einer niederen **Sphäre** festgehalten wird, den Sitten und Lastern der Knechtschaft, deren sie bedarf um sich ihre Situation erträglich zu machen.«

(176.13f.)

(1897) Hilty Jahrbuch 291.5

»Richtig ist ferner im Allgemeinen, was ein moderner Schriftsteller über diesen Gegenstand sagt: es handle sich praktisch zunächst darum, die Frauen in der **Sphäre** des Gemeindestimmrechts zu emanzipiren, sobald sie darüber hinausgehen, so riskiren sie einen Gegenstoss.«

(291.3f.)

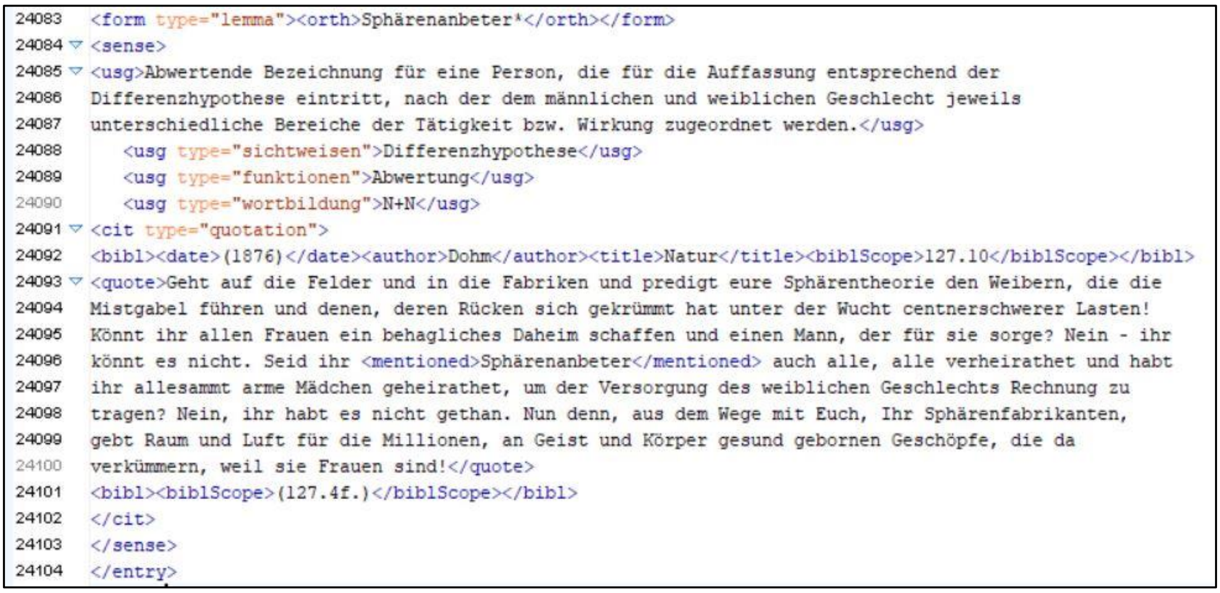
(1908) Freudenberg Manifest 20.48

»Aber das Manifest nimmt freilich an, daß es sich in der Politik nur um Fragen aus der **Sphäre** des Mannes handele. Der moderne Staat (s. b.) beruhe auf der militärischen Macht, zu Lande und zu Wasser; er werde durch die Diplomatie, die Finanzwirtschaft, die Verwaltung der großen Industrien im Betrieb erhalten.«

(20.47f.)

Figure 3: Article “Sphäre”; oXygen’s author view.

The following figure 4 shows the somewhat less complex article “Sphärenanbeter” in oXygen’s text view with the XML encoding:



```
<form type="lemma"><orth>Sphärenanbeter*</orth></form>
<sense>
<usg>Abwertende Bezeichnung für eine Person, die für die Auffassung entsprechend der
Differenzhypothese eintritt, nach der dem männlichen und weiblichen Geschlecht jeweils
unterschiedliche Bereiche der Tätigkeit bzw. Wirkung zugeordnet werden.</usg>
   <usg type="sichtweisen">Differenzhypothese</usg>
   <usg type="funktionen">Abwertung</usg>
   <usg type="wortbildung">N+N</usg>
<cit type="quotation">
<bibl><date>(1876)</date><author>Dohm</author><title>Natur</title><biblScope>127.10</biblScope></bibl>
<quote>Geht auf die Felder und in die Fabriken und predigt eure Sphärentheorie den Weibern, die die
Mistgabel führen und denen, deren Rücken sich gekrümmt hat unter der Wucht centnerschwerer Lasten!
Könnt ihr allen Frauen ein behagliches Daheim schaffen und einen Mann, der für sie sorge? Nein - ihr
könnt es nicht. Seid ihr <mentioned>Sphärenanbeter</mentioned> auch alle, alle verheirathet und habt
ihr allesammt arme Mädchen geheirathet, um der Versorgung des weiblichen Geschlechts Rechnung zu
tragen? Nein, ihr habt es nicht gethan. Nun denn, aus dem Wege mit Euch, Ihr Sphärenfabrikanten,
gebt Raum und Luft für die Millionen, an Geist und Körper gesund geborenen Geschöpfe, die da
verkümmern, weil sie Frauen sind!</quote>
<bibl><biblScope>(127.4f.)</biblScope></bibl>
</cit>
</sense>
</entry>
```

Figure 4: Article “Sphärenanbeter”; XML structure of the entry, oXygen’s text view.

With this kind of analytical markup it is possible to search e.g. for entries that were marked in the following ways:

<usg type="sichtweisen">Differenzhypothese</usg>
<usg type="funktionen">Abwertung</usg>
<usg type="wortbildung">N+N</usg>

To search, e.g., for all entries which are marked to belong to “sichtweise=Differenzhypothese” (point of view=difference hypothesis) one can use oXygen’s xpath query syntax:[14]

//entry[contains(.//usg[@type="sichtweisen"], "Differenzhypothese")]

[14] In a later stage the data will be hosted in a web environment that is able to process xPath-queries and to deliver the results in different formats. Furthermore, it has been pointed out that the “contains”-operator is not the ideal solution for our purpose, compared, e.g., to strict match options. Therefore, we will modify this practice.

The search result will include three entries, which refer to the articles “Sphäre”, “Sphärenanbeter”, and “Sphärenfabrikant”, all related to the difference hypothesis, that is the point of view that women and men by nature belong to different spheres and therefore have different rights and obligations:

/TEI[1]/text[1]/body[1]/div[1]/entry[852]	“Sphäre”
/TEI[1]/text[1]/body[1]/div[1]/entry[853]	“Sphärenanbeter”
/TEI[1]/text[1]/body[1]/div[1]/entry[852]	“Sphärenfabrikant”

To sum up: The lexical documentation system on the one hand serves to analyze and to document individual word usages, on the other hand the system may be addressed from passages in the investigation in order to present further details, to provide the entire subset of relevant material and the textual quotations in their temporal order. The lexical entries are compiled to provide a pragmatic-semantic description of the individual lexical units and their specific uses in the discourse. Thus, the lexical documentation system provides a lexical inventory of the discourse. It is itself based on the TdeF-corpus. The lexical documentation system is to be published in the ZHistLex project by one of the CLARIN-D partners as well.

5 The Generic Potential of Our Infrastructure Strategies

We believe that the research project on word usage in the women’s suffrage discourse has a generic potential in at least three important dimensions:

Firstly, we believe that the integrated coordination of three pillars in a project on lexical research is a fruitful strategy. It combines (a) overall analytical results on vocabulary structure and development (e.g. in a dissertation) with (b) the public presentation of the digital corpus texts that are the basis of these results and (c) a digital lexical documentation system that provides the details of word usage and lexical organisation.

Secondly, we believe that for texts from the 17th to the early 20th centuries the cooperation with CLARIN-D’s German Text Archive (DTA) belongs to the best practices for building up a project specific digital corpus, which we have characterized as a win/win/win-situation for the research project, the DTA and also for other researchers with interests of their own.

Thirdly, the conception of the lexical documentation system with its underlying markup system for multiple dimensions of word usage and vocabulary structure allows to retrieve lexical material according to different criteria (e.g. PoS, communicative function, discourse role, connection to specific points of view in a discourse), to visualize different aspects of vocabulary structure and development, it provides the foundations for writing up the results on specific aspects of vocabulary structure and development in the book or article one has to write. Since there has been a long and controversial debate about the role of alphabetical order[15] and its alternatives[16] in the history of lexicography, the present suggestion for multi-dimensional organization and access via a multi-dimensional markup based on TEI Lex-0 and XPath may well contribute a low-level solution to this controversial point with a long history.

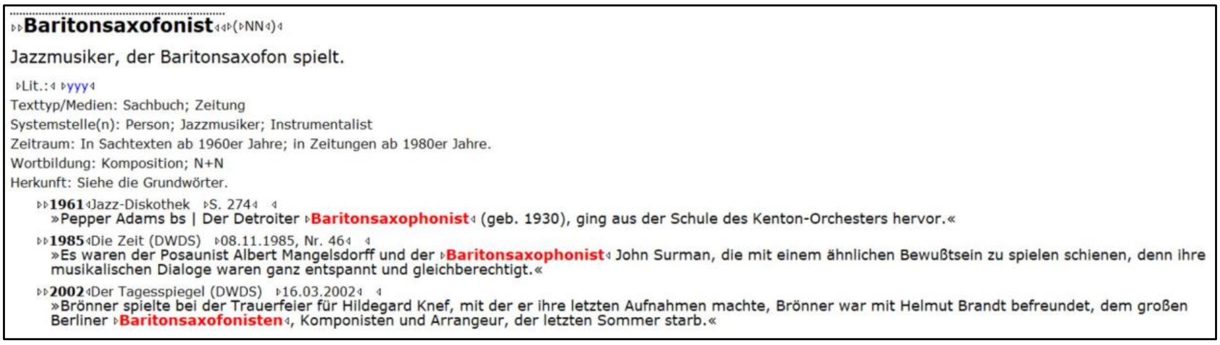
Currently, two other Gießen based projects adopt this kind of research architecture. Andre Pietsch works on word usage and vocabulary structure in texts related to early film and early cinema. Thomas Gloning works on the history and the development of German jazz vocabulary. [17] In respect of the jazz

[15] Jacob Grimm, one of the principal founders of German studies, wrote in the preface to vol. I. of the “Deutsches Wörterbuch” (1854, p. XI): “Nicht minder nothwendig ist dem wörterbuch die alphabetische ordnung und sowol die möglichkeit des vollen eintrags und der abfassung als die sicherheit und schnelle des gebrauchs hängen davon ab. wer reiche beiträge einschalten will, musz die stelle wohin vor augen haben und nicht unschlüssig herum zu suchen, ob das wort schon da sei oder fehle: die biene weisz genau die zelle, zu welcher sie honig einträgt. es würde die arbeit in den wörtern aufheben oder lähmen, wenn man den platz nicht kennt, aus dem sie zu holen sind. schon ihren eingeschränkten samlungen pflegten die alten diese alphabetfolge zum grunde zu legen und wer sie heute nicht handhabt, sondern aufhebt und stört, hat sich an der philologie versündigt.“ The preface is online via woerterbuchnetz.de: `https://tinyurl.com/yaacpscv` or `http://woerterbuchnetz.de/cgi-bin/WBNetz/wbgui_py?sigle=DWB&mainmode=Vorworte&file=vor01_html#abs2` (p. XI) (last access 2020-05-08).

[16] E.g. “conceptual” systems like the one proposed by Hallig and von Wartburg (1952).

[17] This project has been presented at the Göttingen Conference on Historical Lexicography and slightly revised at the 2019 Germanistentag, both in September 2019. The slides are available here: `https://zhistlex.de/folien/Gloning_2019_HistVok-Jazz_Saarbruecken.pptx`.

vocabulary project the lexical documentation system is the core component. The question, whether or not a digital text corpus is feasible at all, is not yet answered. Copyright questions are hard questions when it comes to texts from the 20th and the 21st centuries. Here, the expertise of CLARIN-D might be fruitful, too. Figure 5 shows an early version of the article “Baritonsaxofonist”.



▹▹**Baritonsaxofonist**◃◃▹(▹NN◃)◃

Jazzmusiker, der Baritonsaxofon spielt.

▹Lit.:◃ ▹yyy◃
Texttyp/Medien: Sachbuch; Zeitung
Systemstelle(n): Person; Jazzmusiker; Instrumentalist
Zeitraum: In Sachtexten ab 1960er Jahre; in Zeitungen ab 1980er Jahre.
Wortbildung: Komposition; N+N
Herkunft: Siehe die Grundwörter.

▹▹**1961**◃Jazz-Diskothek ▹S. 274◃ ◃
»Pepper Adams bs | Der Detroiter ▹**Baritonsaxophonist**◃ (geb. 1930), ging aus der Schule des Kenton-Orchesters hervor.«

▹▹**1985**◃Die Zeit (DWDS) ▹08.11.1985, Nr. 46◃ ◃
»Es waren der Posaunist Albert Mangelsdorff und der ▹**Baritonsaxophonist**◃ John Surman, die mit einem ähnlichen Bewußtsein zu spielen schienen, denn ihre musikalischen Dialoge waren ganz entspannt und gleichberechtigt.«

▹▹**2002**◃Der Tagesspiegel (DWDS) ▹16.03.2002◃ ◃
»Brönner spielte bei der Trauerfeier für Hildegard Knef, mit der er ihre letzten Aufnahmen machte, Brönner war mit Helmut Brandt befreundet, dem großen Berliner ▹**Baritonsaxofonisten**◃, Komponisten und Arrangeur, der letzten Sommer starb.«

Figure 5: Current, preliminary state of article “Baritonsaxofonist”; Jazz Vocabulary Project.

One last point on the generic potential of multidimensional background markup. What has been suggested on forms of markup like “points of view=difference hypothesis” in the women’s suffrage discourse can be adapted to dictionary retrodigitisation projects. While many traditional dictionaries have been digitized and made available in lexicographic portals like woerterbuchnetz.de, some of them are not yet available for digital access, e.g. the “Sudetendeutsches Wörterbuch”.[18] Going through the articles of this dictionary one finds many articles that are closely related to a specific cultural background and to aspects of the form of life that is mirrored in the regional vocabulary structure, e.g. in the use of expression for church feasts (“Peter Ketten”), for plant and animal names (e.g. “Brotkäferlein”), evaluative expressions for persons (“Prahl-fotze” ‘boasting female person’) and others. It is obvious that such aspects of vocabulary structure would deserve additional markup in an electronic version.

6 Results and Open Questions

In this paper, we have portrayed a research project with a lexical focus that opens up a number of generic perspectives for the CLARIN-D infrastructure. Anna Pfundt’s research project aims to analyze and describe the vocabulary structure and developments that are connected with the textual cosmos of the discourse on women’s suffrage within the First Women’s Movement in Germany around 1900. Her work comprises three major pillars: (1) the dissertation as an analytical prose text, which will be available both in print and in digital form; (2) a digital text corpus which is built up in an ongoing way within CLARIN’s German Text Archive and which at present comprises around 70 full texts, (3) a digital lexical documentation system, which documents the details of word usages but which allows interlinking to other word articles, to the corpus texts and to the text of the dissertation. This lexical documentation component is to be published with one of the German CLARIN centers as well.

By now, it should have become clear that the integrated combination of (1) investigation, (2) digital text corpus and (3) digital lexical documentation system provides a fruitful architecture for research questions like word usage in the debates on women’s suffrage within the First Women’s Movement (or word usage related to early film and cinema; word usage related to the development of jazz in German). This research architecture provides generic solutions that are closely related to the CLARIN-D infrastructure: building up a project-specific sub-corpus within DTAE is certainly the most important aspect, the support in respect of lexical organization and the respective standards and best practices are equally important.

As a prototypical use case, the project shows how a lexical investigation can simultaneously benefit from the resources of the CLARIN-D infrastructure (here the DTA) and contribute to its further expansion. The compilation of the TdeF corpus is at the same time the textual basis for the research project

[18] `https://www.uni-giessen.de/fbz/fb05/germanistik/forschung/sprache/sdwoerterbuch`; `https://de.wikipedia.org/wiki/Sudetendeutsches_W%C3%B6rterbuch` (last access 2020-05-03).

on word usage in the discourse on women's suffrage and an important contribution to expand the holdings of the DTA in an important phase of German language and discourse history around 1900.

Finally, once a text is integrated in the DTA, it is also available for a broad spectrum of other research questions, other research methods and tools. Apart from the contributing project and the DTA, the win-win situation includes the whole research community and a plethora of possible research questions and methods.

One last important aspect of CLARIN-D's infrastructure must not be forgotten: the ongoing personal support we have experienced over the years, notably from the DTA team.

Acknowledgements

The authors would like to thank Christiane Benetz, Dennis Kaltwasser, Roman Degreif and Andre Pietsch for their support and their suggestions. We equally thank the reviewers of the initial proposal, of the Leipzig poster and of this paper for their questions, their very helpful suggestions, and their constructive criticism. Again: a big "Thanks" to the DTA team for ongoing support on the TdeF corpus! We are also very grateful for support on the lexical documentation system, both from members of CLARIN-D and from the ZHistLex project (`https://zhistlex.de`).

References

Hedwig Dohm. 1876. Der Frauen Natur und Recht. Wedekind & Schwieger, Berlin. http://deutschestextarchiv.de/book/show/dohm_frauenfrage_1876 (last access 2020-02-09).

Ute Frevert. 1995. „Mann und Weib, und Weib und Mann". Geschlechter-Differenzen in der Moderne. Beck, München.

Gerd Fritz. 2006. *Historische Semantik*. second, updated edition. Metzler, Stuttgart, Weimar.

Ute Gerhard. 1990. *Unerhört. Die Geschichte der deutschen Frauenbewegung.* Rowohlt. Reinbek near Hamburg.

Ute Gerhard. 2009. Frauenbewegung und Feminismus. Eine Geschichte seit 1789. Beck, München.

Thomas Gloning. 2003. Organisation und Entwicklung historischer Wortschätze. Lexikologische Konzeption und exemplarische Untersuchungen zum deutschen Wortschatz um 1600. Niemeyer, Tübingen.

Thomas Gloning. 2012. Diskursive Praktiken, Textorganisation und Wortgebrauch im Umkreis der ersten Frauenbewegung um 1900. *Historische Pragmatik*. Ed. Peter Ernst. De Gruyter, Berlin, New York. 127-146.

Thomas Gloning. 2013. Historischer Wortgebrauch und Themengeschichte. Grundfragen, Corpora, Dokumentationsformen. *Perspektiven einer corpusbasierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch" an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12.-13. Dezember 2014.* Ed. Ingelore Hafemann. Berlin. 317-370.

Florence Hervé. 1988. *Geschichte der deutschen Frauenbewegung.* Pahl-Rugenstein, Köln.

Rudi Keller. 1995. *Zeichentheorie. Zu einer Theorie semiotischen Wissens.* Francke, Tübingen, Basel.

Anna Pfundt. 2017. Frauenwahlrecht? Oder Damenwahlrecht? Oder doch ein allgemeines Wahlrecht? – Zum Wortgebrauch in der Diskussion um das Frauenwahlrecht um 1900. *Im Zentrum Sprache*. https://sprache.hypotheses.org/542 (last access 2019-08-23).

Ute Rosenbusch. 1998. Der Weg zum Frauenwahlrecht in Deutschland. Nomos-Verlag, Baden-Baden.

Angelika Schaser. 2006. *Frauenbewegung in Deutschland 1848-1933.* Wissenschaftliche Buchgesellschaft, Darmstadt.

Georg Stötzel and Martin Wengeler. 1995. *Kontroverse Begriffe. Geschichte des öffentlichen Sprachgebrauchs in der Bundesrepublik Deutschland*. De Gruyter, Berlin, New York.

Margrit Twellmann. 1972. Die Deutsche Frauenbewegung. Ihre Anfänge und erste Entwicklung 1843 – 1889. 2 Bände. Hain, Meisenheim am Glan.

Kerstin Wolff, Alexander Geyken, and Thomas Gloning. 2015. Kontroverse Kommunikation im Umkreis der ersten Frauenbewegung. Wie können digitale Ressourcen die sprachliche Untersuchung und die Ergebnisdokumentation verbessern? *Grenzen und Möglichkeiten der Digital Humanities*. Eds. Constanze Baum and Thomas Stäcker. Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1. http://zfdg.de/sb001_010

User Support for the Digital Humanities

Heidemarie Sambale **Hanna Hedeland** **Tommi Antero Pirinen**

Hamburg Centre for Language Corpora (HZSK)
Universität Hamburg, Germany
{firstname.lastname}@uni-hamburg.de

Abstract

In this article, we describe a user support solution for the digital humanities. As a case study, we show the development of the CLARIN-D Helpdesk from 2013 into the current support solution that has been extended for several other CLARIN-related software and projects and the DARIAH-ERIC. Furthermore, we describe a way towards a common support platform for CLARIAH-DE, which is currently in the final phase. We hope to further expand the help desk in the following years in order to act as a hub for user support and a central knowledge resource for the digital humanities not only in the German, but also in the European area and perhaps at some point worldwide.

1 Introduction

For both the ongoing digitalisation of humanities research in general and the CLARIN infrastructure in particular, the non-technical aspects of adequate training and user support are crucial for acceptance and involvement from the research communities. Many humanities researchers come from a rather non-technical background, and the use of digital tools and resources has not yet entered the curriculum to an appropriate extent. Researchers thus face various problems when confronted with tools and platforms for digital humanities research, many of which might not be predictable to the developers. Improving the usability of such kinds of tools and providing comprehensive documentation is undoubtedly very important, but in the end there is no replacement for a reliable help desk to assist users when they for various reasons struggle with digital resources, tools and services or need qualified advice in methodological questions.

Since digital humanities is multidisciplinary, the user support becomes a very central and important resource for exchanging information and experience, and for gathering expertise from various contexts. Apart from providing users with reliable support, the direct interaction with the users also provides valuable input about the users and their behaviour and on existing problems with the tools and platforms for infrastructure providers and developers. In our paper, we describe the development of a comprehensive resource based on a sustainable platform with re-usable workflows, for which we have also developed various strategies for scalability.

The rest of the article is structured as follows: In section 2 we describe the background of our approach to user support and our help desk. In section 3 we present the current developments. In section 4 we introduce the scalability strategies and in section 5 we describe the technical implementation of various challenges regarding management and scalability of the help desk. Finally, in section 6, we describe our plans in the future, based on the current development.

2 The CLARIN-D Helpdesk

The CLARIN-D Helpdesk (Lehmberg, 2014; Lehmberg, 2015) was first launched in 2013 to provide the necessary user support for the emerging CLARIN-D infrastructure. After a thorough review of current ticketing systems, the open source OTRS platform[1] was considered the best solution in order to meet the needs of the infrastructure's users and developers. In OTRS, the support requests are managed by using support *tickets*. A ticket represents and documents the entire conversation between agent and customer chronologically. The support is provided by *agent* users that are active within one or more support areas. The tickets are organised into *queues*, which represent the various support areas. Figure 1 shows the arrangement of the queues: every centre (e.g. EKUT, HZSK) has its own queue with several subqueues for every section they offer support.

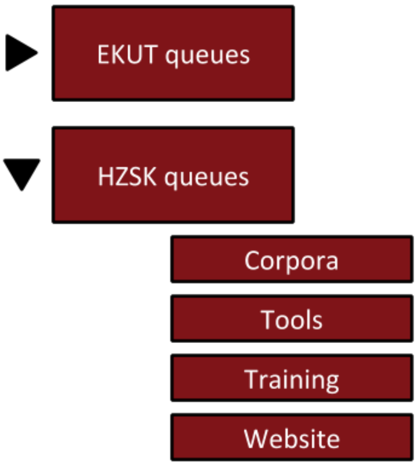


Figure 1: Structure of queues representing support areas in the CLARIN-D Helpdesk.

Two agent types can be assigned to each ticket, responsible and owner. The owner handles the customer requests, whereas the responsible monitors the overall progress. While the owner of the ticket can change in the course of different working stages, the responsible remains the same agent. The information about which agent is responsible and which one is the owner of the ticket is visible to other agents. Apart from the communication with the customer, it is possible to communicate internally in a structured and documented manner within the ticket conversation in order to find an appropriate answer to an inquiry collaboratively.

This technical infrastructure and the related concepts and workflows comprise the relevant functionality required to reliably answer and document incoming queries. The use of a ticketing system solves most of the issues related to providing user support via email, e.g.

- relevant information (such as previous answers, templates etc.) is not accessible for other agents due to the use of private mail accounts,
- duplicate answers are sent when a common mail account is used by several agents and coordination fails,
- comprehensive documentation of every working step is not created,
- workflows cannot be standardised and simplified, e. g. by the use of templates and automatic answers
- long response time owing to holiday or illness.

[1] `https://otrs.com/`

In any case, without some kind of ticketing system, status and responsibilities for inquiries need to be managed individually. By using the OTRS, the support progress is clear to other agents, so that another agent can easily take over if the current owner of the ticket is absent.

As shown in Figure 2, tickets arrive to the help desk from multiple sources. Depending on the origins, the tickets may need further sorting in the queues and assigning to the agents. The ticketing system records metadata about the issues, such as first response times and closing times that are used to gather statistics relevant for the goal of providing efficient user support. The textual content of the help desk, i.e. questions and answers, can also be searched as a semi-structured knowledge base, or used as the basis for edited FAQ articles, which are also distributed from the help desk.

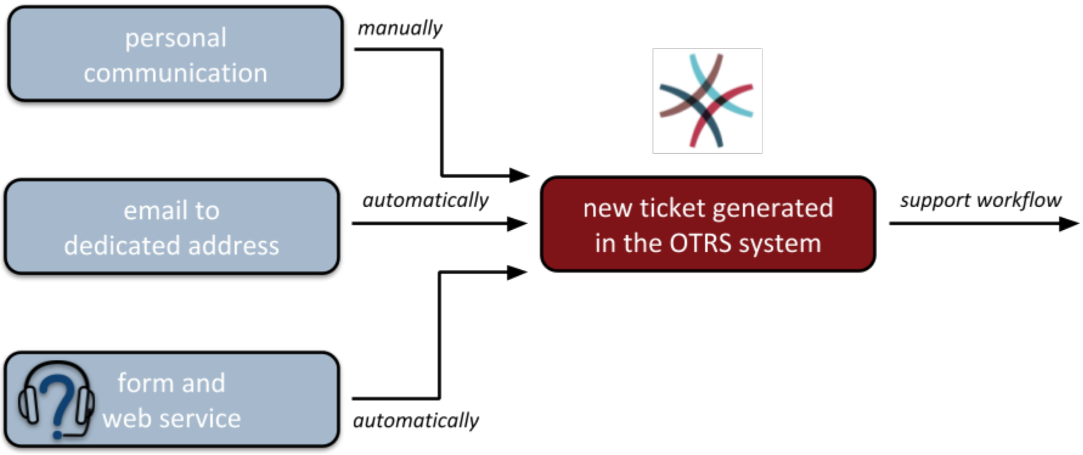


Figure 2: Tickets can be generated in various ways, allowing for pre-sorting and delegation through queue-specific addresses or web service parameters.

Since the CLARIN-D infrastructure is distributed and centre-based, the queue structure of the CLARIN-D Helpdesk models the centres and the services they provide, and the help desk is used to distribute the support requests to the relevant experts. Most of the tickets are automatically sorted and delegated using parameters from email addresses, web forms or keywords. Not automatically sorted tickets are manually assigned to the right queue by the first line support agents, as shown in Figure 3. The first line support is carried out by experienced student assistants, who receive specific training. The training involves monitoring and managing tickets, answering common and general questions, and delegating incoming inquiries that have not already been assigned to a queue automatically. Administrators and first line support have a complete overview of all queues and agents whereas most expert agents of the second line support will only be able to see tickets and queues relevant to them. The experts comprising the second line support are researchers and developers of the participating centres and projects. Before an inquiry is successfully closed, several experts might contribute with their respective expertise on a complex matter, requiring the ticket to be moved across queues and reassigned several times. According to our quality guidelines, at least 95 % of the tickets should be answered within two days. This target was not only met, the percentage of tickets answered within two days has been increasing steadily in recent years.

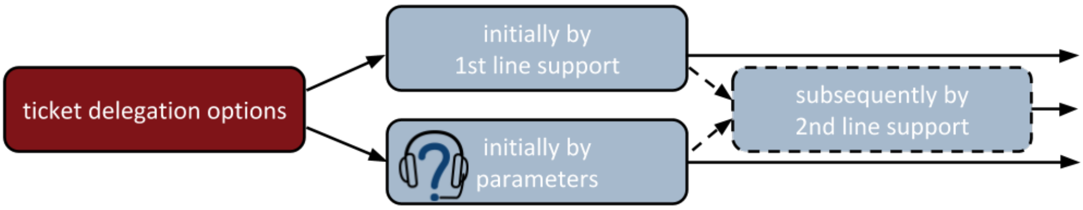

ticket delegation options
initially by 1st line support
initially by parameters
subsequently by 2nd line support

Figure 3: Tickets are sorted by queues and assigned to the right persons automatically or manually.

2.1 Inquiries at the HZSK

There is a huge range of the inquiries content arriving the CLARIN-D Helpdesk and also by looking only on inquiries for the Hamburg Centre for Language Corpora. The type of inquiries at the HZSK can be categorised into four main areas:

- **Corpora** – mainly requests to access one of the corpora in our repository[2], but also inquiries about the handling and terms of use of the corpora
- **Services** – this section contains many different types of questions, mainly from researchers in different projects, about data management, data format, best practice, hosting corpora, corpus creation and corpus curation
- **Tools** – inquiries regarding the EXMARaLDA[3] software, mainly questions about handling, issues and requests for future developments
- **Training** – registrations or questions about trainings, organised and carried out by the HZSK

It is not always easy to draw the line between these four sections. Especially the area service has intersections with every other area. Nevertheless, the categorisation helps to get an overview of the distribution of the inquiries. As we can see in Figure 4, more than 80 % of the 236 incoming inquiries at the HZSK in 2019 belong to the sections tools and corpora. Answers and further communication with the customer are not included in that number, so it is also important to note that this does not reflect the invested time. While inquiries in the sections Corpora, Tools and Training can often be solved within a few messages, the questions in section Service require intensive support over a period of time.

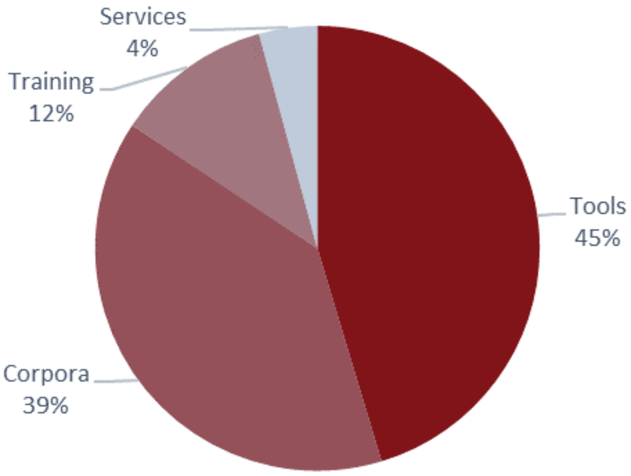


Figure 4: Distribution of inquiries at the HZSK in 2019.

2.2 Distributed Support within CLARIN-D

For complex inquiries, the contribution of several centres with different areas of expertise is useful. An example of an area often requiring coordination between multiple experts and centres is support for projects interested in creating digital spoken language resources using transcription software. The creation and analysis of spoken language resources usually requires a great deal of technical and methodological support, both due to the mostly non-technical background of the users, and due to the highly complex nature of the task at hand. Spoken corpora comprise various interrelated data types and file formats and complex metadata valid across the corpus' components. For the creation of transcripts several software systems exist that are specialised for specific scenarios. Apart from expertise regarding

[2] https://corpora.uni-hamburg.de/hzsk/en/repository-search
[3] https://exmaralda.org

certain tools, centres also provide expertise according to the raw data, e.g. depending on the language or modality of the resource.

Based on the areas of expertise depicted in Figure 5, a simple technical question regarding the transcription software EXMARaLDA might be initially answered by the first line support at the HZSK centre. If the reply of the customer contains a more complex question regarding the software's scope of use, the request would then be forwarded to the second line support, who might find that this user should rather try the tools provided by another centre (e.g. IDS or BAS) and forward the ticket accordingly.

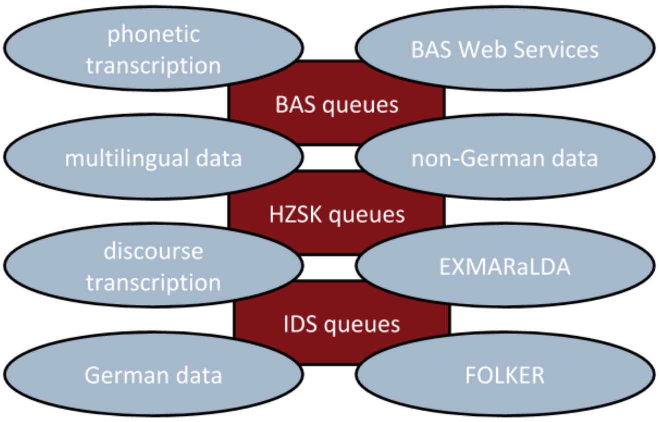


Figure 5: Professional support within the complex area of spoken language resources requires coordination and cooperation between several centres providing complementary expertise.

3 Support beyond CLARIN-D

Beyond CLARIN-D, the CLARIN-D Helpdesk is used to provide support on the European CLARIN level; for the CLARIN Virtual Language Observatory (VLO)[4] user feedback and outreach - including both generic feedback about user experience or reports of erroneous metadata - and for the CLARIN Federated Content Search Aggregator (FCS)[5]. For the VLO, the user feedback is divided into tickets regarding the VLO application and tickets regarding the metadata and the resources it describes. The tickets regarding the VLO application are handled by the VLO developers and the metadata related tickets can be handled by first line support or by the SCCTC Metadata Curation Taskforce as a part of their quality assurance work.

Apart from tools and services directly integrated into CLARIN, some support workflows for related tools from other contexts have also been successfully integrated into the CLARIN-D Helpdesk. This allows for interaction between developers and users of these tools across expertise and support areas and thus for a wider outreach. This could be a first step for these partners to become a part of the emerging Knowledge Sharing Infrastructure[6] even though they are not certified CLARIN centres (yet).

4 CLARIAH-DE - joining forces

In the course of merging CLARIN-D and DARIAH-DE into CLARIAH-DE, the support was also merged into one comprehensive help desk based on the CLARIN-D Helpdesk in November 2019. The integration of the DARIAH-DE support differed from the other extensions in the help desk, hence not only a new module of queues and agents needed to be added, but a fusion has to be made. Apart from uniform external presentation and communication as well as user information, the challenge is to bring the two different internal organisation structures together.

As previously mentioned, CLARIN-D is centre-based. Therefore, in OTRS every CLARIN-D Centre has its own queue, with respective agents for each queue. Some centres have a wide range of tasks or

[4] https://vlo.clarin.eu/

[5] https://www.clarin.eu/content/content-search

[6] https://www.clarin.eu/content/knowledge-centres

multiple software/applications, therefore it is necessary for them to organise the centre queue into several subqueues, as shown in Figure 1. In CLARIN-D, the agents are mainly managed in groups, whereas every centre has its own group. This means that despite the subdivision within the individual centre queues, every agent is part of one centre group and can work with every ticket in that group (see Figure 6). This organisation is possible because there are only a few agents in each centre, belonging to the respective groups.

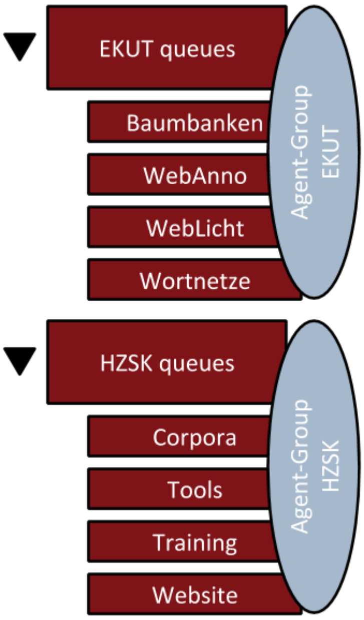


Figure 6: Queues and agent responsibility in CLARIN-D

In contrast, DARIAH-DE is much more centralised which leads to agents with varying overarching responsibilities for different queues. Figure 7 shows a section of the DARIAH-DE queues and responsibilities schematically.

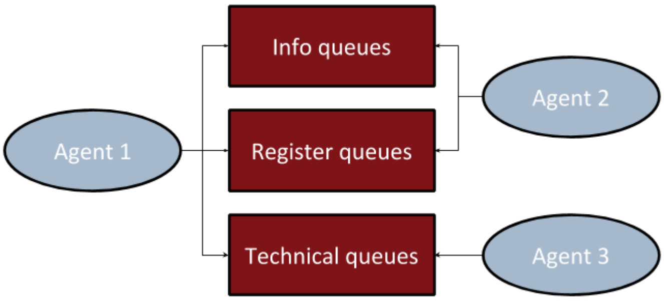


Figure 7: Queues and agent responsibility in DARIAH-DE

If we compare the help desk of CLARIN-D (Figure 6) and DARIAH-DE (Figure 7) there is a structural difference regarding responsibilities of agents within queues. In recent years, both help desk structures had developed their own workflows based on the respective system and both structures have proven their value in user support. So the goal was to keep up both structures. To achieve this, the best way to handle the increased number of agents with its partly overlapping responsibilities was to extend the role system in OTRS. Thereby it was possible to integrate the two different structures within one OTRS system in an elegant way.

The setting options for generating statistics have also been expanded. It is now possible for certain agents without admin status to generate statistics on the tickets in their queues. The use of general statistic templates such as ’tickets created last month/quarter/half/year’ makes it possible to use them regularly.

In addition, special statistic templates can also be created to meet individual needs. Thanks to the restructuring, it is now very easy to expand the group of people who want to generate individual statistics. We can now offer this feature to all previous and future users of the CLARIAH-DE Helpdesk.

The main part of the migration has been completed successfully, but the CLARIAH-DE Helpdesk will be further optimized and adapted to the needs of the users.

5 Customizing ticket handling for increased flexibility

To be able to integrate support workflows for different tools and services from different contexts, we aim to make the help desk platform and all related components highly flexible. The support system is provided to the end users typically as a part of a website or web application. For seamless integration of the support, there are various technologies that can be used for a certain service, tool or other support area, usually the implementation will be carried out by the respective website managers. The most basic model of integration can be achieved with a support email address that forwards to the OTRS system. The CLARIN-D Helpdesk also has a web API based on SOAP that can be used to create and assign support requests. We provide a reference implementation of an HTML form for this API in our public GitHub repository that can be customised according to the requirements for the various support areas.[7] The integration of DARIAH-ERIC has been built on top of this API for WordPress and is likewise available at their GitHub repository.[8] Further information about the implementation of DARIAH-ERIC are available in Raciti et al.(Raciti et al., 2019). Each of the ticket creation workflows in the schematic presentation in Figure 2 come with benefits and costs. Furthermore, some of the technical improvements have been made to enhance the automatic ticket assignment in order to minimise the overhead of the first line support in dealing with spam and to categorise more tickets automatically, e.g. by analysing and state more precise keywords in the filter options. By this, the first line support can use their time more efficiently in actual user support tasks. To retain this in future, the integration of further support areas should be kept as simple as possible.

6 Outlook

We have described a scalable help desk system with sustainable workflows introduced for CLARIN-D but already used far beyond its original designation. In the future, we hope to extend the help desk and the areas of support catered for even further. On the European level, we have already integrated support for several CLARIN services and the DARIAH-ERIC Helpdesk into our system, hence we are currently operating a national and international CLARIAH-DE Helpdesk. While in the past the administration of the agents and their authorisations could largely be managed via groups, the merge and the help desk expansion required the development of an extensive role system, which allows for additional flexibility when integrating further organisational units. Apart from the challenge ahead in optimizing the merged CLARIN-D and DARIAH-DE user support, we are looking forward to integrating complementary support areas and workflows and to further enhancing the usage of the help desk as a central knowledge resource for the digital humanities.

References

Timm Lehmberg. 2014. The CLARIN-D Help Desk. In *Papers, Posters and Demos CAC2014*. CLARIN ERIC: Utrecht, The Netherlands.

Timm Lehmberg. 2015. Wissenstransfer und Wissensressourcen: Support und Helpdesk in den Digital Humanities. In *FORGE*, pages 25 – 27.

Marco Raciti, Yoann Moranville, Raisa Barthauer, Stefan Buddenbohm, and Dorian Seillier. 2019. D5.4 - Implementation of a centralized helpdesk and marketplace mockup. Research report, DARIAH, March.

[7] https://github.com/hzsk/clarind-helpdesk

[8] https://github.com/DARIAH-ERIC/contact-helpdesk

Cross disciplinary overtures with interview data: Integrating digital practices and tools in the scholarly workflow

Stefania Scagliola
University of Luxembourg
stefania.scagliola@uni.lu

Louise Corti
University of Essex
corti@essex.ac.uk

Silvia Calamai
University of Siena
silvia.calamai@unisi.it

Norah Karrouche
Erasmus University Rotterdam
karrouche@eshcc.eur.nl

Jeannine Beeken
University of Essex
jeannine.beeken@essex.ac.uk

Arjan van Hessen
University of Twente
a.j.vanhessen@utwente.nl

Christoph Draxler
University of Muenchen
draxler@phonetik.uni-muenchen.de

Henk van den Heuvel
Radboud University
H.vandenHeuvel@let.ru.nl

Max Broekhuizen
Erasmus University Rotterdam
maksbroekhuizen@gmail.com

Khiet Truong
University of Twente
k.p.truong@utwente.nl

Abstract

There is much talk about the need for multidisciplinary approaches to research and the opportunities that have been created by digital technologies. A good example of this is the CLARIN Portal, that promotes and supports such research by offering a large suite of tools for working with textual and audio-visual data. Yet scholars who work with interview material are largely unaware of this resource and are still predominantly oriented towards familiar traditional research methods. To reach out to these scholars and assess the potential for integration of these new technologies a multidisciplinary international community of experts set out to test CLARIN-type approaches and tools on different scholars by eliciting and documenting their feedback. This was done through a series of workshops held from 2016 to 2019, and funded by CLARIN and affiliated EU funding. This paper presents the goals, the tools that were tested and the evaluation of how they were experienced. It concludes by setting out envisioned pathways for a better use of the CLARIN family of approaches and tools in the area of qualitative and oral history data analysis.

1 Introduction

Although there is much talk about the need to open up cross-disciplinary dialogue and prioritize the use of open-source software, when considering disciplines that work with interview data, we can observe a kind of pillarisation of practices. Support for the multidisciplinary approach to interview data has been endorsed by scholars such as Van den Berg et al. (2011), De Jong et al. (2014), Corti et al. (2016) and Van den Heuvel et al. (2017), but most scholars are completely unfamiliar with each other's approaches, and hesitate to take up technology. When software is used, it is often proprietary and binds scholars to a particular set of practices. This paper sets out to explore how to better exploit the rich multidisciplinary potential of interview data through the use of technology. To that end a multidisciplinary international community of experts organised a series of hands-on workshops with scholars who

Stefania Scagliola, Louise Corti, Silvia Calamai, Norah Karrouche, Jeannine Beeken, Arjan van Hessen, Cristoph Draxler, Henk van den Heuvel, Max Broekhuizen and Khiet Truong 2020. Cross Disciplinary Overtures with Interview Data: Integrating Digital Practices and Tools in the Scholarly Workflow. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 126–136.

work with interview data, and tested the reception of a number of digital tools that are used at various stages of the research process. We engaged with tools for transcription, for annotation, for analysis and for emotion recognition. The workshops were held at Oxford, Utrecht, Arezzo, Munich, Utrecht and Sofia between 2016 and 2019, and were mostly sponsored by CLARIN. Participants were recruited among communities of historians, social science scholars, linguists, speech technologists, phonologists, archivists and information scientists. The website https://oralhistory.eu/ was set up to communicate across disciplinary borders.

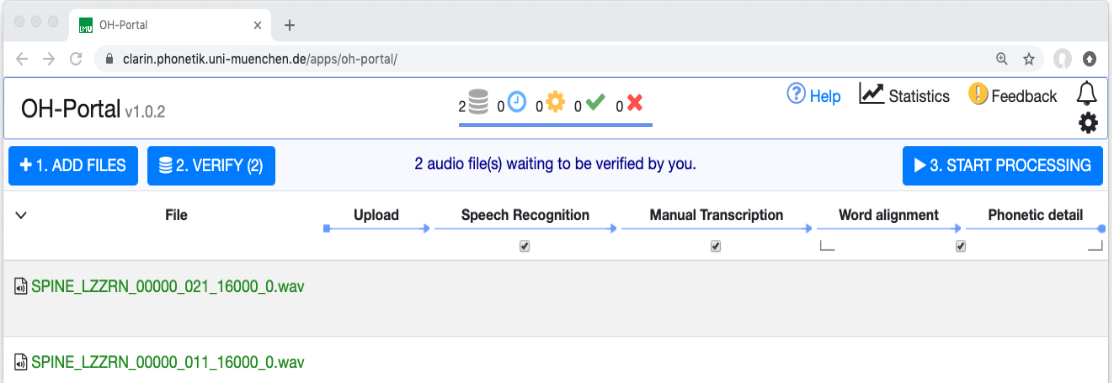
2 Digital tools to work with interview data

A broad diversity of practices can be observed among scholars who work with interview data. Within every discipline distinct sub-disciplines exist, and disciplinary 'silos' certainly complicate collaboration across computer science, humanities and social science. Scholars use the same term for very different practices, or do similar things, but give it different names (De Jong et al., 2011). Frames of interpretations differ. For instance, an oral historian will typically approach a recorded interview as an intersubjective account of a past experience, whereas another historian might consider the same source of interest only because of the factual information it conveys. A social scientist is likely to try to discover common themes and similarities and differences across a whole set of interviews, whereas a computational linguist will rely on counting frequencies and detecting collocations and co-occurrences, for similar purposes. On the other hand sociologists who interview, often seek to understand their interviewees in the same way as (oral) historians. The approaches with regard to re-use of data and anonymisation however are quite different (Van den Berg et al., 2011). The question is how they can benefit from the myriad of freely available transcription, annotation, linguistic and emotion recognition tools. To address this diversity each workshop would start off with an informative session that sketched the various 'landscapes of practices': the different kinds of methodological approaches to attributing meaning to interview data. This exercise in demystification was gratefully received by our audiences, as it offered the opportunity to grasp the essence of the various approaches. After this, sessions were held consisting of a short introduction followed by a step by step tutorial to practice with the various tools. This was done in groups under close supervision of experienced digital humanities scholars. After each session, a short evaluation round was held with group interviews that were recorded.

3 Creating a Transcription Tool: the T-Chain

In the first session participants had a chance to work with functionalities surrounding automatic speech recognition (ASR). Transcription is at the core of research based on interviews and scholars often require full verbatim transcripts. This often means that the focus of attention shifts from the aural dimension of the narrative to its textual representation (Portelli 2006, Boyd 2013). ASR may challenge this practice, as it could take over the laborious practice of manual transcription. Scholars are however sceptical about results of ASR, as they can be disappointing. Accepting incorrect results could however also be considered as an opportunity to rethink the standard practice of relying on full manual transcripts. By aligning audio to the ASR output researchers can easily browse through an entire interview or interview collection, making the audio more accessible and present in the process of analysis. This may lead to a practice in which only particular passages need to be fully described. With this, and other possible uses in mind, the idea of the Transcription Chain, or T-Chain was born (Van der Heuvel 2019). The first workshops were designed to collect requirements for such a tool to be used by a broad diversity of scholars.

A first version of the OH Portal was presented at the Munich workshop in September 2018 (https://clarin.phonetik.uni-muenchen.de/apps/oh-portal/). The workflow consists of the steps *upload, automatic speech recognition, manual correction of the ASR transcript, word segmentation and alignment,* and *phonetic detail* (Figure 1). The T-chain currently processes Dutch, English, Italian and German audio. The portal automatically checks the audio file format (it has to be WAV) and splits stereo recordings into separate mono audio files. These are uploaded to a server, and from there sent to different third-party providers. Thanks to a Google grant, the portal currently supports Google speech recognisers for many languages. To aid in the selection of services, a one-line summary of the service providers' privacy policy may be displayed, together with a link to the full legal text. When the ASR results are sent back the user may choose to check and correct the transcript manually (Figure 2). For this, the transcription editor OCTRA (Pömp et al., 2017) opens within the browser.



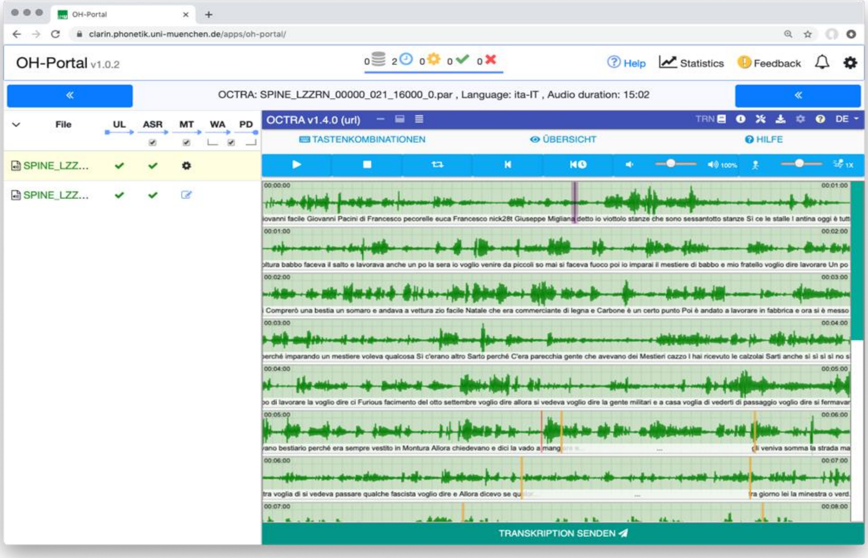
2 audio file(s) waiting to be verified by you.

SPINE_LZZRN_00000_021_16000_0.wav

SPINE_LZZRN_00000_011_16000_0.wav

Figure 1: OH portal with two files in the workspace

This editor features a number of views and can be adapted to various transcription systems.



OCTRA: SPINE_LZZRN_00000_021_16000_0.par , Language: ita-IT , Audio duration: 15:02

iovanni facile Giovanni Pacini di Francesco pecorelle euca Francesco nick28t Giuseppe Migliana detto io viottolo stanze che sono sessantotto stanze Sì ce le stalle l antina oggi è tutt

oltura babbo faceva il salto e lavorava anche un po la sera io voglio venire da piccoli so mai si faceva fuoco poi io imparai il mestiere di babbo e mio fratello voglio dire lavorare Un po

Comprerò una bestia un somaro e andava a vettura zio facile Natale che era commerciante di legna e Carbone è un certo punto Poi è andato a lavorare in fabbrica e ora si è messo

perché imparando un mestiere voleva qualcosa Sì c'erano altro Sarto perché C'era parecchia gente che avevano dei Mestieri cazzo I hai ricevuto le calzolai Sarti anche sì sì sì sì no s

bo di lavorare la voglio dire ci Furious facimento del otto settembre voglio dire allora si vedeva voglio dire la gente militari e a casa voglia di vederti di passaggio voglio dire si fermavar

vano bestiario perché era sempre vestito in Montura Allora chiedevano e dici la vado a mang ... di veniva somma la strada ma

tra voglia di si vedeva passare qualche fascista voglio dire e Allora dicevo se qu ... ra giorno lei la minestra o verd

Figure 2: Transcription editor in the OH Portal window

After the manual correction of the ASR transcript, automatic word segmentation can be performed by using the WebMAUS service (Kisler et al., 2012). The result of this step is a word-based time-aligned transcript of the recording (Figure 3).

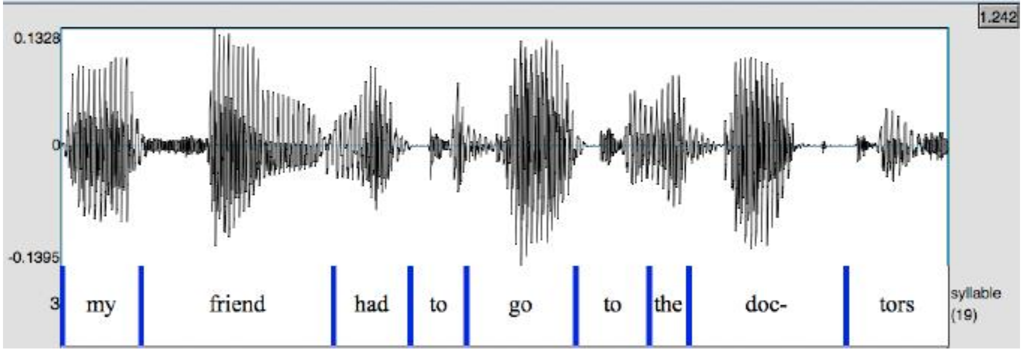


Figure 3: Example of a word alignment after ASR, visualised in Praat

At every step of the process, the output can be downloaded to the local computer in different formats, e.g. plain text, tab- or comma-separated tables, Praat TextGrid, Emu Annot-JSON, and ELAN for further processing and analysis. These exports can be made file-wise or column-wise, i.e. many files at once.

The overall performance and acceptance of the OH Portal depends on a) transcription quality, and b) on user friendliness. Transcription quality depends largely on the acoustic quality, which in turn depends on the recording situation and the speakers, and on the capabilities of the underlying ASR systems. User friendliness depends on the graphical user interface, the speed of interaction and the feeling of being in control of one's own data. We recommend that users convert their audio files using tools such as GoldWave, Audacity or To_Wave_Convertor so that they may be processed by the OH portal. The OH Portal is limited by the amount of memory available in the browser – currently, files up to 300 MB can be processed. This corresponds to approx. 118 minutes of recordings at 22.05 kHz and 16 bit mono. Note that processing long recording files means long waiting times – for a smooth operation we thus recommend that users split their recordings into short meaningful units and process them individually. At hands-on workshops, we provided sample files in different languages with durations of up to 5 minutes so that all participants could work in parallel, and experience the progress of their files through the workflow.

Finally, note that since ASR is provided by external providers, restrictions may apply. This is especially true for commercial providers who impose monthly quotas on accepted file size, recording duration or number of requests per month.

4 Using existing data annotation tools

Most researchers working with interview data use some form of annotation technique. The way these tools are used varies considerably across disciplines and the tools mostly used are proprietary. This can vary from using a pen and paper, coding of digital sources with a tool, to linguistic identification through information extraction tools.

In the area of social science, qualitative data analysis software known as QDAS is popular, and includes market leaders such as NVivo, Atlas.ti and MaxQDA. They allow analysis of text and audio-visual sources, but can be costly and tend to lock-in researchers, with no import or export capability, nor any intermediate 'portable' formats, for example, in XML (Corti and Gregory, 2011).

For the workshop we chose to offer NVivo and the open source annotation tool Elan. The first, designed for social scientists, can collect all kinds of different sources into one project, can classify and group these, mark-up text, images and audio-visual sources with thematic 'nodes', and add notes, known as 'memos'.

ELAN (https://archive.mpi.nl/tla/elan), in contrast, is freely available, and was developed at the Max Planck Institute for use in linguistics. Users can create 'tiers' in which annotations can be added to audio or video files, differentiating types of tiers (for example different speakers) and specifying 'parent tiers' (Wittenburg et al., 2006). The ability to annotate the audio enables users to engage with multiple dimensions of an interview from as early as the point of recording the data. This tool is also seen as very suitable to annotate audiovisual data.

4.1 Linguistic analysis tools

Text mining tools used by computational linguists could also enrich the practice of social science scholars and historians, by offering insight into the structure of language. The semantic contexts in which terms are used, are telling for how social reality is created and remembered. In non-computational methods the analysis of data starts with reading one transcript or listening to one interview at the time. Linguistic tools offer the detection of patterns in language or speech features by looking at the entire collection or to subcollections at once. This might be very useful when re-using a collection from an archive that is new to the researcher. Several features can be explored: concordances and correlations, processing syntactic tree structures, searching for named entities, and applying emotion recognition (Armaselu et al., 2019).

This part of the workshop would start with an introduction to linguistic tools and their functions: lemmatizers, syntactic parsers, named entity recognizers, auto-summarizers, tools for detecting concordances/n-grams and semantic correlations. Participants were then given a live demo of the software tools and then some step by step guided exercises with data.

The first tool to be introduced was Voyant (https://voyant-tools.org/), a lightweight text analysis tool that yields output on the fly (Sinclair et al., 2016). This was followed by Stanford CoreNLP (https://stanfordnlp.github.io/CoreNLP/), a linguistic tool that can automatically tag words in a number of different ways, such as recognizing part of speech, type of proper noun, numeric quantities, and more (Manning et al., 2014). Lastly, participants were encouraged to use Autosummarizer (http://autosummarizer.com/), a website which uses AI to automatically produce summaries of texts.

These tools are relatively lightweight and require little to none installation of programs, so were far more readily amenable to the participants. A more complex tool was TXM, which stands for 'textometry', a methodology allowing quantitative and qualitative analysis of textual corpora, by combining developments in lexometric and statistical research with corpus technologies (http://textometrie.ens-lyon.fr/?lang=en) (Heiden, 2010). It allowed for a more granular analysis of language features, requiring the integration of a specific language model, the splitting of speakers, the conversion of data into computer readable XML language, and the lemmatization of the data.

4.2 Emotion recognition tools

Emotion recognition tools are often developed and used in the field of Social Signal Processing (SSP), where the goal is to investigate and develop machines that are socially intelligent; this implies that they are capable of recognizing and interpreting social and affective signals automatically (Vinciarelli et al., 2009). Tools exist that can extract characteristic speech parameters which can subsequently be used in machine learning software to find and learn new patterns (e.g., emotions) from this data. Applying these tools requires programming, so for social scientists and historians this is only feasible to apply in tandem with a computer scientist. What was key to convey to the participants, is the existence of this non-textual dimension of the recording. Emotion is indeed dealt with in the discipline 'discourse analysis', but not in a multimodal way. A way to profit from this tool, would be to identify particular emotions in an entire corpus, as a basis to go back to the interpretation of a single interview. Participants were first presented with the concept. During the session the use of the linguistic tool Praat (http://www.fon.hum.uva.nl/praat/) was demonstrated, showing the silences in a corpus and how these can be relevant for emotional expression analysis (Boersma, 2020). The hands-on component encouraged participants to get familiar with different speech features, for example, how to digitally detect and analyse these using a voice recording (Truong et al., 2013; Van den Heuvel & Oostdijk, 2016).

5 Pedagogical considerations for preparation and evaluation

Prior to the workshop, participants were invited to reflect on their own research trajectory and provide us with a short narrative of a typical research journey they had undertaken when working with interview data. Based on this homework, we assessed and visualized their workflows, and constructed a series of typical 'research trajectory' flow charts. This enabled us to come up with a high-level simplified trajectory and to identify how and where the digital tools might fit into the researchers' workflow. This is illustrated in Figure 4.

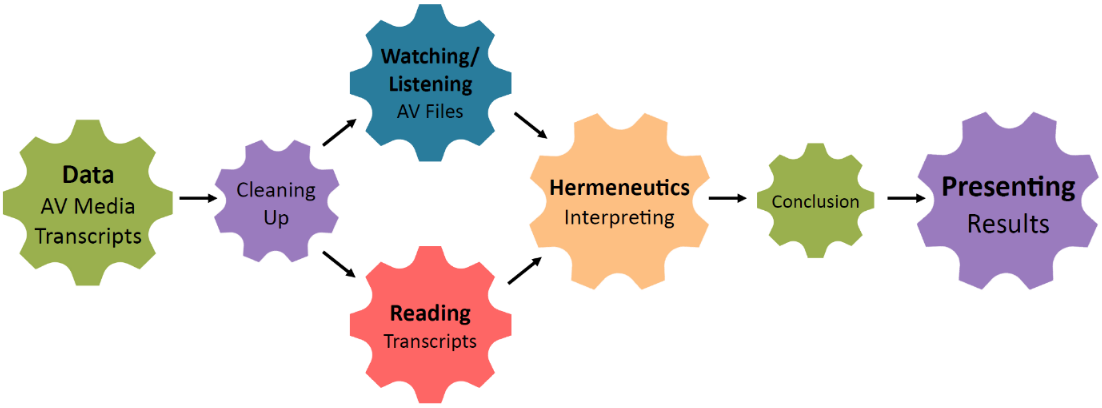


Figure 4: Poster showing typical research trajectory for scholars using interview data (cogs), annotated with specific tools that they worked on in the workshops.

Anticipating that the diversity of participants and tools would make the organization of the workshop complex, it was essential to follow principles in the design of the workflow that ensured 'satisfying user and research experiences'. To this end we took great care of ensuring a basic level for preparation: we gathered information on the participants on their level of digital shrewdness; prepared data familiar to the participants in both a common language (English) and in their native language; assigned homework in order to make participants become familiar with the tools; and ensured that a participant with advanced digital skills was present in each of the language groups. After each session, each language group was asked questions on their experience and their assessment of tools' strengths and weaknesses. We also asked them if they would use the tool in their own research and if so, in what way. Lastly, we asked them for suggestions that we could pass on to the tool's developers.

6 Evaluation of tools

Even before the workshops started, some participants struggled to download unfamiliar software prior to meeting, suggesting a lack of basic technical proficiency. This can turn out to be a significant barrier to the use of open source tools that often require a bit more familiarity with, for example, laptop operating systems. It was useful to have speech technologists sitting amongst the scholars, witnessing first-hand some of the really basic challenges in getting started.

6.1 Evaluation of the Transcription Chain

The first hurdle with which the participants were confronted was the slow pace of the ASR process for particular languages. This meant that not all participants were able to fully experience working with the OH-portal. Those who did get the chance to play around with correcting the ASR results, appreciated the simple design and usability of the T-Chain. One participant remarked that they were glad the tool did not require much technical know-how. A happy surprise was that the software catered to all native languages represented by the different groups.

With regard to criticisms, lack of transparency seemed to be the overarching theme for all groups. During the long waiting time for results to show, participants were bothered by having no insight in what was happening behind the scenes. Several participants expected more information about the ASR-engines in terms of accuracy and speed. The interface in some cases was unclear, for example, there were two options to choose as ASR-engines (Dutch NL and Dutch OH), but the difference between the two was not clear. Similarly, the OCTRA-editor included for correcting ASR results had a lot of shortcuts, but these were unclear to those unfamiliar with speech technology terminology.

Attendees also expressed concern about the issue of privacy in sources that were uploaded to the web-based system, indicating that this may well be a problem in using the OH portal in their own research. Are the data stored somewhere after the processing is completed? They further questioned who owned the ASR-engines. Will the ASR results be used for improvement of the engine? One

British participant remarked that without answers to these questions, it would be hard to justify using the tool in front of an academic research ethics committee.

When asked for possible improvements on the T-Chain, participants stated that more documentation, on the technical, ethical and legal side, was needed. They pointed to the need for dejargonisation of terms or brands (like OCTRA) and further description or help added. One participant had the idea of a quick list on the side of the OCTRA-editor, where all the shortcuts to actions could be listed. File limits were of course also a problem. Participants would have liked to be able to upload not just wave files of a limited length, but also mp3 and different video files. More options for file conversion and export were requested, such as automatic conversion to TEI-encoded documents. Lastly, one feature that a British researcher would appreciate, is diarisation: the automatic detection of a speaker change, as well as a way of visualizing it.

To conclude, most of the participants were intrigued by the general concept of a T-Chain with its speech-to-text and alignment software. Almost all of them saw ASR as a way of potentially easing the transcription process. With some improvements, participants could see themselves using this resource for this purpose, although a long period of acculturation would be needed. Some doubted whether they would use the included OCTRA editor for correcting results, and stated they would feel 'more comfortable' using an external tool or word processor for this purpose.

6.2 Evaluation of text annotation tools

Overall, the familiarity of annotation across disciplines made both NVivo and ELAN accessible to participants. But the vastly different terminology and user interface meant that users had to spend additional time acquainting themselves to the tool's unique layout before being able to annotate. What would help is a uniformity of language and terminology for features that all tools have in common. A useful step to take is the dejargonising of the interface, or creating custom user interfaces for different types of users.

The NVivo tool worked particularly well with written transcripts, and allowed users to actually see mark-up and notes in the context of a transcript. Being able to collate all documents related to a single research project proved to be a clear benefit of the tool, with one user commenting that ELAN had a much more visual display and worked solely with audio and video data sources. But the learning curve of NVivo was steep for all participants unfamiliar with the software. Many experienced it as overwhelming and hard to work with. Another issue was how closed-off the package is. With barely any useful export and import functions, it is nearly impossible to use NVivo in combination with other software. ELAN, on the other hand, had good support for importing and exporting files.

Some users named working with ELAN as a 'pleasant experience', as it had a clear and comprehensible interface (the same on Mac and PC, which NVivo has not). The concept of 'tiered' annotations was found interesting, and a useful way of visualizing codes/notes. With ELAN, codes and annotations are tiered and placed on a timeline, something which NVivo lacks. While ELAN was taken up more quickly by the participants, its focus on the particularities of linguistics was experienced as a hurdle. Specifically, users disliked the lack of the possibility to make a distinction between transcriptions and annotations/codes in the tiers. It seemed that transcribing in ELAN is less suited for the in-depth interviews, typically used in Oral History, than briefer interviews that are often analysed by (socio-)linguists.

All in all, it was hard for participants to imagine getting out of their 'comfort zone'. The amount of time needed to become familiar with the features of the tools, and to actually experience the benefits was too short. Some expressed interest in exploring the tool further while others were turned off by the idea of using such intricate tools in general. The reason is that a choice of a particular annotation tool leads to an engrained practice of research that cannot be easily traded for an alternative. The most open to change were those who already used existing similar software such as ATLAS.ti. An experienced participant pointed to the suitability of ELAN for research where multimodal annotation/coding was required, with video/audio and text in tandem. The less digital savvy researchers suggested offering functionalities with different levels of complexity: a Simple Mode and an Expert Mode. This would make simple actions a lot more accessible to researchers with second to no experience with digital research.

6.3 Evaluation of on-the-fly linguistic analysis tools

Whereas the introduction to gain insight in the generic linguistic tools and their shortcomings/opportunities was very much welcomed, the hands-on components were met with varied reactions. Overall, the participants enjoyed these tools, and referred to them as "easy-to-use", "simple", and "lightweight". *Voyant* and *Autosummarizer*, while not necessarily useful for drawing conclusions, inspired the participants to think about their own process of getting insight into a text and summarizing it. The limited amount of text that can be analysed was perceived as a barrier for the take up. There was an overall need to already be informed in a very concrete way about the added value for use of these tools by non-linguists. Exploring what the possibilities could be, proved to be demanding within the available time slot and only of interest to those interested in experimenting. What became clear is that sociolinguists may benefit from the use of the Voyant word frequency functionality.

Although the use of word frequency raises controversy within linguistics, it is widely accepted that frequent words may influence phonetic change, and also may act as 'locus of style' for a given speaker (Hay, Foulkes 2016, p. 324). At the same time, it seemed that Voyant was not sophisticated enough to process uncleaned transcriptions. Not everyone found the tools easy-to-use. Voyant, with its many different windows, was described by one researcher as "incredibly frustrating", another participant wanted to know how the word clouds were generated. Stanford NLP a tool for more 'heavy-duty' linguistic analysis, was hard for the oral historians to relate to.

6.4 Evaluation of a textometry tool

Among advanced digital scholars, TXM was the most liked tool of all. One participant described it as the "most complete and transparent tool used so far". The combination of both power and transparency gave the researchers lots of inspirations and ideas on how to use the software. Functionalities that were mostly appreciated were the Tree Tagger, the visualisation of concordances and co-occurrences, and the ease with which it was possible to get an overview of the complete corpus. Aside from the tool itself, the clear printouts of instructions in this session that each could follow at his or her own pace, were greatly appreciated. They formed a counterbalance to the complicated nature of TXM.

Interest was expressed for using co-occurrences and concordance in analysing both separate texts and sub-corpora. Some participants thought of this as a wholly new way of interacting with the data and discovering new aspects, while others thought it could prove or disprove 'hunches' they had about data, in that it is possible to quantitatively prove prevalence of certain words in text. A specific use for TXM that was mentioned was analysing differences in gender. TXM allows users to tag lets a user tag the gender of an interviewee and interviewer, which makes it possible to quantitatively analyse difference in the way these people speak, and making it a good tool to evaluate the entire corpus.

For most participants without experience with digital tools, the complexity of TXM was a hurdle. Some encountered problems with the terminology used in the program (e.g. what is a partition?). Several groups found the need to pre-process the data time-consuming and complicated. What could be improved of TXM is making the interface more approachable, adding more colour, turning the tool into a web-based service, and incorporating a glossary in which all the terms used in the software are explained. There were exceptions however: one user noted that "[TXM was] a bit of a struggle at first, but this helps you to do a close reading of an interview, and I think it fits perfectly within my traditional hermeneutical approach".

Overall, it appeared difficult to understand how to attribute meaning to the frequency of a particular term in the entire corpus of interviews, when being used to focus on the interpretation of a single interview. TXM can offer insights in features of the interview process in its entirety, such as: the relation between words expressed by interviewer and interviewee, the difference in active and passive use of verbs between gender, age or profession, or the specificity of certain words for a respondent. In some ways, this might require the scholar to temporally disregard the individuality of the person talking, and switch from close listening to interpretations in which scale and numerical relations can be relevant. This requires a widening of methodological perspective in data analysis.

6.5 Evaluation of emotional recognition tools

The session on emotion recognition was the most remarkable one. While the technology behind it was the most complex, the relevance of identifying emotional features was clear to everyone. Participants first of all highly appreciated the outline of the discipline of Social Signal Processing (SSP). Even though Praat, as a tool for computational linguistics, was perceived as complex, the presentation illustrated its power in a way that all participants could easily relate to it. Here as well, the step by step guide provided was appreciated. One participant observed that she struggled less with terminology in this last session. She felt accomplished in having become familiar with technical terms like "tiers". Still the step of connecting this dimension of the data to their own research practice was difficult to make. Experimenting with Praat was an enriching experience, but even after fully understanding how differences in pitch, speed and silences can be extracted from the data, they were doubtful on whether such applications could be integrated in social science or oral history research. How do you blend in insights with regard to scale, frequency and paralinguistic features into the classic interpretation of the interview data?

7 Conclusion

Interview data provide rich and promising information sources with which to engage in methodological interdisciplinary or multidisciplinary conversations. Our analysis of the user experience before, during and after the workshop suggests that scholars are open to cross-fertilization. At the same time, scholars are only willing to integrate a digital tool into their existing research practice and methodological mindset, if it can easily be used or adapted to their needs. The limited functionality of the free easy-to-use tools, and the observed methodological and technological complexity and jargon-laden nature of the dedicated downloadable tools, were both seen as significant barriers, despite the availability of clear documentation. Time investment is key to really grasp the essence of a tool. This means that addressing the right audience at the right stage of their career is crucial. It could also mean that we have to offer different trajectories with different levels of complexity to different types of scholars. With regard to assumptions we had about the potential of the tools to affect existing research practices - questioning the dogma of full verbatim transcription, exploring a collection with text mining tools, integrating the emotional features of an interview into an analysis - it is too early to be able to draw any conclusion. The workshops offered insights in new approaches to a broad group of researchers, but we have only scratched the surface of what is possible. A much longer and intensive engagement with data and tools in small multidisciplinary teams is necessary to test our assumptions. We intend to continue to publish the results of our endeavours through case studies, that can be selected from the multilingual archive of interviews on the topic of migration, that was created by the team for the workshops.

In terms of improving the take up among social scientists and historians who work with interview data, of what the CLARIN infrastructure has to offer, it is clear that jargon is an obstacle. More user friendly, well-documented and stable tools would be welcomed, especially in a form that makes it possible to skip the lengthy process of installing software. More specifically, it may be worth exploring the integration of open source annotation tools into the latter end of the T-Chain to enable a seamless experience, and thereby moving towards the ever-appealing concept of the workbench. Moreover, as in all ASR services, privacy concerns must be addressed so that users have a very clear understanding of what will happen to a source once uploaded to a tool. In this respect, an explicit GDPR-compliant data processing agreement may allay worries. A closer collaboration with the CLARIN Ethical and Legal Committee (CLIC) is therefore recommended. As different disciplines and tools for interview data use varying metadata schema, work needs to be done on mapping and crosswalks and export of marked up formats. Empowering the curators and publishers of interview data to get their collections 'analysis- and tool ready" is useful, as is encouraging them to also make use of text mining tools, such a term extraction, for enhancing resource discovery.

References

Armaselu, F., Danescu, E., Klein, F. 2019. Oral History and Linguistic Analysis. A Study in Digital and Contemporary European History. *Linköping Electronic Conference Proceedings* (159), 13.

Boersma, P., Weenink, D. 2020. Praat: doing phonetics by computer [Computer program]. Version 6.1.10, retrieved 23 March 2020 from http://www.praat.org/.

Boyd, D. (ed). 2013. Oral History in the Digital Age. *The Oral History Review*, 40(1): i-iii doi:10.1093/ohr/oht038.

Corti, L., Gregory, A. 2011. CAQDAS Comparability. What about CAQDAS Data Exchange? [42 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *12*(1), Art. 35, http://nbn-resolving.de/urn:nbn:de:0114-fqs1101352.

Corti, L., Fielding, N. 2016. *Opportunities From the Digital Revolution: Implications for Researching, Publishing, and Consuming Qualitative Research*. SAGE Open. https://doi.org/10.1177/2158244016678912.

De Jong, F.M.G., van Hessen, A., Petrovic T., Scagliola S. 2014. Croatian Memories: speech, meaning and emotions in a collection of interviews on experiences of war and trauma. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* http://www.lrec-conf.org/proceedings/lrec2014/index.html.

De Jong, F.M.G., Ordelman, R.J.F., Scagliola S. 2011. Audio-visual Collections and the User Needs of Scholars in the Humanities: a Case for Co-Development. *Proceedings of the 2nd Conference on Supporting Digital Humanities (SDH 2011).*

Freund, A. 2009. Oral history as process-generated data, *Historical Social Research*, 34 (1): 22–48.

Heiden, S. 2010. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In K. I. Ryo Otoguro (Ed.) *24th Pacific Asia Conference on Language, Information and Computation:* 389-398. Institute for DigitalEnhancement of Cognitive Development, Waseda University. Online: https://halshs.archives-ouvertes.fr/halshs-00549764.

Hay, J., Foulke, P. 2016. The evolution of medial /t/ over real and remembered time. *Language*, 2016, 92: 298-330.

Kisler, T., Schiel, F., Sloetjes, H. 2012. Signal Processing via Web Services: the Use Case WebMAUS. *Proceedings of DH2012*, 30-34, 2012, Hamburg.

Pömp, J., Draxler, Chr. 2017. OCTRA – A Configurable Browser-Based Editor for Orthograpic Transcription. *Tagungsband Der 13. Tagung Phonetik Und Phonologie Im Deutschsprachigen Raum*, 145-148. Berlin.

Portelli, A. 2006. What makes oral history different, in: R. Perks and A. Thomson, *The Oral History Reader* (New York) 36.

Sinclair, S., Rockwell G., 2016. *Voyant Tools*. Web. http://voyant-tools.org/.

Truong, K. P., Westerhof , G. J., Lamers, S.M.A., de Jong, F.M.G., Sools, A. 2013. Emotional expression in oral history narratives: comparing results of automated verbal and nonverbal analyses. *Proceedings of the Workshop on Computational Models of Narrative CMN 2013 Hamburg, Germany*.

Van den Berg, H., Scagliola, S., Wester, F. (eds) 2010. *Wat veteranen vertellen; verschillende perspectieven op verhalen over ervaringen tijdens militaire operaties (What veterans tell us; different perspectives on biographical interviews about experiences during military operations*, Pallas Publications. http://www.watveteranenvertellen.nl/.

Van den Heuvel, H., Draxler, C., Van Hessen, A., Corti, L., Scagliola, S., Calamai, S., Karrouche, N. 2019. A

Transcription Portal for Oral History Research and Beyond. *Proceedings DH2019, Utrecht,10-12 July 2019.* https://dev.clariah.nl/files/dh2019/boa/0854.html.

Van den Heuvel H., Oostdijk, N.H.J. 2016. Falling silent, lost for words ... Tracing personal involvement in interviews with Dutch war veterans. *Proceedings LREC2016, Portorož, Slovenia*: 998-1001, http://www.lrec-conf.org/proceedings/lrec2016/pdf/104_Paper.pdf.

Van den Heuvel, H., van Hessen, A., Scagliola, S., Draxler, C. 2017. Transcribing Oral History Audio Recordings – the Transcription Chain Workflow. *Poster at EU. Clarin Conference, Budapest, September 18/19- 2017. EU. Clarin Conference.*

Vinciarelli, A., Pantic, M., Bourlard, H. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing*, *27*(12): 1743-1759.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. 2006. ELAN: a professional framework for multimodality research. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. 1556-1559.

Integrated Language and Knowledge Resources for CLaDA-BG

Kiril Simov
LMaRK
IICT-BAS, Bulgaria
kivs@bultreebank.org

Petya Osenova
LMaRK
IICT-BAS, Bulgaria
petya@bultreebank.org

Abstract

This paper presents the envisaged integration of the language resources for Bulgarian with the knowledge sources like ontologies and linked open data to support their joint usage with respect to the cultural and historical heritage (CHH) objects. We started with the knowledge integration of the language resources for Bulgarian. Our plan is to continue with the addition of selected CHH objects to the initial integrated data. Based on the available Bulgarian resources like dictionaries and corpora as well as on the Bulgarian Wikipedia, DBpedia and Wikidata, we have constructed the first version of a Bulgaria-centered Knowledge Graph. It represents the conceptual information for the Bulgarian virtual infrastructure CLaDA-BG.

1 Introduction

Nowadays vast networks with linked objects are dominant in many areas of life, including tools and data in NLP. Among the many prominent initiatives in linking available data in various combinations are the following: CLARIN-ERIC[1] (the infrastructure that combines the strengths of the language resources and technologies), the Linked Open Data Cloud[2] (language resources with ontologies), the Predicate Matrix[3] (a lexical resource that integrates the information from different semantic and syntactic resources such as FrameNet, VerbNet, PropBank, WordNet), BabelNet[4] (a knowledge base with a strong multilingual value), PARTHENOS project[5] (integrates cloud storage with services and tools and support collaborative working on language and CHH data), SSHOC[6] (connecting existing and new infrastructures from the SSH ERICs), ELEXIS[7] (the European Lexicographic Infrastructure) and many others. All these projects and initiatives focus on the idea of linking. This means: linking tools, resources, architectures, knowledge bases and infrastructures. Our work has many lines in common with the mentioned projects. Similarly to CLARIN-ERIC and PARTHENOS, we aim at providing adequate language technology for Linguistic Studies, Humanities, Cultural Heritage, History and related fields, making them mutually understandable and coherent. As in Predicate Matrix, BabelNet and ELEXIS, we combine information from various resources - in our case these are BTB-WordNet, Valency dictionary, Wikipedia, Wiktionary. The differences can be summarized as follows: our focus is set particularly on Bulgaria-related data; apart from integrating resources, we rely on annotated biographical data from historians, librarians, museum workers; our semantic and encyclopaedic resources for Bulgarian do not have the coverage of English or German related ones. Thus, we rely on getting more knowledge through the cross-lingual mappings coming from wordnets, Wikipedia, etc. In the future we will investigate the

[1] https://www.clarin.eu/
[2] https://lod-cloud.net/
[3] http://adimen.si.ehu.es/web/PredicateMatrix
[4] https://babelnet.org/
[5] http://www.parthenos-project.eu/
[6] https://sshopencloud.eu/
[7] https://elex.is/

Kiril Simov and Petya Osenova 2020. Integrated Language and Knowledge Resources for CLaDA-BG. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 137–144.

possibility to incorporate in our work also ideas from VerbAtlas - di Fabio et al. (2019). VerbAtlas is a manually-crafted verbal semantic resource structured into frames. It groups semantically-coherent synsets from WordNet. It is the first resource enriched with semantic information about implicit, shadow and default arguments following Pustejovsky (1995). VerbAtlas aims at improving the main features of the existing verbal inventories (FrameNet, PropBank, VerbNet), while also adding new semantic information.

CLaDA-BG[8] is the Bulgarian National Interdisciplinary Research E-Infrastructure for Bulgarian Language and Cultural Heritage Resources and Technologies. In contrast to other EU infrastructures that started separately as CLARIN and DARIAH, and later on in some countries (Austria, the Netherlands, Greece and others) combined or started to work in a closer cooperation, in Bulgaria the joint infrastructure started as a joint endeavour from the beginning. In the spirit of European CLARIN and DARIAH, the mission of CLaDA-BG is to establish a national technological infrastructure of language resources and technologies (LRT), as well as cultural and historical heritage (CHH) resources and technologies in a connected framework. The consortium of CLaDA-BG comprises 15 organizations including research institutes at the Bulgarian Academy of Sciences, several universities, the National Library “Ivan Vazov” in Plovdiv, and two museums. Thus, the consortium does not include only technological partners, but also content providers and experts in history, library studies, arts.

The main goal of the infrastructure is to provide public access and an integrated version of the available resources and technologies for various societal tasks, targeted at a wider audience. The infrastructure aims to support primarily researchers in Art, Humanities and Social Sciences to process Bulgarian language texts and CHH datasets necessary for their research. However, the real applications are envisaged to go beyond the research framework, since many areas can profit from linked knowledge. Thus, the results will be applicable also to education and industry.

Needless to say, linking data in a broader sense is a challenge. First of all, due to the fact that data itself is diverse. Second, because these data exist in various formats and representations. Third, different tools are necessary for manipulating the data. Last but not least, the data has to be made to communicate to each other, since it supplies similar or different pieces of knowledge that might contradict or remain incomplete thus leading to misunderstanding or wrong assumptions.

For all the reasons mentioned above, we focus on putting the varying types of data into the context of each other. The approach for interlinking of the data is called *contextualization*. The different types of objects of study, representation and search are integrated on the basis of common metadata categories and via textual descriptions. The language resources and the textual descriptions of other objects are integrated with the help of a common Bulgaria-centred knowledge graph - *BGKG*. Thus, the language description has become the main brick for creating the knowledge graph. We also plan to integrate links to images and digitized/3D-scanned objects.

The existing open data for Bulgarian, such as Wikipedia, DBpedia and Wikidata[9] are still scarce and/or not completely reliable. For that reason, we provide a) linking with our in-house lexicons and corpora, and b) gather data from our content providing partners that are of high quality.

In this paper we present the core sets of language resources that in our view are necessary to support research in social sciences and humanities. We also show how they are integrated in order to support the semantic annotation of texts with conceptual information from the knowledge graph with the aim to ensure: extraction of new knowledge from text, querying over the knowledge graph, and indexing of texts within the CLaDA-BG repository.

2 Integrated Bulgarian Language Resources

Since it was decided that language data and language descriptions of library/museum objects will be the connecting parts within the CLaDA-BG, we have to ensure the necessary framework and technology. For the necessary language resources set as a prerequisite we rely on the Basic Language and Resource Kit – BLaRK – Krauwer (2003). Below is the initial list of these language resources. It has to be noted that this data has been available for Bulgarian for years, but they have to reach the designated size in the

[8] http://clada-bg.eu/

[9] Here we consider the part of Wikidata that is equipped with Bulgarian labels.

brackets and to become easily searchable on the web and within the CLARIN repository. The basic language resources are:[10]

Corpora

- Text Archive for Bulgarian (minimum 100 million running words);
- Morphologically Annotated Corpus (1 million running words);
- Syntactically annotated corpus (1 million running words);
- Semantically annotated corpus with ontological and fact information (1 million running words);
- Domain corpora (minimum 100 000 running words per domain)

Lexicons

- Bulgarian Wordnet (BTB-WN) (50 000 synsets of coverage of the lemma senses in a related semantically annotated corpus)
- Valency lexicon (coverage of the verbs in BTB-WN)
- Domain dictionaries (minimum 100 000 running words per domain)
- Representative lists of Bulgarian names (coverage of the names of the public figures, location and organization names. Additionally, they will include relations to the Bulgarian Wikipedia)

The language processing tools include minimally the following ones: morphological, shallow syntactic, deep syntactic, and semantic analyzers, named entities recognition and identification modules.

During the first year of the CLaDA-BG project we focused on the integration of the various existing language resources and performed only minimal extension in order to make them usable.

As a basis for the manually annotated corpus we use the texts included in BulTreeBank - an HPSG-based treebank for Bulgarian - comprising about 260 000 running tokens. These texts were annotated before the start of CLaDA-BG with senses from BTB-WN and instances from DBpedia, URLs from WikiPedia and classes from DBpedia ontology (see Popov et. al, (2014)). The original annotation is an HPSG-based constituent structure with marked up the head in each phrase. It was converted automatically to a dependency format. The dependency annotation follows the Universal Dependency guidelines.[11] The original Treebank is also manually annotated on morphosyntactic level with a rich tagset (680 tags – Simov et al. (2004)) and lemmas.

The BTB-WN currently contains 22 000 synsets which cover all the words within BulTreeBank and most frequent words over the Bulgarian national reference corpus (about 100 million running words). We started with the extension of the information within BTB-WN by adding inflectional paradigms to each lemma in the synsets and with their mapping to articles from the Bulgarian Wikipedia – see Simov et al. (2019). The inflectional information is important because many lemmas in BTB-WN can belong to different inflectional paradigms. The information from Bulgarian Wikipedia provides not only additional encyclopedic information for the named entities, but also a terminological one. During the process of mapping BTB-WN to Wikipedia, new senses have been added to the lexical resource. In addition, the mapping to Wikipedia provides a source for new relations between the synsets in BTB-WN. On the basis of the current synsets in BTB-WN, we extracted about 13 000 Wikipedia articles. These articles were manually inspected and mappings between the synsets and the articles were established. The mapping follows the approach of McCrae (2018). In addition to the Wikipedia articles that correspond to the synsets in BTB-WN, we selected and extracted 10 899 Wikipedia articles that relate to the names in a Bulgarian gazetteer. This gazetteer represents the most important names in the Bulgarian National Reference Corpus. From them 1 515 pages were already extracted on the basis of the lemmas within BTB-WN. The remaining 9 384 pages were classified as Bulgarian locations, other locations, people, organizations, and other. In this way we extend BTB-WN with important information for Bulgaria named entities.

[10] In brackets we put the desirable minimal size of the corresponding resource that would make it applicable in many areas of usage.

[11] https://universaldependencies.org/

The creation of the Valency lexicon started by generalization over the annotations within BulTreeBank. After the extraction of verbs with their arguments from the treebank we classified the verbs by their senses within BTB-WN and then the arguments were also mapped to the corresponding synsets. Thus, one syntactic frame could result in several semantic frames. Here is one example on Fig. 1:

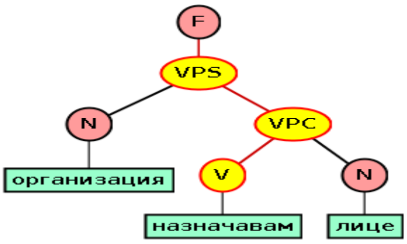


Figure 1. An example of a generalized syntactic frame.

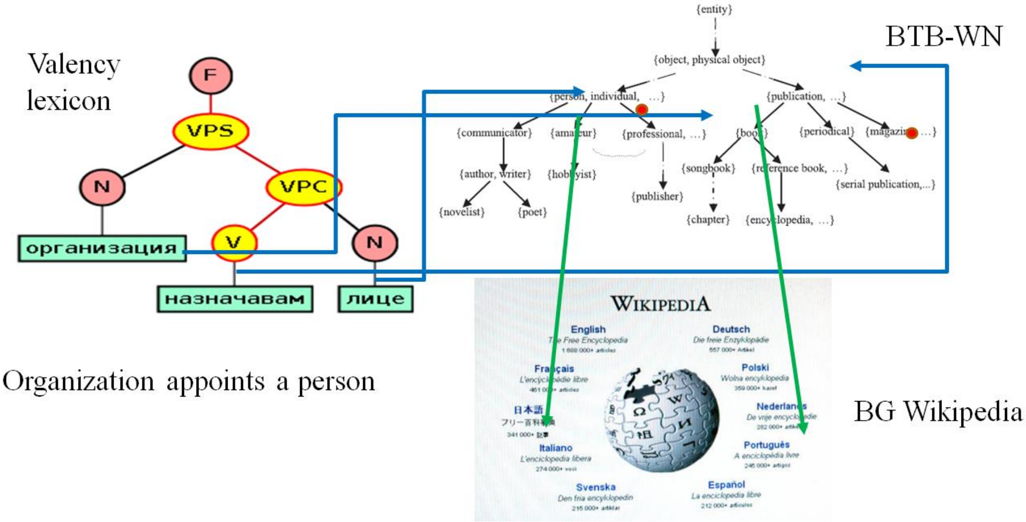


Figure 2. Mapping between the Valency lexicon, BTB-WN and BG Wikipedia.

It represents the case when an *organization* (subject noun - N) *appoints* (V) *a person* (object noun - N). During the first phase we did not modify this lexicon, but focused on ensuring the connection of each verb with the corresponding frame. This was possible since the treebank was annotated with senses from BTB-WN and the frames were originally extracted from the treebank. Later on, a number of missing senses were added as well. The rich annotation of the treebank allows it to be used for the training of machine learning techniques that assign the correct frame for each verb depending on the context. In the next phases of CLaDA-BG we plan to extend the Valency lexicon to also cover the verbs within BTB-WN. On the other hand, the mapping between BTB-WN and encyclopedic knowledge ensures a mapping between the Valency lexicon and encyclopedic knowledge. Currently, the encyclopedic knowledge is being extracted from Wikipedia, DBpedia, Wiktionary as well as expertise data (biographies of important Bulgarian people, descriptions of significant events, etc.) but during the next phases of the project it is envisaged to cover the whole knowledge graph. Thus, on the lexical level we have the mappings between the Valency lexicon, the Bulgarian Wordnet (BTB-WN) and Bulgarian Wikipedia as depicted on Fig. 2 above.

On the corpora level all the information needed for an end-to-end representation was integrated: tokens, grammatical annotation, lemmatization, syntax, word senses and Named Entity categories. Fig. 3 shows an example of a sentence annotated with senses from the Bulgarian Wordnet BTB-WN and URLs from Bulgarian Wikipedia on top of the morphosyntactic analysis.

The sentence is: "*Водещ на купона беше Тома Спространов.*" ("The host of the party was Toma Sprostranov.") The two open class words are connected with the respective synsets from BTB-WN, represented here by their definitions. The word "*Водещ*" ("The host") is a participle of the verb "*водя*" ("to organize") and is annotated with that sense of the related verb. From the fact that it is a participle, present tense, it follows that the word denotes the person who is organizing the event. The word "*купона*" ("the party") is connected to the definition "*Организирано увеселение* …" ("Organized entertainment ..."). The host was the DJ Toma Sprostranov who has a Wikipedia page:

`https://bg.wikipedia.org/wiki/Тома_Спространов`

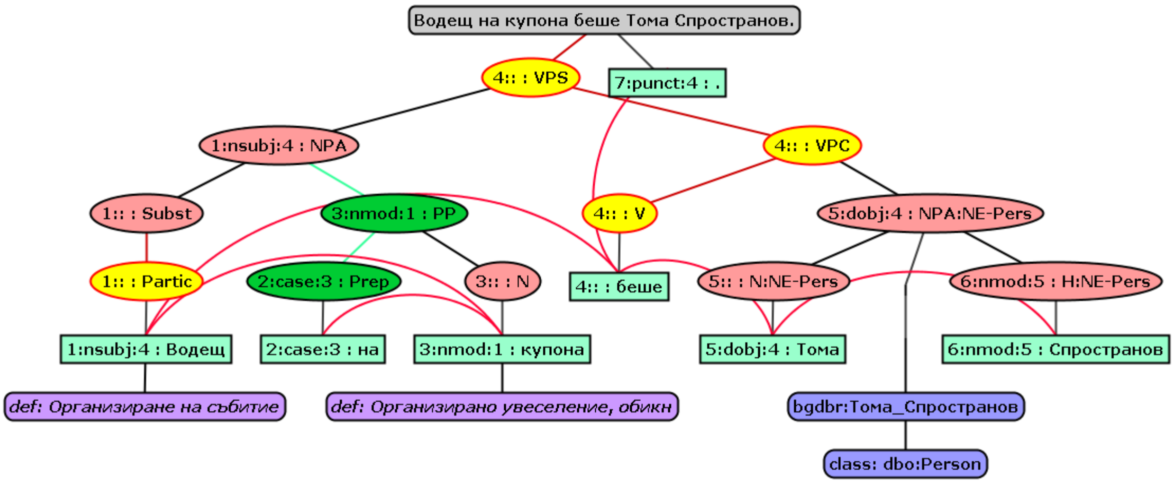


Figure 3. An example of a sentence annotated with morphosyntactic and semantic annotation.

In the image we used the namespace `bgdbr:` defined at `http://prefix.cc/bgdbr`. In the cases when there is no Wikipedia page for the corresponding named entity we add only a class from the DBpedia ontology, such as `Person`, `Politician`, `Musician`, `Country`, `City`, `Document`, etc. These annotations provide access from the treebank to the knowledge within BTB-WN and Wikipedia (later from the knowledge graph). Thus, the users of these integrated resources have access to the knowledge not only in the annotated corpus, but also within the lexicons and encyclopedic resources. The annotation of all these language levels over the same text documents provides a good basis for the widely used end-to-end neural models.

The integration of the language resources will be used at least in two directions (1) training of a wider set of processing modules, and (2) contextualization through the relations from the text to the encyclopaedic information. The latter is considered very important for the connection between language processing and suitable information extraction from textual descriptions of cultural and historical objects.

The integration of language resources and encyclopaedic knowledge is the first step in the direction of constructing a knowledge graph for CLaDA-BG aligned to language resources for Bulgarian.

3 Towards a Bulgaria-Centric Knowledge Graph

We aim at creating a semantically integrated environment for maintaining possibilities of referring to texts and descriptions of cultural or historic objects. For this purpose, the texts and descriptions of collections should be first annotated with an appropriate ontology, and then the annotation should be uploaded into an RDF repository. The first version of BGKG is based on the Bulgarian DBpedia knowledge graph. In the process of implementation of CLaDA-BG we will gradually add knowledge from other sources. Besides Wikipedia and DBpedia we envisage the inclusion of Wikidata as part of the initial knowledge graph. The integration of these sources of knowledge is guaranteed by their design. Wikidata as a knowledge source is considered with a higher level of quality because it follows rigorous rules for the construction and manual inspection phases.

As one step in the process of doing research within social sciences and humanities (SSH) we consider the identification of information of interest and its simultaneous observation within the same context. In order to support the research within SSH, CLaDA-BG needs to provide management of information of a huge variety of research objects including different kinds of texts (various genres, domains, time periods), artefact models, art masterpieces representations and descriptions, etc. The top unification of this data is the metadata,of course, but in fact very little common information can be represented in this way. In order to escape from the problem, we consider a new layer of information between the metadata and the actual datasets within SSH. Fig. 4 depicts the architecture we want to achieve.

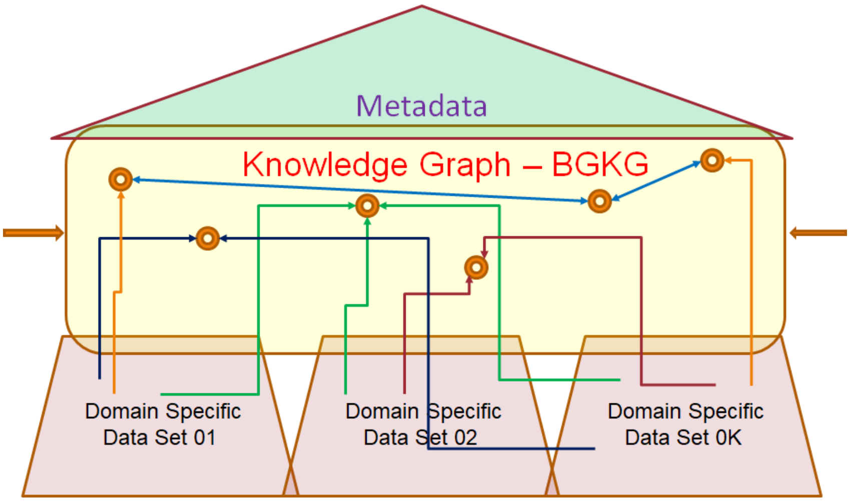

Metadata
Knowledge Graph – BGKG
Domain Specific Data Set 01
Domain Specific Data Set 02
Domain Specific Data Set 0K

Figure 4. The knowledge graph abstract architecture.

The newly introduced layer comes between the metadata layer where information about the objects is represented within the domain datasets. In many cases the researchers that use metadata to find the necessary information have to know additional information about the content of the domain data. Also, they need to find accordingly this information in the domain datasets. In order to provide such a functionality, we propose to construct the *Bulgarian-Centric Knowledge Graph* (BGKG) for supporting access to heterogeneous datasets. The approach for interlinking of these data was named *contextualization*. The aim of the integrated language resources is to provide a language layer for accessing the BGKG.

The main characteristics of the contextualization are time and space – what events happened at the same time or in the same space. The additional characteristics include also: the participants in a given event; the similar constructions of physical objects like form, size, material; the similar style of representation in images, text and sounds; the same school of production; etc. Thus, our motto is: *Everything in our world is connected and appears in a context*. The knowledge graph is a network of interlinked descriptions of people, events, geographical entities, objects, documents, authors, opinions, etc.:

- People – biographical data – events in their life, their roles
- Geographical entities – history of cities, etc.
- Objects – creation, materials, form, discovery
- Events – place, time, participants, connection to other events
- Documents – authors, contents, opinion about peoples, events, …

As already mentioned above, we consider the text as the main source of information for the represented objects. Linked Open Data and knowledge graphs were chosen for its representation Ehrlinger and Wöß (2016).

We first started with the existing DBpedia knowledge graph, constructed on the basis of Bulgarian Wikipedia. This knowledge graph is used currently in two ways: (1) for the semantic annotation of a huge web-based corpus; and (2) as an initial source of identifiers and facts for the construction of BGKG.

For the application in (1) we rely on the existing NLP pipeline for Bulgarian to annotate the documents within the web-based corpus mentioned above with Named Entities and identifiers from the knowledge graph. This will allow searching via entities from the knowledge graph.

The usage (2) of the initial knowledge graph is to support the creation of BGKG. The approach we selected relies on knowledge extraction from text. Thus, we started the creation of a semantically integrated environment for maintaining possibilities of referring to texts and descriptions of cultural or historic objects. For this purpose, the texts and descriptions of cultural or historical collections should be annotated with an appropriate ontology, entity identifiers and then the annotation should be converted to RDF triples and uploaded into an RDF repository. The whole process includes the following steps:

- Selection of appropriate ontologies
- Mapping of the ontologies to the integrated language resources
- Semi-automatic annotation of domain documents with Named Entities, their identifiers, concepts, relations between them
- Extraction of RDF triples from the annotations
- Manual assessment of the extracted facts and adding them to BGKG

The selection of the ontologies to be used in the creation of BGKG depends on the data available to the partners within CLaDA-BG. Each selected ontology will be aligned to the Bulgarian Wordnet BTB-WN. This will provide a better understanding of the ontology through appropriate lexicalizations. The classes of the ontology will be aligned to the synsets within BTB-WN. The properties will be aligned to triples of synsets (one or more triples of this kind) where the properties correspond to event synsets (usually verbal, but noun and adjectival synsets are also possible) and the subject and object of the properties will be mapped to the relevant synsets. In this way, the related properties will be mapped to the same event. For example, properties like `date-of-birth`, `place-of-birth`, `mother-of`, and `father-of` will be mapped to the synset for the event `birth`. The alignment of the ontology to BTB-WN will give the possibility of automatic processing by the available NLP pipelines. The actual integration could also require additions to the inflectional lexicons for the new words as well as annotation of new texts.

The annotation of new texts will be done semi-automatically, thus including human inspection. Human intervention will undergo various changes during the annotation process. At the beginning it will be during the entities annotation, the identification selection, and the relation annotation. When there are enough manually annotated documents and a subsequent improvement of the automatic annotation, the human attention will be directed to the extracted facts for the knowledge graph.

The knowledge graph will be available via a search tool. The search tool will provide the following search possibilities: a) concept search; b) facet search (integrating several concepts) and c) combined search (integrating concepts with random key words). This will ensure similar search for mentions of conceptual information in the tests and in the semantic description of the cultural and historical objects. The inclusion of the language, cultural and historical information into a common knowledge graph will provide one of the main mechanisms to support the research through the contextualization of each object of analysis.

4 Conclusion

In the paper we present the ongoing development of language and semantic resources within CLaDA-BG to support research in Humanities and Social Studies. This is done through the exploration of text corpora and the description of cultural and historical objects. In the area of language resources, the integration of Bulgarian language resources through BTB-WN and the Bulgarian Wikipedia provides a basis for training of text indexing with instances and classes from a Knowledge Graph. The actual knowledge graph is based on DBpedia and Wikidata (including also information from Wikipedia). Besides the textual information, descriptions of cultural and historical heritage objects are expected to be mapped to the knowledge graph as well. This step will allow a joint search including a SPARQL endpoint. When developed enough, the knowledge graph will be provided freely for download as a linked open dataset.

The initial knowledge graph will be further extended by specific ontologies for modelling the specific classification schemata, or time and space, events, facts.

Acknowledgements

This research was partially funded by the *Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH* – CLaDA-BG, Grant number DO01-272/16.12.2019.

References

Lisa Ehrlinger and Wolfram Wöß, W. 2016. *Towards a Definition of Knowledge Graphs*. SEMANTICS 2016: Posters and Demos Track. September 13-14, 2016, Leipzig, Germany

Andrea Di Fabio, Simone Conia and Roberto Navigli. 2019. *VerbAtlas: a Novel Large-Scale Verbal Semantic Resource and Its Application to Semantic Role Labeling.* Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), Hong Kong, China, November 3-7, 2019, 627–637.

Steven Krauwer. 2003. *The Basic Language Resource Kit (BLARK) as the first milestone for the language resources roadmap*. In Proceedings of the 2nd International Conference on Speech and Computer (SPECOM2003), 8–15.

John P. McCrae. 2018. *Mapping WordNet Instances to Wikipedia*. Proceedings of the 9th Global WordNet Conference (GWC 2018), 62–69.

Alexander Popov, Stanislava Kancheva, Svetlomira Manova, Ivajlo Radev, Kiril Simov and Petya Osenova, 2014. *The Sense Annotation of BulTreeBank*. Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13), 2014, 127–136.

James Pustejovsky. 1995. The Generative Lexicon. MIT Press, Cambridge MA.

Kiril Simov, Petya Osenova, and Milena Slavcheva. 2004. *BTB-TR03: BulTreeBank Morphosyntactic Tagset*. BulTreeBank Project Technical Report № 03. 2004.

Kiril Simov, Petya Osenova, Laska Laskova, Ivajlo Radev, and Zara Kancheva. 2019. *Aligning the Bulgarian BTB WordNet with the Bulgarian Wikipedia*. Proceedings of the 10th Global WordNet Conference. Wroclaw, Poland, 290–297.

Topic Modelling Applied to a Second Language: A Language Adaptation and Tool Evaluation Study

Maria Skeppstedt[1], **Magnus Ahltorp**[1], **Kostiantyn Kucher**[2],
Andreas Kerren[2], **Rafal Rzepka**[3,4], **Kenji Araki**[3]
[1]The Language Council of Sweden, the Institute for Language and Folklore, Sweden
{maria.skeppstedt,magnus.ahltorp}@isof.se
[2]Department of Computer Science and Media Technology, Linnaeus University, Växjö, Sweden
{kostiantyn.kucher,andreas.kerren}@lnu.se
[3]Faculty of Information Science and Technology, Hokkaido University, Sapporo, Japan
{rzepka,araki}@ist.hokudai.ac.jp
[4]RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan

Abstract

The Topics2Themes tool, which enables text analysis on the output of topic modelling, was originally developed for the English language. In this study, we explored and evaluated adaptations required for applying the tool to Japanese texts. That is, we adapted Topics2Themes to a language that is very different from the one for which the tool was originally developed. To apply Topics2Themes to Japanese texts, in which white space is not used for indicating word boundaries, the texts had to be pre-tokenised and white space inserted to indicate a token segmentation. Topics2Themes was also extended by the addition of word translations and phonetic readings to support users who are second-language speakers of Japanese. To evaluate the adaptation to a second language, as well as the reading support, we applied the tool to a corpus consisting of short Japanese texts. Twelve different topics were automatically identified, and a total of 183 texts representative for the twelve topics were extracted. A learner of Japanese carried out a manual analysis of these representative texts, and identified 35 reoccurring, fine-grained themes.

1 Introduction and background

Topic modelling provides a means of extracting a relevant subset of texts from a document collection that is too large to make a fully manual analysis of all its texts feasible. The extracted texts are organised into groups by the topic modelling algorithm, each group corresponding to an automatically detected topic that occurs frequently in the document collection (Blei et al., 2003; Blei, 2012; Jelodar et al., 2019). In addition to being associated with a group of extracted texts, the topics detected are also represented by a list of terms that are associated with the topics. This ability to extract and topically sort relevant texts in an unsupervised fashion has been used to perform qualitative text analysis in social science and humanities research (Baumer et al., 2017).

There are several tools for visualising topic modelling output, for instance with the focus on assessing and improving the quality of the topic model produced (Chuang et al., 2012; Lee et al., 2012; Choo et al., 2013; Hoque and Carenini, 2015; Lee et al., 2017; Cai et al., 2018; Smith et al., 2018), and with the focus on supporting the user in exploring and interpreting the texts included in the document collection (Alexander et al., 2014). A popular topic modelling visualisation approach is, for example, to display the topics and their associated texts or terms in a grid, and to use visual markers such as circles of different sizes and colours to indicate the level of association between a topic and a text or term (Chuang et al., 2012; Alexander et al., 2014).

Maria Skeppstedt, Magnus Ahltorp, Kostiantyn Kucher, Andreas Kerren, Rafal Rzepka and Kenji Araki 2020. Topic Modelling Applied to a Second Language: A Language Adaptation and Tool Evaluation Study. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 145–156.

The output of topic models, in the form of an automatic selection of subsets of texts and terms from a large text collection, has been shown useful for speeding up and facilitating qualitative text analysis (Baumer et al., 2017). Previous research has, however, also demonstrated that relying only on extracted terms—without also analysing extracted texts—has led to misunderstandings regarding the content of the text collection (Lee et al., 2017). In addition, there is not always a one-to-one correspondence between (i) the topics automatically extracted by the topic modelling algorithm, and (ii) what the user identifies as interesting, reoccurring categories of information when analysing a text collection (Baumer et al., 2017). Baumer et al. compared two methods for extracting reoccurring information in 2,190 free-text survey responses: (i) the use of topic modelling for selecting reoccurring topics, and (ii) a fully manual approach, in which a grounded theory-based analysis was applied. For the manual approach, all survey response texts were analysed in the search for what the authors call *themes*, i.e., categories formed by reoccurring information found in the texts. When comparing the output from the topic modelling and the grounded theory-based analysis, it could be concluded that the "topic modeling results captured to a surprising degree many of the themes identified in grounded theory, and vice versa." However, topics produced by the topic modelling algorithm often corresponded to several of the themes detected in the manual analysis, and some of the manually detected themes could be associated with several topic modelling-produced topics. With the aim of helping the user deal with this possible difference in granularity between automatically detected topics and manually detected themes, we have previously developed the Topics2Themes visualisation tool. The tool facilitates a manual search for themes, among the texts selected by the topic modelling algorithm (Skeppstedt et al., 2018a; Skeppstedt et al., 2018b).

With the Topics2Themes tool, we have thereby expanded the functionality typically provided by previous tools. The user can not only explore and interpret the automatically extracted topics and texts, but also add, and subsequently explore, an additional layer of analysis. This is carried out by enabling the creation of user-defined themes that can be associated with the texts extracted by the topic modelling algorithm. These user-defined themes and their text and topic associations, as well as their associations to automatically extracted terms, can then be explored in the tool. Thereby, an overview of the text analysis can be obtained, in which the automatically extracted information is integrated with the output of the manual analysis performed by the user. We also provide functionality for including metadata in the topic model visualisation in the form of text labels. The labels are either static or take the form of dynamic text labels that can be changed by the user.

We originally created Topics2Themes for English texts. Despite the unsupervised nature of the topic modelling algorithm, which makes the functionality of Topics2Themes fairly language-independent, it is not self-evident that the tool can be applied as-is to text written in a language that is typologically very different from English. To investigate this, we applied the tool to texts written in Japanese, i.e., a language that is both morphologically and orthographically different from English.

In addition, we envisioned the situation in which the text analysis of the Japanese texts would be performed by an analyst that would require some level of language support for fully understanding the texts. Such a situation would most naturally occur in a language learning situation, i.e., a situation in which the interaction with the texts is the primary reason to use the tool, and the output of the analysis is only of secondary importance. This situation could, however, also occur in the case in which a second-language speaker needs an understanding of the important content of a document collection, without having the means of employing the help of a more proficient speaker of the language. With the situation of a language learner in mind, we incorporated a system into Topics2Themes that helps second-language speakers of Japanese to understand Japanese text. We are not aware of any previous tools that combine the possibilities of using topic modelling for extracting and sorting the most relevant information from large text collections, with the functionality of providing reading support for language learners.

We here describe (i) the adaptation of the Topics2Themes tool to Japanese and to the situation in which the tool would be used by a second-language learner of Japanese, and (ii) the evaluation of the adapted tool on a Japanese document collection. The study has resulted in a new, language-independent version of the Topics2Themes tool, in which reading support can be provided. General usability issues, detected when the tool was evaluated on Japanese texts, were also corrected in the new version of the tool.

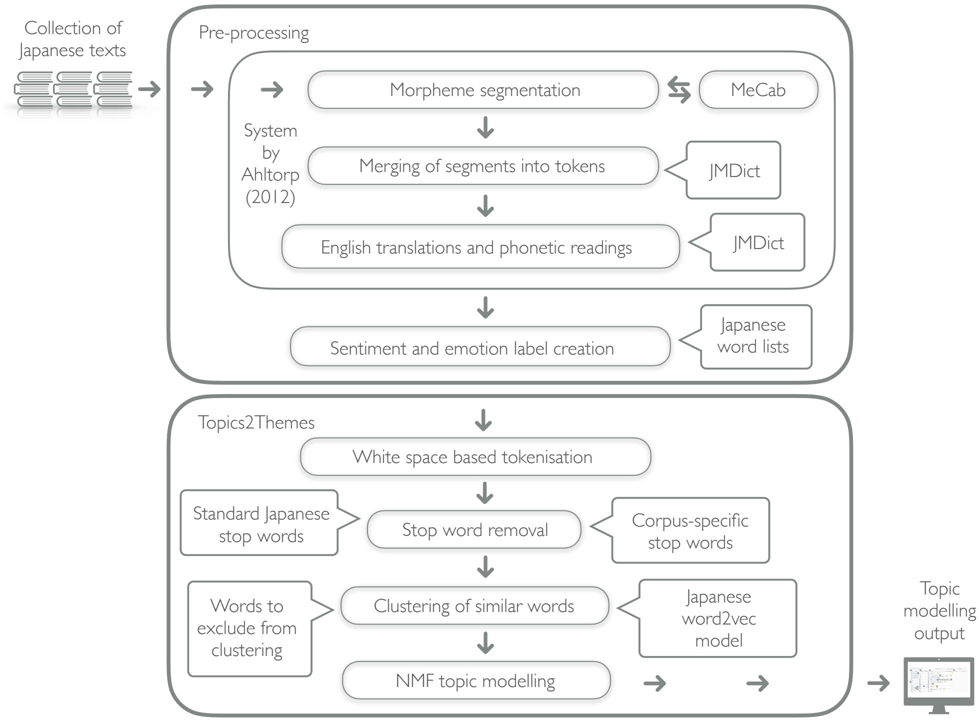

Collection of Japanese texts
Pre-processing
Morpheme segmentation
MeCab
System by Ahltorp (2012)
Merging of segments into tokens
JMDict
English translations and phonetic readings
JMDict
Sentiment and emotion label creation
Japanese word lists
Topics2Themes
White space based tokenisation
Standard Japanese stop words
Stop word removal
Corpus-specific stop words
Words to exclude from clustering
Clustering of similar words
Japanese word2vec model
NMF topic modelling
Topic modelling output

Figure 1: The components of the pre-processing and of the Topics2Themes tool adapted to Japanese. The language-specific parts consist of the entire pre-processing functionality, of the three word lists used (stop word lists, and words to exclude from clustering), and of the word2vec model used by Topics2Themes.

The development of Topics2Themes[1] was initiated by research funding from the Swedish Research Council, and the adaptation to Japanese and the evaluation on Japanese texts was funded by the Japan Society for the Promotion of Science. The updated version of the tool, in which usability issues were corrected, was developed within the Språkbanken and SWE-CLARIN infrastructures.

2 Method

The adaptation to Japanese consisted of adding an additional step in the process of using Topics2Themes, in the form of a pre-tokenisation of the texts before they were imported into the tool. Topics2Themes was also configured to use Japanese stop words and a word2vec (Mikolov et al., 2013) model trained on Japanese in the topic modelling process, as well as Japanese word lists for adding automatic labelling of the texts. Finally, a system for reading support for second-language speakers was incorporated.

We thereafter applied the adapted tool to a Japanese corpus, and the texts extracted by the topic modelling algorithm were then manually analysed by a learner of Japanese.

2.1 Adaptation to Japanese and the addition of reading support

Topics2Themes uses a very simple tokenisation based on the occurrence of white space. As white space is not normally used in Japanese to indicate word boundaries, another tokenisation technique is needed. We decided not to change the tokenisation method built into Topics2Themes, but to instead require the texts imported into the tool to be pre-tokenised and white space inserted into the texts to indicate token segmentation. The tokenisation included in Topics2Themes could therefore be used as-is.

For the pre-tokenisation, we segmented the text into morphemes using the MeCab tool (Kudo, 2006), and then merged morphemes into tokens by matching them to the JMDict dictionary (JMdict, 2013), as implemented by Ahltorp (2012).

[1] The code for the Topics2Themes tool is available at: `https://github.com/mariask2/topics2themes`, and the code for the Japanese pre-processing at: `https://github.com/mariask2/T2T_pre-processing_ja`.

The Topics2Themes tool can be configured to apply DBSCAN (Ester et al., 1996) clustering on word2vec vectors that correspond to the words in the corpus. Words belonging to the same cluster can thereby be collapsed into one concept, before the text is submitted to the topic modelling algorithm. The maximum distance between two words for them to be counted as the same concept can be adjusted by the user. That is, a large maximum distance allows for not only synonyms and different morphological instantiations of the same concept to be clustered together, but also creates groups in the form of semantically related concepts. To be able to perform the clustering on Japanese, we configured Topics2Themes to use vectors from a word2vec model[2] that had been trained on Japanese texts. The texts had been segmented by MeCab, and the segments had been merged into tokens with the help of a dictionary. Further on, a list of 111 words to exclude from the automatic clustering was manually created, since the clustering grouped these words together with semantically distant ones.

We also configured the tool to use Japanese stop words. Firstly, we used a Japanese stop words list available online[3]. This list was then extended by adding 150 frequent Japanese non-content words that occurred in the corpus to which the tool was applied.

For reading support, we incorporated a system constructed for Japanese language learning that provides a ranked list of English translations for each token included in the text, as well as a phonetic reading (*furigana*) for each Japanese *kanji*[4] character in the text. This tool has been developed for, and evaluated on, beginner learners of Japanese as well as learners on an intermediate level[5] (Ahltorp, 2012).

Topics2Themes was extended to use the *ruby*-tag provided in HTML to display the phonetic reading and the top-ranked English translation in a small font above each token. In addition, when the user hovers the mouse over a token, all available English translations are shown in the form of a tooltip. The functionality provided is not specific to Japanese. Instead, the extended version of Topics2Themes will provide this kind of reading support to any input text that indicates translations and/or phonetic readings using the same HTML-format.

In an attempt to further help the reader to understand and analyse the texts, we also created metadata labels by matching the texts to Japanese sentiment and emotion word lists (Nakamura, 1993; Takamura et al., 2005; Rzepka and Araki, 2012; Rzepka and Araki, 2017). Texts that contained words present in the lists were given static labels to indicate in which list they were present, and the sentiment and emotion words present in the text were also marked with a green or red background, for positive or negative words, respectively.

Figure 1 gives an overview of the components of the pre-processing and of the resources required to run Topics2Themes on the Japanese text collection.

2.2 Application of the adapted tool to a Japanese corpus

We applied the extended version of Topics2Themes to a corpus consisting of around 1,000 microblogs[6] collected with the criterium that they should contain the same content written in Japanese and in English (Ling et al., 2014). The tool was applied to the Japanese part of the microblogs, and the English part of the texts was not used in this study.

We configured Topics2Themes to try to find 15 topics using the NMF (non-negative matrix factorisation) topic modelling algorithm (Lee and Seung, 2001). The tool was further configured to run the topic modelling algorithm 100 times on the text collection, and to only keep topics that were stable enough to occur in all re-runs. This resulted in 12 stable topics being identified. The most prominent among those can be seen in the *Topics* panel in Figure 2, where each topic is represented by its three most closely associated terms.

[2] https://github.com/shiroyagicorp/japanese-word2vec-model-builder

[3] https://github.com/stopwords/japanese-stopwords/blob/master/data/japanese-stopwords.txt

[4] The logographic Chinese characters adapted to and used in Japanese.

[5] More specifically, the levels A1–B1 according to the Council of Europe CEFR levels.

[6] The corpus used is listed as a CLARIN resource at: https://www.clarin.eu/resource-families/parallel-corpora, and is also available at: http://www.cs.cmu.edu/~lingwang/microtopia/#twittergold.

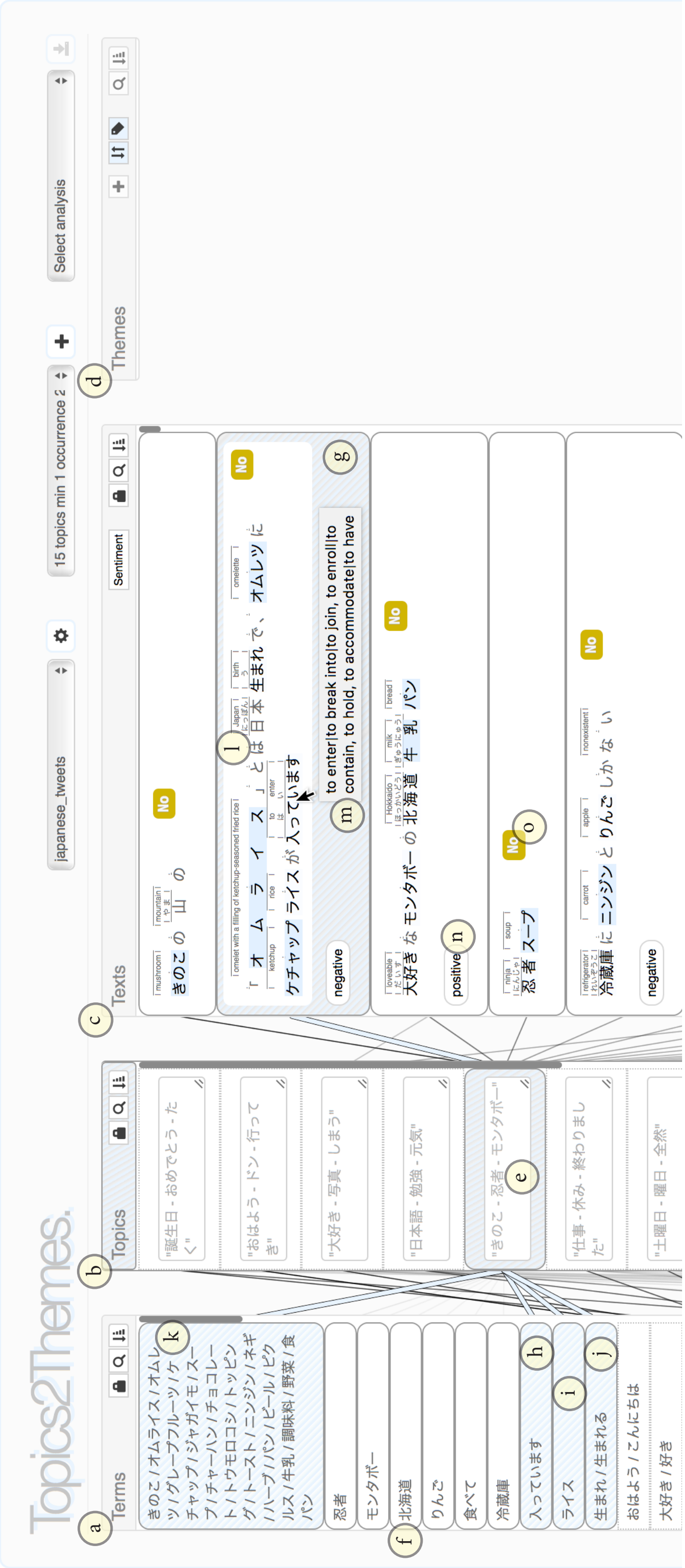

Topics2Themes.
Select analysis
japanese_tweets
15 topics min 1 occurrence 2
Terms
Topics
Texts
Themes
Sentiment
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
きのこ / オムライス / オムレツ / グレープフルーツ / ケチャップ / ジャガイモ / スープ / チャーハン / チョコレート / トウモロコシ / トッピング / トースト / ニンジン / ネギ / ハーブ / パン / ビール / ピクルス / 牛乳 / 調味料 / 野菜 / 食パン
忍者
モンタポー
北海道
りんご
食べて
冷蔵庫
入っています
ライス
生まれ / 生まれる
おはよう / こんにちは
大好き / 好き
"誕生日 - おめでとう - たく"
"おはよう - ドン - 行って き"
"大好き - 写真 - しまう"
"日本語 - 勉強 - 元気"
"きのこ - 忍者 - モンタポー"
"仕事 - 休み - 終わりまし た"
"土曜日 - 曜日 - 全然"
No
きのこ の 山 の
「オムライス」とは 日本 生まれ で、オムレツ に ケチャップ ライス が 入っています
to enter|to break into|to join, to enroll|to contain, to hold, to accommodate|to have
negative
大好き な モンタポー の 北海道 牛乳 パン
positive
忍 者 スープ
冷蔵庫 に ニンジン と りんご しか な い
negative

Figure 2: User interface at the early stages of analysis. (a–d) The *Terms*/*Topics*/*Texts*/*Themes* panels. (e) The selected topic. (f) Rounded border indicating terms and texts associated with the selected topic. (g) The text over which the mouse hovers. (h–k) Terms associated with the text over which the mouse hovers. (k) Cluster of food-related words. (l) Language support in the form of phonetic reading and English translation. (m) Additional English translations for the word over which the mouse hovers. (n) Static label from automatic word list matching. (o) Dynamic label that can be changed by the user.

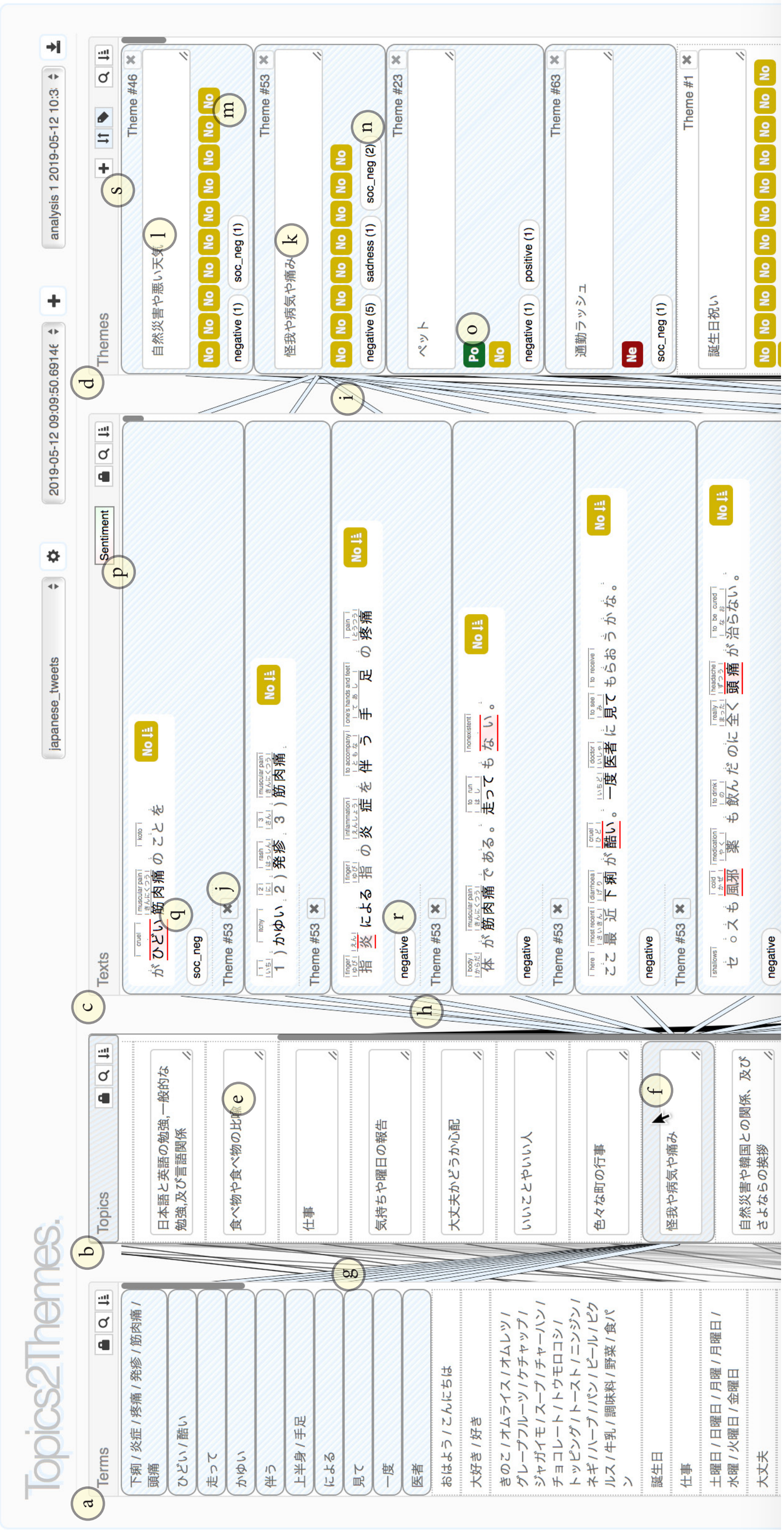


Figure 3: User interface at the later stages of analysis. (a–d) The *Terms*/*Topics*/*Texts*/*Themes* panels. (e) Example of a topic description written by the user. (f) The topic selected by the user. (g–h) Lines showing associations created by the topic modelling algorithm. (i) Lines showing associations assigned by the user, i.e., associations between texts and themes. (j) Labels indicating which themes the text has been assigned to. (k–l) Two themes assigned to texts associated with the selected topic. (m) The number of dynamic labels indicates the number of texts in which the theme occurs. (n) Static labels associated with texts in which the theme occurs. (o) One of the assigned texts contains a *positive* evaluation. (p–q) Automatic sentiment word list matching. (q–r) The red sentiment markings and the static labels indicate negativity of the selected topic. (s) Button for creating a new theme.

The relatively small size of the corpus used, and the small size of each text in the corpus, might make it difficult for the topic modelling algorithm to find reoccurring topics. We therefore configured the tool to allow a large maximum distance[7] for the word2vec-based concept clustering. This makes it possible for the topic model algorithm to find topics based on semantically related words. One example of such a cluster of semantically related words is the cluster of food-related words shown in the top element of the *Terms* panel in Figure 2, which e.g., includes "mushroom", "grapefruit", "soup", "toast", "carrot", "bread", "beer", "milk" and "vegetables". Another example is given by the cluster of words for pain and diseases, which is shown in the top element in the *Terms* panel in Figure 3.

Figure 2 also indicates how the results can be explored by the Topics2Themes tool. In the situation shown in the figure, the user has double-clicked on, and thereby selected, the fifth topic in the *Topics* panel. This has had the effect that the terms most closely associated with the selected topic have been sorted as the top-ranked elements in the *Terms* panel, and that the texts most closely associated with the topic have been sorted as the top-ranked elements in the *Texts* panel. The elements associated with the selected topic have also been given a bold, rounded border. The figure further shows how the user hovers the mouse over one of the texts, which has the effect that the terms included in this text, as well as the topic(s) to which the text is associated, are highlighted with a blue colour.

The language support, in the form of phonetic reading and English translation, is shown in a small font above the Japanese texts, as well as in the form of a tooltip for the word over which the mouse hovers. The *Texts* panel also displays the output of the sentiment and emotion word list matching in the form of labels attached to the texts.

2.3 Manual analysis of the texts extracted

Topics2Themes automatically extracted a total of 183 texts from the text collection as typical for the twelve topics identified. These texts were manually analysed by a learner of Japanese. The learner of Japanese (one of the authors) had very limited experience in reading authentic Japanese texts and had previously mainly read texts from beginner-level textbooks. The goal of the analysis was to find examples of themes that reoccur in these types of Japanese-English bilingual microblogs.

The texts extracted were, for each topic, analysed with the help of the language support provided. The analysis was first carried out individually by the language learner. The same texts were, thereafter, read by the language learner with the help of a Japanese language teacher. The teacher was a native speaker of Japanese and had no previous experience in using tools similar to Topics2Themes. The content of the texts and the identified themes were discussed, and misunderstandings that had led to incorrectly identified or overlooked themes were corrected.

In addition to analysing the texts for reoccurring themes, the language learner also used the dynamic labelling functionality included in Topics2Themes for attaching the sentiment labels *positive* or *negative* to texts whose content included a positive or negative evaluation.

While using the tool, reflections on its usefulness were made, and notes regarding usability issues were taken.

3 Results and reflections

Results of the study consist of examples of themes that are reoccurring in the corpus, as well as of reflections on the usefulness and usability of the tool and its language support, when applied to Japanese texts.

3.1 Outcome of the manual analysis

Table 1 shows the outcome of the analysis task given, i.e., the task of finding examples of reoccurring themes in the collection of microblogs. The table shows the final analysis, after the Japanese teacher had corrected the analysis carried out individually by the learner. This analysis resulted in that a total of 78 themes were identified, of which 35 occurred at least twice, and 19 at least three times among the texts analysed.

[7] Two words with a Minkowski distance of up to 0.7 could be included in the same cluster.

Topic description	**Theme descriptions**	**Nr of occ.**
Birthdays	• **Birthday greetings**	12 (12)
Good morning greetings	• ***Good morning greetings***	*12 (12)*
	• *Reports on/confirmations regarding weekdays*	*1 (5)*
	• *Reports of going to work/work starting*	*1 (3)*
Expression of liking	• **Expression of liking towards a person**	6 (6)
	• **Images/photos that someone likes**	4 (4)
	• *Food that someone likes*	*1 (3)*
	• *Wise sayings and advice on how to live*	*1 (8)*
	• *Injury, illness or pain*	*1 (8)*
Studies of the Japanese and English languages, studies in general and matters related to language	• **Questions about Japanese studies**	2 (2)
	• **Doubts regarding English studies**	2 (2)
	• **Someone reports to study Japanese**	2 (2)
	• **Someone's level of English**	2 (2)
	• **Changes of texts into Japanese**	2 (2)
Food and food metaphors	• ***Food in general***	*7 (9)*
	• **Cooking and food ingredients**	4 (4)
	• ***Food that someone likes***	*3 (3)*
	• ***Food metaphors***	*3 (3)*
	• *Wise sayings and advice on how to live*	*1 (8)*
Work	• **Rest from work/reports of work that ends**	5 (5)
	• ***Reports of going to work/work starting***	*3 (3)*
	• *Good morning greetings*	*1 (12)*
Feelings and days of the week	• ***Reports on feelings***	*5 (7)*
	• ***Reports on/confirmations regarding weekdays***	*5 (5)*
	• *Good morning greetings*	*1 (12)*
	• *Information about events*	*1 (7)*
Okay	• ***Worries of whether something/oneone is okay***	*7 (7)*
	• ***Natural disasters and bad weather***	*2 (10)*
Good things and good people	• ***Wise sayings and advice on how to live***	*5 (8)*
	• **Questions on appearances/methods**	4 (4)
	• ***Reports on feelings.***	*2 (7)*
	• *Food in general*	*1 (9)*
	• *Natural disasters and bad weather*	*1 (10)*
	• *Worries of whether something/someone is okay*	*1 (7)*
	• *Food metaphors*	*1 (3)*
Information about events taking place in different cities	• ***Information about events taking place in different Japanese cities***	*6 (7)*
Injury, illness or pain	• ***Injury, illness or pain***	*7 (8)*
	• *Natural disasters and bad weather*	*1 (10)*
Natural disasters, relations with Korea and goodbye greetings	• ***Natural disasters and bad weather***	*8 (10)*
	• **Korea-Japan relations**	3 (3)
	• ***Worries of whether something is okay***	*2 (7)*
	• *Wise sayings and advice on how to live*	*1 (8)*

Table 1: Themes found in texts associated with each one of the automatically detected topics. *Nr of occ.* indicates the number of texts, associated with this topic, in which the theme was found. The number shown in parenthesis indicates the total number of texts in which this theme occurred. Bold indicates that a theme has occurred at least twice in texts associated with the topic. Italics indicates that a theme has also been found in texts associated with other topics, and thereby is listed multiple times in the table. (Note that the same text can be associated with several topics and assigned to several themes.)

The table includes all the themes that occurred at least three times, as well as the themes that occurred at least twice for the fourth topic (for which no themes occurring more than twice were identified). Themes that were assigned to texts associated with several topics are listed multiple times in the table, once for each topic. When only taking themes that occurred at least twice for a topic into account, it can be seen that most of the twelve topics extracted contained semantically coherent themes. Examples include the themes related to (i) birthday greetings, (ii) good morning greetings, (iii) food, (iv) language, (v) work starting and ending, and (vi) injury, illness and pain. There were, however, also topics with non-coherent themes, e.g., the topic "Feelings and days of the week".

Most texts did not contain a positive or negative evaluation, and hence were not assigned a positive or negative label with the manual labelling functionality provided by the dynamic labels. Exceptions were texts belonging to the themes "Images/photos that someone likes" (4 positive), "Expression of liking towards a person" (6 positive), "Food that someone likes" (4 positive), and "Food in general" (1 positive).

From the re-analysis performed together with the teacher, it could be concluded that a total of 25 texts among the 183 texts analysed had been misunderstood by the learner of Japanese. For these texts, themes assigned were removed or new theme assignments were created.

Figure 3 gives an example of what the interface of the Topics2Themes tool looks like when themes have been added. The figure shows texts associated with the topic "Injury, illness or pain" that have been assigned to four different themes created in the *Themes* panel. The figure also shows that the topics have been given user-defined descriptions, which have replaced the initial default names.

3.2 Reflections on usefulness

The subjective reflection of the analysis led to the insight that the application of Topics2Themes to the text collection enabled users to access text content that otherwise would have been very difficult to access for someone not used to reading authentic Japanese texts. Although some level of Japanese language skills were still required to access the text content, the reading support provided made it possible to focus on the search for reoccurring themes while reading the texts. That is, it was feasible to focus on the content of the texts, instead of having to use effort for manually tokenising it, for figuring out the dictionary form of the tokens, and for looking them up in a dictionary.

The automatic selection of a subset of the texts for manual analysis was also perceived as an indispensable feature for accessing the content of the document collection. It would have taken a very long time for the learner of Japanese to manually analyse all posts in the collection, in order to find examples of reoccurring themes, even with the help of the language support provided. To manually read 183 short texts with automatic reading support was, however, perceived as a feasible task for the language learner.

These subjective reflections were supported by the more objective facts that (i) only around 14 percent of the texts analysed had been misunderstood, despite the language learner's previous unfamiliarity with reading authentic Japanese texts, and (ii) by the detection of the 35 examples of reoccurring themes among the subset of 183 texts selected by the tool.

The output of the word list matching, in the form of texts being given static labels as well as in the form of polarity words in the text being highlighted in green or red, was useful for gaining a first impression of the topic. For instance, the static labels and the red highlighting of words signifying negative polarity, which are shown for the texts associated with the topic "Injury, illness or pain" in Figure 3, indicate that this topic includes negative content. The output of the word list matching was, however, not perceived as useful for performing the text analysis of each individual text.

3.3 Usability issues

Usability issues and missing features detected while performing the analysis were mostly related to the interaction with the interface of the Topics2Themes tool, as well as to how information was presented and sorted in the tool. The analysis of the Japanese texts performed in this study is one of the first authentic use cases to which the Topics2Themes tool has been applied, and we therefore expected that usability issues not stemming from the adaptations performed for Japanese would be detected.

Three large usability issues—described in the following paragraphs—were detected, and the user interface was updated to resolve these issues.

Previously created themes relevant for the topic and texts that are being analysed are typically sorted as the top-ranked elements in the *Themes* panel. However, when the user creates a new theme element, this new element was positioned at the bottom of the *Themes* panel in the version of Topics2Themes used for the evaluation. This causes unnecessary scrolling when the newly created theme is to be used, and the tool's functionality was therefore changed to position newly created themes as the first element in the *Themes* panel.

The content of the *Terms* panel was used in the process of determining (i) which additional words should be added to the stop word list, (ii) which words to add to the list of words to be excluded from the automatic clustering process. This panel was also used for gaining an initial overview of the output of the topic modelling algorithm. However, the *Terms* panel was not perceived as useful when performing the actual analysis of the texts. At the same time, the horizontal space available for writing theme descriptions was perceived as too small. Functionality for minimising the *Terms* panel was therefore added, to make it possible for the user to choose to have more screen space available for the *Themes* panel while performing the text analysis.

The evaluated version of the tool did not include any functionality for determining whether there were texts in the *Texts* panel, for which no theme assignments had yet been made by the user. Omitted texts were therefore difficult to detect. A functionality for sorting texts according to their number of assigned themes was therefore added to the *Texts* panel, i.e., a functionality that can be used for easily finding texts without any theme associations.

There were also a number of smaller usability issues associated with the lists being automatically resorted or being automatically scrolled to the top element in cases when these events should not take place. These issues have all been corrected, and a new version of Topics2Themes has been released with the implemented corrections.

4 Conclusions

The aim of the Topics2Themes tool is to provide functionality for automatic extraction and sorting of a subset of texts from a text collection too large for a fully manual analysis. The automatically extracted texts are to contain examples of themes that reoccur in the text collection, and Topics2Themes also provides functionality for documenting reoccurring themes detected when these texts are manually analysed. The main goal of the current study was to investigate whether it is possible to achieve this aim when Topics2Themes is applied to a language very different from English, i.e., different from the language for which the tool was originally developed. When applying the tool to a collection of around 1,000 short Japanese texts, and manually analysing 183 of these texts, 35 examples of reoccurring themes were identified. 19 of these themes occurred at least three times among the extracted texts. This shows that the functionality of extracting relevant texts can be achieved also when Topics2Themes is applied to texts written in another language.

The current study also included an evaluation of the usefulness of the Japanese extension of Topics2Themes for making it possible for a learner of Japanese to access the content of a collection of authentic Japanese texts. The learner perceived the Japanese reading support provided by the tool, and the fact that only a subset of a large text collection was extracted for manual analysis, as necessary for accessing the content of the text collection. These subjective reflections were supported by the fact that only around 14 percent of the texts analysed had been misunderstood by the learner of Japanese, as well as by the fact that the learner was able to carry out the task of identifying reoccurring themes, despite limited previous experience in reading authentic Japanese texts.

A number of general usability issues were detected when Topics2Themes was used for analysing the Japanese texts. For instance, issues related to the functionality of presenting and resorting the information. These usability issues have been addressed, and a new version of the tool has been released. This new version of Topics2Themes will be used when the tool is applied to other text types.

Acknowledgements

The adaptation of Topics2Themes to Japanese, as well as the evaluation on Japanese texts, was funded by the Japan Society for the Promotion of Science, in the form of a "Postdoctoral Fellowships for Research in Japan (Short-term)" research grant.

The updated version of the tool, in which usability issues were corrected, was developed within the Språkbanken and SWE-CLARIN infrastructures, supported by the Swedish Research Council (2017-00626). Further developments and evaluations of the tool will also continue within the Språkbanken and SWE-CLARIN infrastructures.

We would also like to thank the reviewers, for their useful input, and the teachers at the IAY International Academy Japanese Language School in Sapporo, for making it possible to evaluate our tool.

References

Magnus Ahltorp. 2012. A personalizable reading aid for second language learners of Japanese. Master's thesis, Royal Institute of Technology, Sweden.

Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. 2014. Serendip: Topic model-driven visual exploration of text corpora. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, VAST '14, pages 173–182. IEEE.

Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6):1397–1410, June.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, January.

David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, April.

Guoray Cai, Feng Sun, and Yongzhong Sha. 2018. Interactive visualization for topic model curation. In *Proceedings of the ACM IUI 2018 Workshop on Exploratory Search and Interactive Data Analytics*, ESIDA '18. CEUR-WS.org.

Jaegul Choo, Changhyun Lee, Chandan K. Reddy, and Haesun Park. 2013. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1992–2001, December.

Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, pages 74–77. ACM.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, KDD '96, pages 226–231. AAAI Press.

Enamul Hoque and Giuseppe Carenini. 2015. ConVisIT: Interactive topic modeling for exploring asynchronous online conversations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 169–180. ACM.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent Dirichlet Allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, June.

JMdict. 2013. The JMDict Project. `http://www.edrdg.org/jmdict/j_jmdict.html`.

Taku Kudo. 2006. MeCab: Yet another part-of-speech and morphological analyzer. `https://ci.nii.ac.jp/naid/10027284215/en/`.

Daniel D. Lee and H. Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*, NIPS '00, pages 556–562. MIT Press. Proceedings of the Neural Information Processing Systems Conference 2000.

Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. 2012. iVisClustering: An interactive visual document clustering via topic modeling. *Computer Graphics Forum*, 31(3pt3):1155–1164, June.

Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105:28–42, September.

Wang Ling, Luis Marujo, Chris Dyer, Alan Black, and Isabel Trancoso. 2014. Crowdsourcing high-quality parallel data extraction from Twitter. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, WMT '14. ACL.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. `https://arxiv.org/abs/1301.3781`.

Akira Nakamura. 1993. *Kanjo hyogen jiten [Dictionary of Emotive Expressions]*. Tokyodo Publishing, Tokyo, Japan.

Rafal Rzepka and Kenji Araki. 2012. Polarization of consequence expressions for an automatic ethical judgment based on moral stages theory. *IPSJ SIG Notes*, 14(2012-NL-207):1–4, July.

Rafal Rzepka and Kenji Araki. 2017. What people say? Web-based casuistry for artificial morality experiments. In *Proceedings of the 10th International Conference on Artificial General Intelligence (AGI '17)*, volume 10414 of *LNCS*, pages 178–187. Springer International Publishing.

Maria Skeppstedt, Andreas Kerren, and Manfred Stede. 2018a. Vaccine hesitancy in discussion forums: Computer-assisted argument mining with topic models. *Studies in Health Technology and Informatics*, 247:366–370. Proceedings of the 29th Medical Informatics Europe Conference (MIE '18) — Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth.

Maria Skeppstedt, Kostiantyn Kucher, Manfred Stede, and Andreas Kerren. 2018b. Topics2Themes: Computer-assisted argument extraction by visual analysis of important topics. In *Proceedings of the 3rd Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources at LREC '18*, VisLR III, pages 9–16. ELRA.

Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, IUI '18, pages 293–304. ACM.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 133–140. ACL.

CLARIN

Common Language Resources and Technology Infrastructure

Linköping Electronic Conference Proceedings 172
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2020
ISBN 978-91-7929-807-4

Front Cover Illustration:
Picture Composition by CLARIN ERIC