An infrastructure for Historical Dutch Corpus Development

Katrien Depuydt
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
katrien.depuydt@ivdnt.org

Jesse de Does
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
jesse.dedoes@ivdnt.org

Vincent Prins
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
vincent.prins@ivdnt.org

Mathieu Fannee
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
mathieu.fannee@ivdnt.org

Roland de Bonth
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
roland.debonth@ivdnt.org

Thomas Haga
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
thomas.haga@ivdnt.org

Abstract

We describe an infrastructure for linguistic annotation of historical Dutch texts, consisting of tagging and lemmatisation guidelines, gold standard data, trained tagging models, the GaLAHaD platform for automatic linguistic annotation and evaluation and the LAnCeLoT manual annotation tool. Users can upload unannotated materials to the GaLAHaD platform, where trained PoS taggers and lemmatizers (including modern deep learning models) annotate the data. Results can be evaluated against newly expanded gold standard corpora (13^{th} to 19^{th} centuries) with various metrics, then exported for further analysis. The LAnCeLoT tool supports manual annotation correction at both type and token level, thus supporting the development of in-domain gold standard material. The infrastructure was developed at the Instituut voor de Nederlandse Taal (INT, Dutch Language Institute) and first released in 2025.

1 Introduction

Historical texts are essential source material for both linguistic and digital humanities research. Adding linguistic annotation layers (e.g. part of speech and modern Dutch lemma) to historical text helps to make the data more accessible. These layers support linguistic analysis, and general users need not be concerned with historical spelling variation when querying and analysing the data.

Linguistic annotation of historical Dutch is far from trivial. Consider the following variants of the lemma *wereld* (”world”), taken from the central database of the INT and covering 15 centuries of Dutch language (the modern Dutch forms are in italics):

> uuerildis, uue roldi, uueroldi, uuerildi, uuerolt, uuerolde, werelde, vuerolti, werelt, weerelt, *wereld*, weerelds, wereldt, *werelden*, weereld, werrelts, waerelds, weerlyt, wereldts, vveerelts, waereld, weerelden, waerelden, weerlt, werlt, werelds, sweerels, zwerlys, swarels, swerelts, werelts, swerrels, weirelts, tsweerelds, werret, vverelt, werlts, werrelt, worreld, werlden, wareld, weirelt, weireld, waerelt, werreld, werld, vvereld, weerelts, werlde, tswerels, werreldts, weereldt, *wereldje*, waereldje, weurlt, wald, weëled

The formal variation (spelling and inflection) is abundant and is not easy to capture in patterns. An additional complication is the irregularity of word separation. What would be considered a single word from the point of view of modern Dutch can be split into several tokens (*te rugh* for modern *terug*) or vice versa, what would be written as separate words is realized as a single token (*sprackse*, *ensagh*).

Providing the tools to produce high quality annotation of historical Dutch language is therefore challenging, and it is equally challenging to do so in a such a manner that the tools are also accessible for non-technical users, and we can support metholodologically sound tool use. This involves much more

than merely providing a processing facility. First, linguistic principles (tagset and tagging and lemmatization guidelines) need to be explicit and suitable for historical Dutch. The training and evaluation material adhering to these guidelines should be publicly available and should cover all historical periods. Moreover, without awareness of the noise and bias that the tools inevitable introduce, a methodologically sound use of these tools is hardly possible. Therefore, instead of treating the tools as black boxes (with at best a generic accuracy score on a benchmark set), users should be assisted in making informed processing tool choices and in analyzing tool performance in detail and inspecting relevant results.

Supported by additional funding of both CLARIAH$^+$ and currently SSHOC-NL, we have been developing an infrastructure, officially released in April 2025, to tackle the above-mentioned issues.

The infrastructure consists of:

- A diachronic tagset (TDN) and lemmatization guidelines for historical and contemporary Dutch
- Gold standard training and evaluation material (13th–19th century)
- The GiGaNT-Hilex historical lexicon[1]
- Trained tagger-lemmatizers (currently PIE framework and Hugging Face transformers framework)
- Detailed evaluation of PoS tagging and lemmatization to allow for a methodologically sound choice of the appropriate annotation tool
- A user-friendly platform, GaLAHaD for corpus annotation and evaluation
- LAnCeLoT, a service (web application) for creating gold standard corpus data

2 Related work

2.1 Linguistic annotation of Historical Dutch

Prior to the development of the infrastructure presented here, several tools were already available for the linguistic annotation of historical Dutch. However, efforts were fragmented, training material was scarce (even absent for some periods) and diverse in adopted tagset and annotation guidelines. A common strategy was lacking to deal with issues such as nontrivial word segmentation and the substantial degree of formal variation in spelling and inflection which is hard to classify. This became very clear in the Nederlab project[2], where an attempt was made to annotate a large diachronic corpus of Dutch from the 6th - 21st century. The quality of the linguistic annotation was insufficient, and the tagset and tagging principles, originally developed for modern Dutch, turned out to be unsuitable for historical Dutch.

We briefly discuss some of the systems that have been developed.

- The well-known Frog[3] tagger (van den Bosch et al., 2007) has been used in two ways: using the standard Dutch models after normalizing spelling to modern Dutch, cf. (Tjong Kim Sang et al., 2017), and by training specific models: a model for Middle Dutch (trained on CRM and Gysseling) has been developed.
- Adelheid[4], developed by Hans van Halteren. Adelheid is a tagger-lemmatizer for fourteenth century Dutch charters (as found in the corpus van Reenen/Mulder), with a special focus on the treatment of the extensive orthographic variation in this material, achieving about 95% accuracy in a tenfold cross-validation experiment for both tagging and lemmatization. For variation in spelling, expanded lexicon with predicted spelling variants is used. For those forms which can also not be found in this expanded lexicon, the word class is derived first, and subsequently a search for the most similar lexicon word is performed.
- The web service and application INL labs incorporated a supervised SVM-based tagger trained on the Letters as Loot corpus[5]. The GiGaNT-Hilex historical lexicon is used for lemmatization, in conjunction with a noisy-channel spelling variation model.
- PIE[6], developed by Enrique Manjavacas, Ákos Kádár, and Mike Kestemont, (Manjavacas et al.,

[1] A morphosyntactic lexicon of historical Dutch, derived from the scholarly dictionaries of historical Dutch, ONW, VMNW, MNW and WNT, cf. https://ivdnt.org/corpora-lexica/gigant/, https://gtb.ivdnt.org

[2] https://www.nederlab.nl/, a project funded by NWO (Dutch research council) that ran from 2013-2017

[3] http://languagemachines.github.io/frog/

[4] http://adelheid.ruhosting.nl/

[5] http://inl-labs.inl.nl/, no longer available

[6] https://github.com/emanjavacas/pie

2019). A deep learning approach is used. Lemmatization is approached as a string transduction task with an encoder-decoder architecture enriched with sentence context information using a hierarchical sentence encoder. They report significant improvements over the state-of-the-art when training the sentence encoder jointly for lemmatization and language modeling.

- RNNTagger[7], (Schmid, 2019). Part of speech tagging is based on bidirectional long short-term memory networks (LSTMs), with character-based word representations. Lemmatization uses an encoder-decoder system with attention.

2.2 NLP Processing services

A substantial amount of work has been devoted to developing web services and web-based interfaces for NLP tasks. Prominent examples include Stanford CoreNLP and the UDPipe web service. Some systems are centered around a specific core task, such as tagging or syntactic parsing, while also integrating auxiliary functionality, including format conversions and preliminary tasks like tokenization. Other platforms support the chaining of multiple web services, as exemplified by Weblicht[8], or facilitate tool selection based on task requirements, as with the CLARIN Language Resource Switchboard[9]. At a larger infrastructural level, the European Language Grid[10] provides a unified environment for accessing, combining, and deploying NLP services across languages. For Dutch, we mention the CLAM[11]-based Language and Speech Tools[12] hosted by the Centre for Language and Speech Technology at Radboud University.

None of the discussed platforms directly supports Historical Dutch, or supports an infrastructure integrating processing, gold-standard data creation, corpus building, and evaluation.

2.3 Tools for manual annotation of corpus data

There are various environments for manual corpus annotation available. We mention the INCEpTION platform[13] (Klie et al., 2018), the FoLiA Linguistic Annotation Tool FLAT[14], and the Arborator Grew[15] system, with which we have in common that it is part of an effort to create an infrastructure for a research community.

The main reason we have developed a separate tool for manual annotation tasks is that we found the purely token-based workflow in the available tools lacking in efficiency. The main features of LAnCeLoT - quickly reviewing all occurrences of a word, and bulk assignment of annotations turned out to be indispensable for efficient corpus processing.

3 Requirements

3.1 Support the research community

As mentioned in the introduction, the availability of processing tools is just a part of an ecosystem that supports historical corpus research. We aim to enable researchers from the Humanities and Social Sciences to make use of corpus data and linguistic annotation tools in a methodologically sound manner, supporting data formats prevalent in these communities. More specifically, we need to support several scenario's in which users are enabled to:

- choose the right tagger for a dataset, with or without having relevant gold standard data available.
- develop gold standard data for evaluation and/or training.
- evaluate taggers on gold standard data.
- inspect the tagging results.
- annotate data while preserving document structure annotation.

[7] https://www.cis.uni-muenchen.de/~schmid/tools/RNNTagger/
[8] https://weblicht.sfs.uni-tuebingen.de/weblicht/
[9] https://www.clarin.eu/content/language-resource-switchboard
[10] https://live.european-language-grid.eu/
[11] https://proycon.github.io/clam/
[12] https://webservices.cls.ru.nl/portal/
[13] https://inception-project.github.io/
[14] https://github.com/proycon/flat
[15] https://arborator.grew.fr/

This translates to the following requirements:

- Ensure that high quality tools are available.
- Develop clear and documented annotation guidelines, tested on a substantial amount of historical corpus data.
- Support a methodologically sound approach to historical corpus processing and analysis by providing extensive evaluation functionality.
- Support the development of gold standard data.
- Make benchmark data available to support tagger choice.
- Provide training data for tagger development for all stages of historical Dutch.
- Support digital humanities research by keeping the XML-encoding (e.g. TEI) of the uploaded text intact.

Finally , in terms of technical requirements, the following should be met: 1. Simplicity: do not implement a more complex pipelines than strictly necessary. In our experience, complex pipelines are sensitive to errors and results are not transparent. For this reason, we have not included BlackLab (de Does et al., 2017)[16] corpus creation as a step in the GaLAHaD application. 2. Robustness for larger uploads. 3. Separation of backend (web service with a documented API) and user interface.

4 Tagset, tagging guidelines and lemmatisation principles

There are several tagsets for linguistic annotation of text corpora containing historical and contemporary Dutch, such as the tagset of GiGaNT (Groot Geïntegreerd lexicon van de Nederlandse Taal, Ruitenberg et al., 2012), the corpus tagset CGN/D-Coi (Van Eynde, 2005), the tagset used in both the Corpus Gysseling (van Dalen Oskam and Depuydt, 2000) and the Corpus Van Reenen-Mulder, and the tagset for the Corpus Oudnederlands[17]. These tagsets differ in the method of tagging (lexical or functional), in naming (what is called A in one tagset is called B in another) and in degree of detail (which features are tagged). Due to these differences, it is complicated to compare linguistically annotated corpora with one other.

In order to make it possible to study language usage throughout the centuries, we have created a new tagset. This Tagset for Diachronic corpus data of Dutch (Tagset voor Diachroon corpusmateriaal van het Nederlands: TDN, Haga et al., 2024) was developed as a result of the re-evaluation of the GiGaNT tagset and after careful analysis of the above-mentioned corpus tagsets.

A core requirement for its development was that the TDN would be mappable to the existing tagsets, that TDN should be applicable to all stages of Dutch language - from the sixth century to the present day - and that a core tagset should be defined to allow production of large amounts of training and evaluation data. The proposed tagset was submitted to a large number of (historical) linguists who critically read along and provided valuable comments. The core tagset is applied in the gold standard corpora, cf. section 5.

The tagset was compiled on the basis of a number of design principles.

- TDN should enable both coarse tagging (main PoS and lemma) and detailed annotation, including the annotation of difficult features such as case. By accommodating different levels of linguistic detail, we ensure that projects with distinct objectives can nonetheless benefit from one another's efforts through a shared common core.
- TDN is based on words, not tokens. A token is defined here as a string of characters separated from other tokens by a space (and possibly a punctuation mark). Some tagsets apply the principle that one token corresponds to one part of speech and one lemma. This principle cannot be applied to historical language material. In older language stages, words sometimes include whitespace, while tokens can consist of multiple words (as in the case of clitic combinations).
- TDN is based on *functional tagging* (as opposed to *lexical tagging*). Lexical tagging aims to assign the same part of speech to what is considered the same word across different contexts, typically using a dictionary or lexicon as a reference. Functional tagging, by contrast, assigns part of speech based on a word's syntactic role in a specific context. Its main advantage is that a word form can

[16] https://blacklab.ivdnt.org/

[17] https://corpusoudnederlands.ivdnt.org

be tagged according to actual usage, without making claims about how lexicalized that usage is. For historical language, decisons on lexicalisation can only be made after extensive analysis of annotated corpus data.

- TDN should enable mapping to other tagsets. Where tagsets differ in approach (e.g. lexical versus functional), it should in principle be possible to use the information available from TDN, with or without the aid of lexicon information, to translate it into other tagsets, such as Corpus Gysseling/CRM tagset, CGN/D-COI, Universal Dependencies.

The assignment of a modern Dutch equivalent to historical Dutch language data is based on etymological criteria. Words that do not occur in Modern Dutch any more are also assigned a Modern Dutch equivalent. The lemmatisation principles are described in Depuydt et al., 2024.

5 Gold standard corpora

The main gold standard corpora of historical Dutch available before the project were: Corpus of Old Dutch[18], 500–1200; Corpus Gysseling[19], 1200–1300; Corpus van Reenen-Mulder[20], 1300-1400; Corpus Letters as Loot[21], 1661-1783. Obviously, there were significant gaps in the coverage of these corpora. An additional complication was the fact that three different tagsets were used with different leves of detail. The annotation of the Corpus of Old Dutch is the most detailed and provides fine-grained features like case for nouns and mood for verbs, which were ommitted for reasons of feasibility in Gysseling and CRM. The Letters as Loot Corpus provides only the part of speech label. We have taken samples from Gysseling, CRM, and Letters as Loot corpora and have adapted the lemmatisation and part of speech tagging to the principles that were defined. We list the training corpora developed in table 1, and the distribution of the material per period in table 2.

Corpus	**Tokens**	**Tokens (%)**	**Documents**
total	*863,033*	*100.00%*	*13,788*
crm[22]	123,913	14.36%	597
gentse-spelen[23]	87,913	10.19%	1
gysseling-literair[24]	73,195	8.48%	29
gysseling-ambtelijk	66,853	7.75%	195
kranten-19[25]	63,279	7.33%	27
letters-as-loot[26]	63,227	7.33%	112
dictionary-quotations-14[27]	53,058	6.15%	3,191
dictionary-quotations-15[28]	42,537	4.93%	2,316
dictionary-quotations-16[29]	46,034	5.33%	1,838
dictionary-quotations-17[30]	46,287	5.36%	1,919
dictionary-quotations-18[31]	47,485	5.50%	1,794
dictionary-quotations-19[32]	35,006	4.06%	1,550
couranten[33]	29,674	3.44%	119
clvn[34]	27,204	3.15%	88
dbnl-excerpts-15[35]	15,460	1.79%	3
dbnl-excerpts-16	11,714	1.36%	3
dbnl-excerpts-17	10,126	1.17%	2
dbnl-excerpts-18	10,067	1.17%	2
dbnl-excerpts-19	10,001	1.16%	2

Table 1: Distribution of tokens per corpus

[18] https://taalmaterialen.ivdnt.org/download/corpus-oudnederlands-online/
[19] https://taalmaterialen.ivdnt.org/download/tstc-corpus-gysseling/
[20] https://www.middelnederlands.nl/corpora/crm14/
[21] https://taalmaterialen.ivdnt.org/download/tstc-bab-gouden-standaard/

Period	Tokens	Tokens (%)	Documents
total	863,033	100.00%	13,788
1200–1300	140,048	16.23%	224
1300–1400	176,971	20.51%	3,788
1400–1500	52,538	6.09%	2,318
1500–1600	171,277	19.85%	1,929
1600–1700	87,675	10.16%	2,041
1600–1800	63,227	7.33%	112
1700–1800	57,552	6.67%	1,796
1800–1900	113,745	13.18%	1,580

Table 2: Overview of gold standard material by period

In order to fill the gaps, we have extended the available material. Our aim was to ensure that at least 100,000 tokens per century are available, tagged according to the defined lemmatisation principles and the "core" level of the TDN tagset. This has been achieved for all centuries[36] but the fifteenth. Cf. table 2.

To the material that was already annotated in some form we added a small set of excerpts from the DBNL digital library, several per-century sets of dictionary quotations from the scholarly historical dictionaries of Dutch, a sample set from the corpus of Late Middle and early Modern Dutch (CLVN) and the Couranten Corpus, the completely tagged Gentse Spelen collection, and a set of 19th century newspaper issues based on KB (National Library of the Netherlands) data.

6 Linguistic annotation: the taggers/lemmatizers

Currently, GaLAHaD provides tagging and lemmatization models for the PIE framework (slightly adapted for the GaLAHaD platform) and for the transformer-based taggers discussed in the next subsection.

Transformer-based tagging with pretrained language models

In order to benefit from advances in deep learning, we have developed a tagger-lemmatizer based on the BERT token classifier, implemented using the Hugging Face framework[38], and the GysBERT historical Dutch language model[39]. The lemmatizer uses the GiGaNT-Hilex historical lexicon, and a ByT5 (Xue

[22]Corpus van Reenen-Mulder, selection

[23]Cf. https://www.dbnl.org/tekst/_gen001gent01_01/_gen001gent01_01_0004.php

[24]Corpus gysseling, selection

[25]19th century newspaper issues

[26]The Letters as Loot / Brieven als Buit-corpus. Leiden University. Compiled by Marijke van der Wal (Programme leader), Gijsbert Rutten, Judith Nobels and Tanja Simons, with the assistance of volunteers of the Leiden-based Wikiscripta Neerlandica transcription project, and lemmatised, tagged and provided with search facilities by the Institute for Dutch Lexicology (INL). 3rd release januari 2021. http://hdl.handle.net/10032/tm-a2-s4

[27]Dictionary quotations (MNW: Dictionary of Middle Dutch, https://ivdnt.org/woordenboeken/historische-woordenboeken/middelnederlandsch-woordenboek/), 14th century

[28]Dictionary quotations (MNW)

[29]Dictionary quotations (MNW)

[30]Dictionary quotations (WNT: Dictionary of the Dutch Language, https://ivdnt.org/woordenboeken/historische-woordenboeken/woordenboek-der-nederlandsche-taal/)

[31]Dictionary quotations (WNT)

[32]Dictionary quotations (WNT)

[33]Couranten Corpus (version 2.0) (July 2025) [Online Service]. Available at the Dutch Language Institute: https://hdl.handle.net/10032/tm-a3-c2.

[34](Corpus Laatmiddel- en Vroegnieuwnederlands, van der Sijs et al., 2018

[35]DBNL excerpts, 19th century

[36]All available data for the Old Dutch stage (before 1200) is included in the Corpus Of Old Dutch[37], which is tagged manually. Currently, there is no Old Dutch material for which automatic tagging would make sense.

[37]https://corpusoudnederlands.ivdnt.org/

[38]https://github.com/INL/int-huggingface-tagger

[39]https://huggingface.co/emanjavacas/GysBERT

et al., 2022) pre-trained byte-to-byte model, finetuned on data from this lexicon, as a fallback for out-of-vocabulary words.

Annotation results

We list the evaluation results of some of the subcorpora below. Results are taken from the GaLAHaD online application version of January 30, 2026. The *pie-* label refers to PIE trained on different datasets (*all* for the complete training set, *1640-1600* and *1600-1900* refer to the combination of all training sets belonging to these periods); likewise *hug-* refers to the transformer-based taggers. The label *-enhanced* refers to hidden tag enhancements used internally, which are used to connect tokens which are considered parts of the same word by the TDN guidelines.

CLVN

PoS

tagger	macro f1	micro accuracy
hug-tdn-1400-1600	0.52	0.93
hug-tdn-all-enhanced	0.53	0.92

Lemma

tagger	macro f1	micro accuracy
hug-tdn-all-enhanced	0.64	0.86
pie-tdn-all	0.55	0.86

Couranten Corpus

PoS

tagger	macro f1	micro accuracy
hug-tdn-all-enhanced	0.70	0.96
hug-tdn-1600-1900	0.70	0.95

Lemma

tagger	macro f1	micro accuracy
hug-tdn-all-enhanced	0.72	0.92
pie-tdn-all	0.60	0.89

Letters as Loot

PoS

tagger	macro f1	micro accuracy
hug-tdn-all	0.72	0.91
hug-tdn-1600-1900	0.67	0.91

Lemma

tagger	macro f1	micro accuracy
hug-tdn-all-enhanced	0.56	0.85
pie-tdn-all	0.48	0.82

WNT Dictionary quotations, 19th century

PoS

tagger	macro f1	micro accuracy
hug-tdn-all	0.64	0.97
hug-tdn-1600-1900	0.65	0.97

Lemma

tagger	macro f1	micro accuracy
hug-tdn-all-enhanced	0.85	0.96
hug-tdn-1600-1900	0.81	0.94

7 GaLAHaD: Generating Linguistic Annotations for Historical Dutch

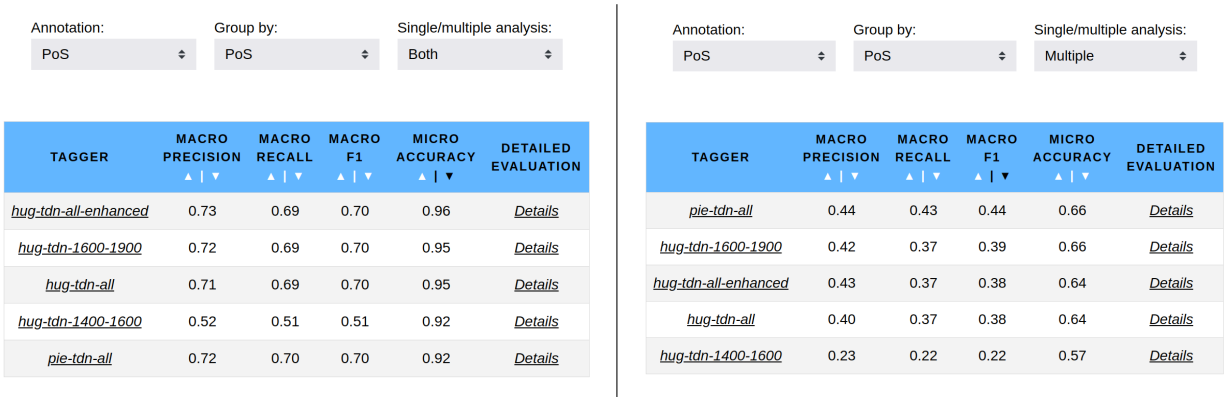
The platform serves two purposes. One is to make annotation and tool evaluation easily accessible to researchers, the other to make it easy for developers to contribute their tools and models in the platform, and thus compare them to other tools with gold standard material included in the platform.

Apart from the basic task of uploading and annotating corpus material, *GaLAHaD*[40] is designed to enable end users to choose the optimal path for their material. The platform provides options to inspect and evaluate the result of the annotation process, in order to raise the awareness of typical errors and biases in the tools. The functionality of comparing annotation layers enables users to assess the accuracy of different tools on their data. It can be used both to evaluate a layer added by an automatic tagger with respect to a gold standard reference layer, or to compare layers added by different taggers. Disagreement between layers is not only represented by global statistics, but also illustrated by examples which are immediately visible in the tool. Different metrics and targeted evaluations for certain token types may lead to different conclusions as to the path of choice, cf. fig. 7, where the evaluation of clitic combinations with macro F1 (on the right) contradicts the rosy picture of the standard micro accuracy metric on the left. The resulting annotated material can be uploaded to the *Autosearch* corpus exploration environment and to the *LAnCeLoT* tool for manual correction of linguistic annotation.

For tool developers, the docker-based application architecture[41] ensures easy contribution of tools to the platform. The application and taggers are hosted by the INT and are accessible with any CLARIN account. There is also the option to self-host an instance using the publicly available docker images from the INT docker hub or the open source code available on GitHub.

[40] https://portal.clarin.ivdnt.org/galahad/, https://github.com/INL/galahad

[41] https://github.com/INL/galahad-taggers-dockerized



Annotation: PoS | Group by: PoS | Single/multiple analysis: Both

TAGGER	MACRO PRECISION	MACRO RECALL	MACRO F1	MICRO ACCURACY	DETAILED EVALUATION
hug-tdn-all-enhanced	0.73	0.69	0.70	0.96	Details
hug-tdn-1600-1900	0.72	0.69	0.70	0.95	Details
hug-tdn-all	0.71	0.69	0.70	0.95	Details
hug-tdn-1400-1600	0.52	0.51	0.51	0.92	Details
pie-tdn-all	0.72	0.70	0.70	0.92	Details

Annotation: PoS | Group by: PoS | Single/multiple analysis: Multiple

TAGGER	MACRO PRECISION	MACRO RECALL	MACRO F1	MICRO ACCURACY	DETAILED EVALUATION
pie-tdn-all	0.44	0.43	0.44	0.66	Details
hug-tdn-1600-1900	0.42	0.37	0.39	0.66	Details
hug-tdn-all-enhanced	0.43	0.37	0.38	0.64	Details
hug-tdn-all	0.40	0.37	0.38	0.64	Details
hug-tdn-1400-1600	0.23	0.22	0.22	0.57	Details

Figure 1: "Best" *Couranten corpus*[42] tagger depending on evaluation type and token type

8 LAnCeLoT: Linguistic Annotation Corpus Laundry Tool

LAnCeLoT is an online tool[43] (service) to manually verify and correct corpora that are linguistically annotated with part of speech and lemma. The core characteristic of the tool is that manual verification is not done token by token in running text, but at both the type and token level. Users are presented with a list of types with the accompanying annotations occurring in the corpus. For each type, the corresponding concordances can be listed to enable verification of the analyses and their correction at token level. When working at token level, one can make a subselection of tokens (concordances) and assign a specific annotation to every instance in this subselection all at once. An example of this bulk annotation can be found in figure 3: the six occurrences type of the *accoord* can be simultaneously tagged and verified in a single action. As additional support to the annotator, suggestions for possible analyses are provided for each type using the historical computational lexicon GiGaNT-HILEX[44], which covers Dutch from ca. 1250 to 1976. The current version of GiGaNT-HILEX contains the lexicon modules based on the Dictionary of the Dutch Language (Woordenboek der Nederlandsche Taal, WNT) and the Dictionary of Middle Dutch (Middelnederlandsch Woordenboek, MNW) . GiGaNT Hilex follows the linguistic principles for part of speech tagging (TDN) and lemmatisation as discussed in section 4.

LAnCeLoT works as a standoff annotation tool on data indexed by BlackLab, the api of which is used both to construct the PostgreSQL database in which the corrected annotations are stored and to retrieve the concordance for display. Therefore, in order to use LAnCeLoT, the corpus that needs manual verification has to be uploaded in LAnCeLoT Search, a customised version of the AutoSearch application which allows nontechnical linguistic researchers to index and search their own data with BlackLab as a corpus search engine. One can navigate from a concordance to LAnCeLoT Search to present more context when this is necessary to assign the correct annotation to a certain token. LAnCeLoT Search can also be used to search within the automatically annotated corpus. For both import and export, LAnCeLoT currently supports the TEI p5 format used by GaLAHaD.

[42] https://couranten.ivdnt.org
[43] It is a successor of CoBaLT (Kenter et al., 2012)
[44] https://ivdnt.org/corpora-lexica/gigant/

9 Technical infrastructure architecture

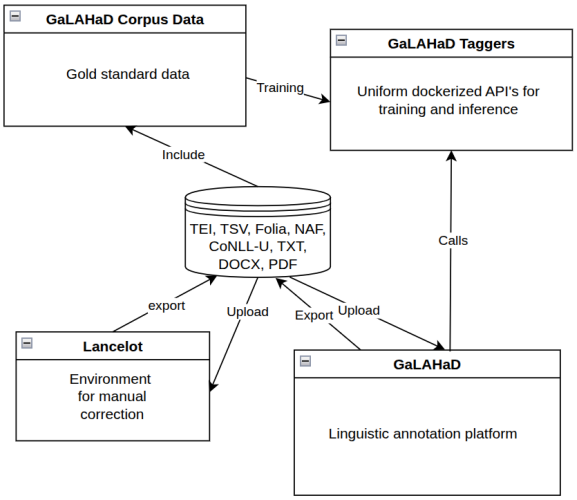


Figure 2: Infrastructure architecture

GaLAHaD consists of a Kotlin Spring Boot backend, a REST backend with Swagger UI API documentation, and a Vue 3 frontend. The backend supports the parsing of various file formats via Aalto XML[45] and other parsers, enabling the extraction of plain text, lemmas, and POS annotations. Parsed data and corpus metadata is stored on disk as JSON, together with the original files. The backend communicates with various dockerized tagger-lemmatizers that turn plaintext into annotated TSV. This protocol has been standardised as *GaLAHaD Taggers Dockerized*[46]. Currently, all taggers are implemented in Python and use Bottle for their web service, but any web service wrapper implementing the simple communication protocol suffices.

The backend is then capable of either exporting these newly produced annotations to various formats, or merging it with the original encoding of the document. As mentioned in section 7, annotation layers of various taggers and the source document can be compared and evaluated. As token-level evaluation can be computationally expensive on large corpora, GaLAHaD uses Ben Manes' Caffeine cache[47] and stores intermediate document results as JSON to disk. As documents are processed independently, multiprocessing is used to improve performance.

LAnCeLoT is built on the Lex'it (Lexicon Interactive Tool)[48] database application development platform, the BlackLab[49] corpus search engine, a Java Jakarta RESTful Web Service, and the PostgreSQL database. File uploading and corpus querying is handled by BlackLab. TEI files exported by GaLAHaD can be uploaded in the LAnCeLoT Search environment, a reskin of the BlackLab frontend. Annotations are corrected in the Lex'it frontend, which is extended for LAnCeLoT using JavaScript/jQuery. Via its Jakarta web service, LAnCeLoT keeps track of lemma and POS annotations that have been corrected or validated in its work table in the PostgreSQL database, and can export these by merging them with the original TEI encoding.

For creating our GaLAHaD gold standard corpus data, we use a couple of python scripts to detect duplicates, split data into machine learning sets, and generate statistics. Once a corpus is exported from LAnCeLoT, it is converted to TSV using the GaLAHaD webservice. We provide both TEI and TSV, as TSV is the preferred data format for machine learning.

Next we check for files with textual overlap. These may be the A and B version of a manuscript, or two news articles in different newspapers based on the same correspondence source. Overlaps are detected

[45] https://github.com/FasterXML/aalto-xml
[46] https://github.com/instituutnederlandsetaal/galahad-taggers-dockerized
[47] https://github.com/ben-manes/caffeine
[48] https://github.com/instituutnederlandsetaal/Lexit
[49] https://blacklab.ivdnt.org/

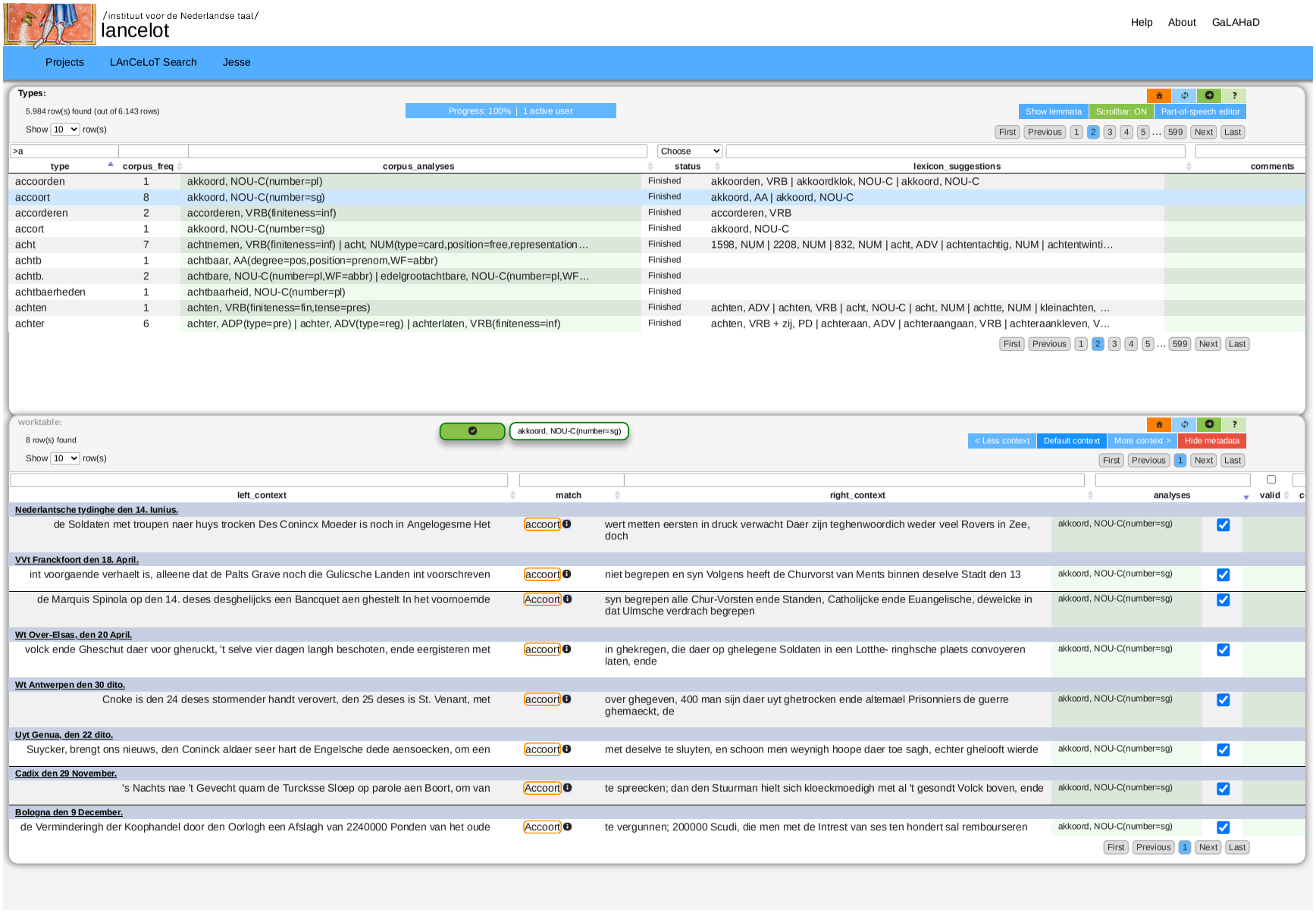


Figure 3: Lancelot: bulk tagging of the word *accoord* in the Couranten training corpus

based on edit distance using the Edlib[50] library. A text is first converted to a single string of alphabetical characters (that includes removing spaces). The string is then split into chunks of 20 characters by default. For each chunk, we then check if it occurs in any other document, in any position, within an edit distance of 5. If a high percentage (e.g. 90%) of all chunks in a document occurs in another document as well, the two are considered duplicates.

The corpus is then split into a train, test, and validation set such that any duplicates are grouped into the same set to avoid bias. Provenance of the files per split is stored in a JSON file. Finally, we calculate some statistics about the corpus, such as token size per split, century, and subcorpus, as well as the frequencies of each lemma and POS annotation. We also try to detect suspicious annotations: annotations that are unlikely to occur. If a certain token is assigned a POS of NOU-C a thousand times, and VRB only once, chances are it was tagged incorrectly. This feedback is then processed by the annotators.

10 Future work

In the SSHOC-NL project, we are working on extending the platform with more linguistic annotation tools and evaluation methods.

We are extending the annotation tools to modern Dutch by integrating spaCy and Stanza. Simultaneously, we are adding support for annotations other than lemma and PoS, namely named entities and dependency relations as they are produced by spaCy and Stanza models.

The existing evaluation methods will be extended to support the new annotation types. Additionally, we are working on new evaluation methods for span-based annotations (e.g. named entities), evaluation metrics grouped by frequency (e.g. hapax accuracy), and document-level evaluation (e.g. text-inline visualisation of errors).

We will extend the gold standard data with data from the Corpus of Middle Dutch[51] and contemporary corpus data.

Of course, we hope that the state of the art in historical language processing will continue to improve to yield better taggers, lemmatizers and parsers to be integrated in the infrastructure. Working with the spaCy, Stanza and UDPipe models for modern Dutch, we found that the lemmatization accuracy of these models is often below par for less frequent lemmata, and can be improved by using the GiGaNT-Molex[52] morphosyntactic lexicon of modern Dutch. We will develop a lemmatizer that incorporates the lexicon.

11 Acknowledgments

The team behind the infrastructure consists, besides the authors, of Vincent Prins, Roland de Bonth, Tim Brouwer, Mathieu Fannee and Thomas Haga, all INT. Since 2024 Bram Vanroy (INT), Eleanor Smith and Antske Fokkens from the VU have also been involved in the further development of GaLAHaD. The work is carried out with funding from the Netherlands Organisation for Scientific Research (NWO) in the CLARIAH-PLUS project (Grant 184.034.023) and the SSHOC-NL project (Grant 184.036.020) , with further support from the Dutch Language Union.

References

de Does, J., Niestadt, J., & Depuydt, K. (2017). Creating research environments with BlackLab. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries*. Ubiquity Press.

Depuydt, K., Haga, T., & Mooijaart, M. (2024). *Lemmatiseerprincipes voor GiGaNT, het centrale lexicon van het INT*. https://ivdnt.org/wp-content/uploads/2024/11/lemmatiseerprincipesV2_combi.pdf

Haga, T., Depuydt, K., de Does, J., de Bonth, R., & Geirnaert, D. (2024). *Tagset voor Diachroon corpusmateriaal van het Nederlands (TDN)*. https://ivdnt.org/wp-content/uploads/2024/11/TDNV2_combi.pdf

[50] https://github.com/Martinsos/edlib

[51] https://corpusmiddelnederlands.ivdnt.org/

[52] https://taalmaterialen.ivdnt.org/download/gigant-molex2-0/

Kenter, T., Erjavec, T., Žorga Dulmin, M., & Fišer, D. (2012, April). Lexicon construction and corpus annotation of historical language with the CoBaLT editor. In K. Zervanou & A. van den Bosch (Eds.), *Proceedings of the 6th workshop on language technology for cultural heritage, social sciences, and humanities* (pp. 1–6). Association for Computational Linguistics. https://aclanthology.org/W12-1001/

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., & Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation [Event Title: The 27th International Conference on Computational Linguistics (COLING 2018)]. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. http://tubiblio.ulb.tu-darmstadt.de/106270/

Manjavacas, E., Kádár, Á., & Kestemont, M. (2019). Improving lemmatization of non-standard languages with joint learning. *arXiv preprint arXiv:1903.06939*.

Ruitenberg, T., van pellicom K., de Does, J., & Depuydt, K. (2012). De morfosyntactische module van het GiGaNT-lexicon. *INL Working Papers - Taalbank Nederlands*.

Schmid, H. (2019). Deep learning-based morphological taggers and lemmatizers for annotating historical texts. *Proceedings of the 3rd international conference on digital access to textual cultural heritage*, 133–137.

Tjong Kim Sang, E., Bollmann, M., Boschker, R., Casacuberta, F., Dietz, F., Dipper, S., Domingo, M., van der Goot, R., van Koppen, M., Ljubešić, N., et al. (2017). The CLIN27 shared task: Translating historical text to contemporary language for improving automatic linguistic annotation. *Computational Linguistics in the Netherlands Journal*, *7*, 53–64.

van Dalen Oskam, K., & Depuydt, K. (2000). Lemmatisering en codering in het VMNW-corpus II Codering in het VMNW-woordenboek. *nstituut voor Nederlandse Lexicologie, Internal report*.

van den Bosch, A., Busser, B., Canisius, S., & Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. *LOT Occasional Series*, *7*, 191–206.

van der Sijs, N., van Kemenade, A., & Rem, M. (2018). Corpus laatmiddel- en vroegnieuwnederlands (clvn) (onderdeel Nederlab).

Van Eynde, F. (2005). Part of Speech tagging en lemmatisering van het D-COI corpus. *Centrum voor Computerlinguïstiek K.U.Leuven. Report.*

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., & Raffel, C. (2022). ByT5: Towards a token-free future with pre-trained byte-to-byte models (B. Roark & A. Nenkova, Eds.). *Transactions of the Association for Computational Linguistics*, *10*, 291–306. https://doi.org/10.1162/tacl_a_00461