TextLens & LeTTuce: Automated Corpus Annotation and Multilingual Tagging as a Service

Cynthia Van Hee[‡], Jonas Doumen[*], Vincent Prins[†], Pranaydeep Singh[‡], Vincent Vandeghinste[*†] and Els Lefever[‡]

[*] University of Leuven, Belgium; [†] Instituut voor de Nederlandse Taal, The Netherlands;
[‡] LT3, Ghent University, Belgium
first.lastname@kuleuven.be, first.lastname@ivdnt.org, first.lastname@ugent.be

Abstract

We present **TextLens**, a web-based platform for automated linguistic annotation designed to lower technical barriers for researchers in digital humanities, linguistics and translation studies. Hosted by the Dutch Language Institute (INT), TextLens allows users to upload and annotate corpora in a variety of formats (.txt, .tsv, CoNLL-U, FoLiA, TEI, and NAF) using state-of-the-art NLP tools, without the need for local installation or computational resources. The platform supports multilingual data processing and provides a persistent dashboard for managing, monitoring and sharing annotation projects. Alongside this service, we introduce the **LeTTuce-PoS Dataset**, a new multilingual, manually annotated dataset for part-of-speech tagging in English, French, Dutch and German, covering multiple genres and offering a valuable resource to the research community. This paper also reports benchmark results for different PoS taggers (LeTs Preprocess, LeTTuce, spaCy and Stanza) on the dataset. Together, TextLens and the LeTTuce-PoS Dataset provide an accessible, scalable platform for high-quality annotation and a robust multilingual dataset that support comparable and reproducible research in multilingual contexts.

Keywords: Digital Humanities, linguistic annotation, resource creation, digital text analysis, preprocessing, PoS tagging, Lemmatising, Named Entity Recognition, Dependency Parsing, Web interface, online NLP tools

1. Introduction

In the field of Natural Language Processing (NLP), numerous tools have been developed to address diverse tasks. With the rapid expansion of user communities and application domains, comparative evaluations have become essential for identifying the most suitable tool for a given dataset or language. To support such evaluations and facilitate replication studies, we introduce **TextLens**, a scalable platform for data annotation, together with a collection of resources that enable comparative analysis and foster experimentation within the Digital Humanities community.

TextLens is a web-based platform designed to lower the technical and usability barriers that researchers in the digital humanities and linguistics face when applying automatic annotation tools. Such tools, including part-of-speech (PoS) taggers, lemmatizers, named entity recognizers and syntactic parsers, are vital for analyzing large corpora but often require programming expertise and substantial computational resources. Transformer-based models in particular demand large amounts of storage and memory, frequently monopolizing local machines and limiting accessibility for non-technical users.

TextLens addresses these challenges by offering a browser-based interface for advanced linguistic analysis without the need for local installation or configuration. Hosted by the *Instituut voor de Nederlandse Taal*[1] (INT), a technical centre in the CLARIN infrastructure (Eskevich et al., 2020), it allows researchers to upload data, run annotation tasks and download enriched corpora in structured formats such as CoNLL-U (Nivre et al., 2016), FoLiA (van Gompel and Reynaert, 2013), TEI (Ide and Véronis, 1995) and NAF (Fokkens et al., 2014).

The novelty of TextLens lies in integrating state-of-the-art annotation pipelines into an intuitive dashboard that prioritizes accessibility and transparency. Users can monitor processing progress, inspect results and share annotated corpora with collaborators via email-based access control. This makes TextLens particularly valuable for digital humanities scholars, linguists and translation researchers working with multilingual data who require cross-linguistic annotations.

With the **LeTTuce-PoS Dataset**, we introduce a novel multilingual corpus designed for the digital humanities community and for researchers interested in evaluating PoS tagging models. We also provide benchmark results from experiments comparing the performance of established taggers: spaCy (Honnibal and Montani, 2017), Stanza (Qi et al., 2020), LeTs Preprocess (Van de Kauter et al., 2013), and our novel model LeTTuce (Van Hee et al., 2025), on the same datasets.

Together, TextLens and the LeTTuce-PoS Dataset align with the goals of infrastructures like CLARIN by helping to **democratize access to language technologies and enabling research**

[1] Dutch Language Institute

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11574–11584
11–16 May 2026.

replication and comparison. They offer a streamlined, cross-platform environment for reproducible, collaborative and data-driven research, supported by a robust benchmark dataset.

The remainder of this paper is structured as follows. Section 2 provides a brief overview of related work. Section 3 presents the TextLens platform, describing its functionality and evaluation. Section 4 introduces the LeTTuce-PoS Dataset, a new multilingual corpus used in this study to benchmark the performance of established PoS taggers. Finally, Section 5 concludes the paper and discusses future work directions.

2. Related Work

A variety of platforms have been developed to make linguistic annotation tools more accessible. The CLARIN Language Resource Switchboard (Zinn, 2018) recommends suitable tools after automatically identifying the language of the input data. WebLicht (Hinrichs et al., 2010) enables users to construct annotation workflows through a graphical interface, supporting complex pipelines without scripting. Text Tonsorium (Jongejan, 2021) provides a workflow management system that integrates modules for state-of-the-art NLP tools, with automated workflow design. UDPipe (Straka, 2018) offers pre-trained models for over 100 languages and is available through a web interface, command line and API, making it one of the most widely used solutions.

Building on this landscape, TextLens provides a persistent dashboard where users can upload, process, monitor and share corpora without maintaining an active browser session. It supports multiple data formats, includes automatic conversion and simplifies the reuse and distribution of annotated corpora, thereby facilitating both individual and collaborative research workflows. TextLens is a fork of GaLAHaD (Depuydt and de Does, 2025),[2] originally developed for historical Dutch texts, and inherits its features for handling various file formats, file uploads and exports, and authentication.

3. TextLens Functionality

3.1. Using TextLens

TextLens is accessible at `http://hdl.handle.net/10032/tm-a3-c4`, with authentication provided via the CLARIN federated login system, which covers thousands of institutions worldwide.[3]

As shown in Appendix 8.1, users start by creating a new corpus and uploading data in commonly used formats, including plain text (`.txt`), tab-separated values (`.tsv`), CoNLL-U, FoLiA, TEI and NAF. They subsequently select one or more annotation tools to initiate processing tasks. The dashboard, illustrated in Figure 1, provides real-time progress updates and allows users to download completed annotations in multiple formats.

In contrast to other online annotation platforms, TextLens eliminates the need for a persistent session. Users can close their browser or log out after initiating a task and return later to continue annotating or to access their results. Both corpora and their corresponding annotations can also be easily shared with collaborators by entering their email addresses, which fosters collaboration and encourages data reuse.

Tool	PoS	Lemma	NE	Dep
spaCy	✓	✓	✓	✓
Stanza	✓	✓	✓	✓
Lets	✓	✓	✗	✗
LeTTuce	✓	✗	✗	✗

Table 1: Functionality of the tools currently included in TextLens. NE stands for *Named Entity Recognition*, Dep for *Dependency Relations*.

3.2. Architecture

TextLens is built on a modular, containerized architecture. Both front-end and back-end components run in Docker containers, as do the annotation tools, which communicate via the standardised *Taggers-Dockerized* API developed by the GaLAHaD team. This design facilitates easy deployment of custom and local instances.

New annotation tools can be added by extending a simple Python template provided by the *TaggersDockerized* framework. The front-end is developed in Vue.js and the back-end in Kotlin. A fully documented API supports integration with external tools and automated testing. The source code of TextLens is available under the open source Apache 2.0 license on GitHub: `https://github.com/CCL-KULeuven/textlens`.

Table 1 lists the tools currently available in TextLens, along with their respective functionalities. Additional tools and functionalities are planned for inclusion in the coming years.

3.3. Usability Evaluation

To assess the usability and performance of TextLens, we conducted internal tests on processing speed, task reliability and output consistency across different formats. Annotation tasks were tested using corpora in Dutch, English, French and German across a variety of genres, including re-

[2] `http://hdl.handle.net/10032/tm-a3-b2`

[3] External users can create a guest account through the same login interface.

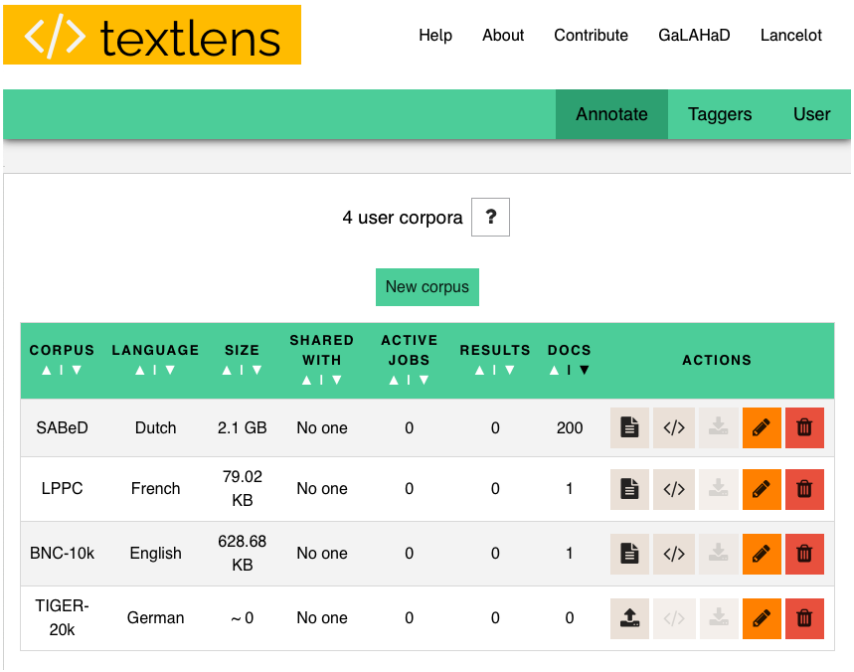


Figure 1: The TextLens dashboard.

views, newswire and social media texts, and historical texts. Preliminary feedback from pilot studies with digital humanities scholars and translation researchers has been positive, highlighting the platform's ease of use, format flexibility and the advantages of remote processing. Ongoing user studies will evaluate long-term adoption and impact in real-world research scenarios, especially as the platform's tools and features continue to expand.

4. The LeTTuce-PoS Dataset: A New Multilingual Benchmark Corpus for PoS Tagging

To address persistent challenges in part-of-speech tagging, we present the LeTTuce-PoS Dataset, a new multilingual benchmark corpus designed to facilitate systematic evaluation across languages, genres and domains. The corpus covers English, French, Dutch and German and comprises hand-annotated data drawn from a variety of sources. For the reviews domain, texts were collected from platforms including bol.com, Trustpilot and TripAdvisor (representing the FMCG[4], airline, and hotel sectors, respectively). Technical data consist of in-house datasets previously provided by industry partners, covering sectors such as dredging and human resources. For the social domain, tweets were scraped from X (formerly Twitter). Finally, for Dutch historical data, texts were retrieved from the Guido Gezelle Archief, a Flemish initiative that digitizes Gezelle's correspondences and makes them publicly accessible.

By combining high-quality annotation with cross-linguistic and cross-genre diversity, the LeTTuce-PoS Dataset provides a valuable resource for assessing existing tools and fostering reproducible, comparable research in multilingual contexts.

4.1. State-of-the-art in PoS Tagging

Part-of-speech (PoS) tagging is a foundational task in NLP, supporting downstream applications such as syntactic parsing, information extraction, machine translation, sentiment analysis and corpus linguistics. Approaches to PoS tagging have evolved considerably. Early systems relied on rule-based and statistical models, including Hidden Markov Models (HMMs), Maximum Entropy models and Conditional Random Fields (CRFs). With the rise of deep learning, recurrent neural networks (RNNs), long short-term memory networks (LSTMs) and Transformer architectures substantially improved context modeling and the handling of morphological complexity and unknown words (e.g., Yang et al., 2018; Pota et al., 2019). Transfer learning further enhanced performance in noisy, low-resource and historical settings (e.g., Meftah and Semmar, 2018; Kim et al., 2017; Smidt et al., 2024; Szawerna, 2024). More recently, large language models (LLMs) have been explored for PoS tagging, leveraging their broad contextual knowledge

[4]FMCG, or Fast-moving Consumer Goods, refers to products that you can sell quickly at relatively low cost. Examples include food items, drinks and toiletries.

to improve robustness across language varieties and under-represented domains (e.g., Fang et al., 2025; Subedi et al., 2024).

Despite these advances, PoS tagging continues to face challenges and remains a key source of parsing errors (Foster et al., 2011). Persistent difficulties include: noisy web and social media text (Foster et al., 2011; Giesbrecht and Evert, 2009), morphologically rich languages, limited availability of annotated data for under-represented languages, creative or non-standard language usage, code-switching or code-mixing and historical texts.

To address these challenges, we introduce the LeTTuce-PoS Dataset, a multilingual dataset for PoS tagging, and benchmark it against a number of established taggers. Our evaluation spans traditional sequence labeling methods and neural architectures across multiple languages, genres and domains, with the aim of identifying performance gaps and highlighting remaining obstacles for future research.

4.2. Data Collection and Annotation

The data for the LeTTuce-PoS Dataset were collected in two ways: (i) through collaborative research projects with industry and institutions (technical and historical data) and (ii) by crawling the web (social media and review data). Table 2 provides an overview of token counts per genre and language. At present, only technical data have been collected for German and historical data only for Dutch, but additional data are planned and will be released in future corpus updates.

Language	Reviews	Social media	Technical	Historical
English	11,186	4527	17,418	-
French	5646	2896	19,081	-
German	-	-	8698	-
Dutch	16,961	4570	16,517	1582

Table 2: Token counts per language and domain in the LeTTuce-PoS Dataset.

The PoS annotation was performed by three trained linguists with C1 or C2 proficiency (CEFR level) in the target language, between August 28 and September 18, 2023. The annotators received training and supervision from the first author and were encouraged to discuss problematic cases among themselves or seek guidance when needed. Annotations were regularly spot-checked by the supervisor and ambiguous cases were resolved through group discussion to ensure consistency. Language-specific tag sets were applied: the Penn Treebank for English (Marcus et al., 1993), the French TreeTagger for French (Schmid, 1994), the Stuttgart-Tübingen Tag Set (STTS) for German (Schiller et al., 1999) and the CGN tag set for Dutch (Van Eynde, 2003).

4.3. The LeTTuce-PoS Dataset: A Benchmark Dataset for Part-of-Speech Tagging Evaluation

The newly compiled dataset was used to benchmark the PoS tagging performance of several tools, including LeTs Preprocess (Van de Kauter et al., 2013), Stanza (Qi et al., 2020), spaCy (Honnibal and Montani, 2017) and LeTTuce (Van Hee et al., 2025).

For LeTTuce, the training data consist of newswire texts in four languages, drawn from the Dutch Parallel Corpus (NL, FR, EN) (Paulussen et al., 2013), the Lassy Small Treebank (NL) (van Noord et al., 2013), the Penn Treebank (EN) (Marcus et al., 1993), and the TIGER Treebank (DE) (Brants et al., 2002). Further details are provided in Van de Kauter et al. (2013).

The training corpus was used to fine-tune two pre-trained models per language: a language-specific BERT and a cross-lingual XLM encoder (Conneau and Lample, 2019), each extended with a linear classification layer for token-level tagging.[5]

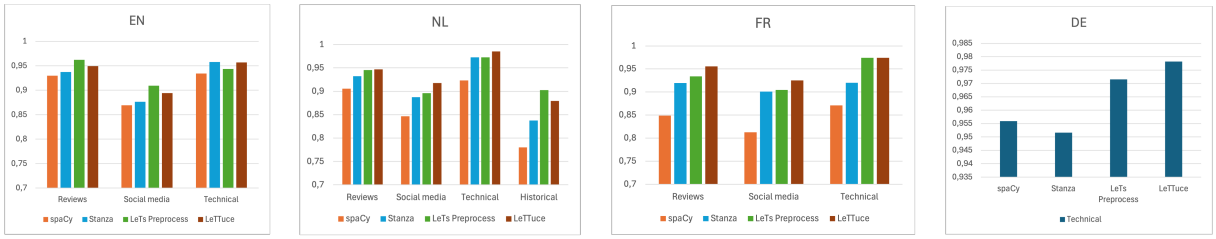
They were trained on 4 Nvidia A-100 GPUs with a batch size of 128 per GPU for 40 epochs with a starting learning rate of $5e^{-5}$ for 500 warm-up steps followed by a linear decay. The maximum sequence length of all models was restricted to 128 tokens, since less than 0.01% of the training data exceeded the 128 length limit.

Training was performed using the native, language-specific tag sets provided by the respective treebanks. To enable fair comparison with spaCy and Stanza, their tags were mapped to UD, as was done for the tags in our gold standard benchmark. Details of this mapping, along with the full tag sets, can be found in Appendix 8.2. Overall results per language are summarized in Tables 3–5, with genre-specific scores displayed in Figure 2.

Since performance differences between the cross-lingual XLM encoders and the language-specific BERT models were not statistically significant ($p > 0.05$), we only report the results of the language-specific models in this study. Future work will extend our benchmarking to additional languages, at which point we will further investigate the trade-off between cross-lingual and language-specific models. For Dutch, only coarse-grained results are reported to maintain comparability with the other languages and models.

Performance differences across PoS tagging tools and data genres are shown in Figure 2. The bar charts reveal a consistent drop in accuracy on social media data across all languages, as well as on historical texts for Dutch. By contrast, technical texts yield the highest tagging performance overall.

[5]The models are available at https://github.com/lt3/Lettuce.



(a) English (b) Dutch (c) French (d) German

Figure 2: Tagging performance across data domains for four PoS tagging toolkits.

English

PoS tagger	Reviews	Social	Technical
spaCy	0.930	0.870	0.934
Stanza	0.937	0.876	**0.958**
LeTs Prepro	**0.962**	**0.909**	0.944
LeTTuce	0.949	0.894	0.957

Table 3: Cross-domain performance (F_1) comparison between state-of-the-art PoS taggers for English. High scores per domain are in bold.

Dutch

PoS tagger	Rev.	Soc.	Tech.	Historical
spaCy	0.906	0.847	0.923	0.780
Stanza	0.933	0.887	0.973	0.837
LeTs Prepro	0.945	0.896	0.973	**0.903**
LeTTuce	**0.947**	**0.918**	**0.985**	0.879

Table 5: Cross-domain performance (F_1) comparison between state-of-the-art PoS taggers for Dutch.

French

PoS tagger	Reviews	Social	Technical
spaCy	0.849	0.813	0.871
Stanza	0.920	0.901	0.920
LeTs Preprocess	0.934	0.905	**0.974**
LeTTuce	**0.956**	**0.925**	**0.974**

Table 4: Cross-domain performance (F_1) comparison between state-of-the-art PoS taggers for French.

German

PoS tagger	Technical
spaCy	0.956
Stanza	0.952
LeTs Prepro	0.972
LeTTuce	**0.978**

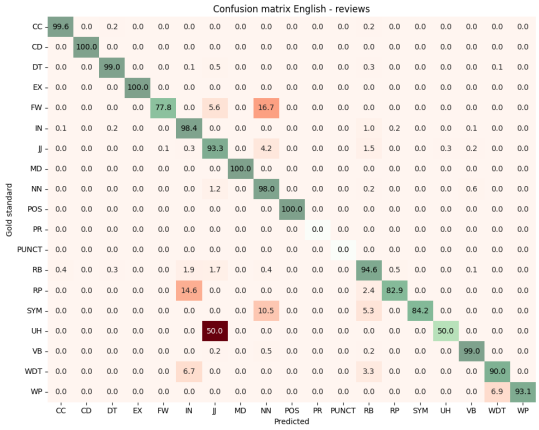
Table 6: Performance (F_1) comparison between state-of-the-art PoS taggers for German.

Across tools, the LeTTuce tagger and LeTs Preprocess consistently achieve the best results, outperforming the established taggers spaCy and Stanza on our benchmark dataset. Since the training procedures behind spaCy and Stanza are not fully documented, it is difficult to pinpoint the exact cause of their lower robustness. Nonetheless, the findings suggest that both the LeTTuce tagger and LeTs Preprocess generalize more effectively across diverse genres. This comparison not only provides insights into PoS tagging accuracy across tools, but also supports NLP practitioners in selecting suitable taggers depending on the language and domain at hand.

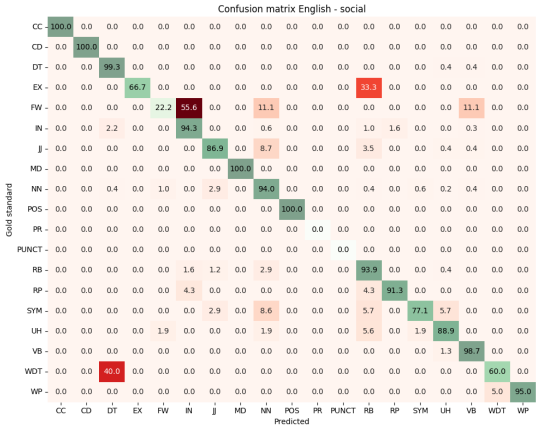
To obtain more fine-grained insights into PoS tagging performance across domains, we further examined the output of our best-performing model, LeTTuce. Figure 3 shows the normalized confusion matrix for all English subdomains in the corpus. To improve readability, PoS categories with few or no confusions (e.g., verb past participle and verb gerund) were merged into broader classes. This prevents overemphasis on noise from rare tags and instead highlights more meaningful confusions between categories. The diagonal cells (highlighted in green) represent correctly predicted tags, while the most prominent off-diagonal cells point to systematic misclassifications. All values are normalized (percentages instead of absolute counts) to enable fair comparisons across domains.

The confusion matrices demonstrate that, across all domains, common PoS tags such as CC (coordinating conjunction), CD (cardinal number), POS (possessive ending), DT (determiner), IN (preposition or subordinating conjunction) and MD (modal) are predicted with high accuracy (98–100%). The domains *social* and *reviews*, however, display greater variability and higher confusion rates compared to *technical*, which aligns with their lower overall prediction accuracy (see Table 3).

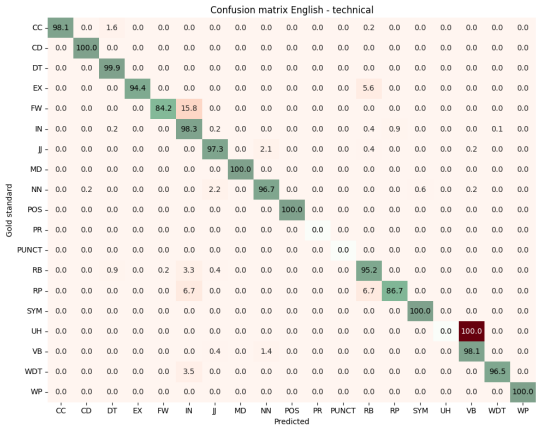
This reduced performance is linked to more informal language use. For example, higher confusion was observed for the tags FW (foreign words) and UH (interjections) in social and review data. These domains also show moderate to high confusion (6.7-40%) between WDT (wh-determiner) and DT (determiner). Interestingly, in the technical domain, all 7 instances of “please” (UH - interjection) were misclassified as verbs (VB).



(a) Confusions Reviews



(b) Confusions Social



(c) Confusions Technical

Figure 3: Confusion matrices for the English subdomains, highlighting the off-diagonal cells with high confusion.

The LeTTuce-PoS benchmark dataset is distributed via the CLARIN infrastructure[6] and released under the Creative Commons License CC BY-NC-ND 4.0.

[6]Available at http://hdl.handle.net/10032/tm-a3-d2.

5. Conclusion and Future Work

In this paper, we introduced TextLens, a flexible and user-friendly platform for linguistic annotation that bridges the gap between advanced NLP tools and non-technical users. Its centralized, browser-based design ensures cross-platform accessibility and supports collaborative and reproducible research. Future developments will include additional annotation layers such as coreference resolution and semantic role labeling, expanded language coverage, enhanced visualization features and broader user evaluation.

In this paper, we also introduced the LeTTuce-PoS Dataset, a new multilingual dataset for benchmarking PoS tagging. Future work will expand the corpus to include additional languages and more diverse data domains (e.g., newswire text, legal documents, creative or literary texts). We will also extend the benchmarking study accordingly, enabling a more comprehensive evaluation of NLP tools and techniques across varied contexts.

Ethical considerations

All data used in this study were obtained from publicly available sources or were collected under formal collaboration agreements. No sensitive data were included. Annotation was carried out by trained linguists, who were compensated and supervised appropriately. The authors are not aware of any ethical risks associated with the release of the LeTTuce-PoS Dataset. However, as with any linguistic resource, biases present in the source material may affect downstream applications.

Limitations

A current limitation of TextLens is its default application of sentence segmentation, as is standard in tools such as spaCy and Stanza. However, this segmentation is not consistently visible across all output formats, being explicitly marked only in CoNLL-U. For future work, we aim to give users more control over this setting and ensure uniform sentence boundary representation across all supported formats. Another limitation is that TextLens currently activates and exports all annotation layers (i.e., lemmas, PoS tags, NE tags, and dependency relations) by default. Future versions will introduce finer-grained control over layer selection, thereby giving users more flexibility in configuring the output files.

For the LeTTuce-PoS Dataset, only four languages and domains are currently included. Expanding coverage to additional languages and domains is planned for future releases.

Acknowledgements

TextLens is based on GaLAHaD, a tool for the automated linguistic annotation of historical Dutch, created by the Dutch Language Institute. The development of TextLens was created in the CLARIAH-VL project funded by the FWO International Research Infrastructures with grant number I000921N as part of the CLARIAH-VL project. GaLaHaD in its turn is funded by NWO under the CLARIAH plus project (grant number 184.034.023) and the SSHOC-NL project (grant number 184.036.020). The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation Flanders (FWO) and the Flemish Government – department EWI.

6. Bibliographical References

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Katrien Depuydt and Jesse de Does. 2025. An infrastructure for Historical Dutch Corpus Development. In *CLARIN Annual Conference Proceedings*.

Maria Eskevich, Franciska de Jong, Alexander König, Darja Fišer, Dieter Van Uytvanck, Tero Aalto, Lars Borin, Olga Gerassimenko, Jan Hajic, Henk van den Heuvel, Neeme Kahusk, Krista Liin, Martin Matthiesen, Stelios Piperidis, and Kadri Vider. 2020. CLARIN: Distributed language resources and technology in a European infrastructure. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 28–34, Marseille, France. European Language Resources Association.

Zhao Fang, Liang-Chun Wu, Xuening Kong, and Spencer Dean Stewart. 2025. A comparative analysis of word segmentation, part-of-speech tagging, and named entity recognition for historical Chinese sources, 1900-1950. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 1–6, Albuquerque, USA. Association for Computational Linguistics.

A.S. Fokkens, A. Soroa, Z. Beloki, C.J. Ockeloen, G. Rigau, W.R. van Hage, and P.T.J.M. Vossen. 2014. Naf and gaf: Linking linguistic annotations. In *Proceedings of the 10th Joint ACL – ISO Workshop on Interoperable Semantic Annotation*. NAF and GAF: Linking Linguistic Annotations.

Jennifer Foster, Özlem Çetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011. #hardtoparse: POS tagging and parsing the twitterverse. In *Analyzing Microtext, Papers from the 2011 AAAI Workshop, San Francisco, California, USA, August 8, 2011*, volume WS-11-05 of *AAAI Technical Report*. AAAI.

Eugenie Giesbrecht and Stefan Evert. 2009. Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, San Sebastian, Spain.

Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29, Uppsala, Sweden. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Nancy Ide and Jean Véronis, editors. 1995. *Text Encoding Initiative: Background and Context*. Springer, Dordrecht.

Bart Jongejan. 2021. The CLARIN-DK Text Tonsorium. In *Selected papers from the CLARIN Annual Conference 2020*, pages 111–121.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838, Copenhagen, Denmark. ACL.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Sara Meftah and Nasredine Semmar. 2018. A neural network model for part-of-speech tagging of social media texts. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo

Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Marco Pota, Fiammetta Marulli, Massimo Esposito, Giuseppe De Pietro, and Hamido Fujita. 2019. Multilingual pos tagging by a composite deep architecture based on character-level features and on-the-fly enriched word embeddings. *Knowledge-Based Systems*, 164:309–323.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the ACL: System Demonstrations*.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das tagging deutscher textcorpora mit stts. In *Technical Report, University of Stuttgart and University of Tübingen*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Gustav Ryberg Smidt, Els Lefever, and Katrien de Graef. 2024. At the crossroad of cuneiform and NLP: Challenges for fine-grained part-of-speech tagging. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1745–1755, Torino, Italia. ELRA and ICCL.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Bipesh Subedi, Sunil Regmi, Bal Krishna Bal, and Praveen Acharya. 2024. Exploring the potential of large language models (LLMs) for low-resource languages: A study on named-entity recognition (NER) and part-of-speech (POS) tagging for Nepali language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6974–6979, Torino, Italia. ELRA and ICCL.

Maria Irena Szawerna. 2024. Can stanza be used for part-of-speech tagging historical Polish? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 44–49, St. Julian's, Malta. Association for Computational Linguistics.

Marjan Van de Kauter, Geert Coorman, Els Lefever, Bart Desmet, Lieve Macken, and Véronique Hoste. 2013. LeTs preprocess: the multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal*, 3:103–120.

Frank Van Eynde. 2003. Part of Speech Tagging en Lemmatisering van het Corpus Gesporken Nederlands. Technical report, Centrum voor Computerlinguïstiek.

Maarten van Gompel and Martin Reynaert. 2013. Folia: A practical xml format for linguistic annotation – a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81.

Cynthia Van Hee, Pranaydeep Singh, and Els Lefever. 2025. LeTTuce PoS-Tagger : a sprout of innovation in multilingual NLP. In *DH Benelux 2025, Abstracts*.

Liner Yang, Meishan Zhang, Yang Liu, Maosong Sun, Nan Yu, and Guohong Fu. 2018. Joint POS Tagging and Dependence Parsing With Transition-Based Neural Networks. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 26(8):1352–1358.

Claus Zinn. 2018. Squib: The language resource switchboard. *Computational Linguistics*, 44(4):631–639.

7. Language Resource References

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The tiger treebank. In *Proc. of the workshop on treebanks and linguistic theories*, pages 24–41.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Hans Paulussen, Lieve Macken, Willy Vandeweghe, and Piet Desmet. 2013. Dutch parallel corpus: a balanced parallel corpus for dutch-english and dutch-french. In Peter Spyns and Jan Odijk, editors, *Essential speech and language technology*

for Dutch: results by the STEVIN-programme, Theory and Applications of NLP, pages 185–199. Springer.

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: Results by the STEVIN programme*, pages 147–164. Springer Berlin Heidelberg, Berlin, Heidelberg.

8. Appendices

8.1. Appendix A: TextLens screenshots

8.2. Appendix B: Tag sets

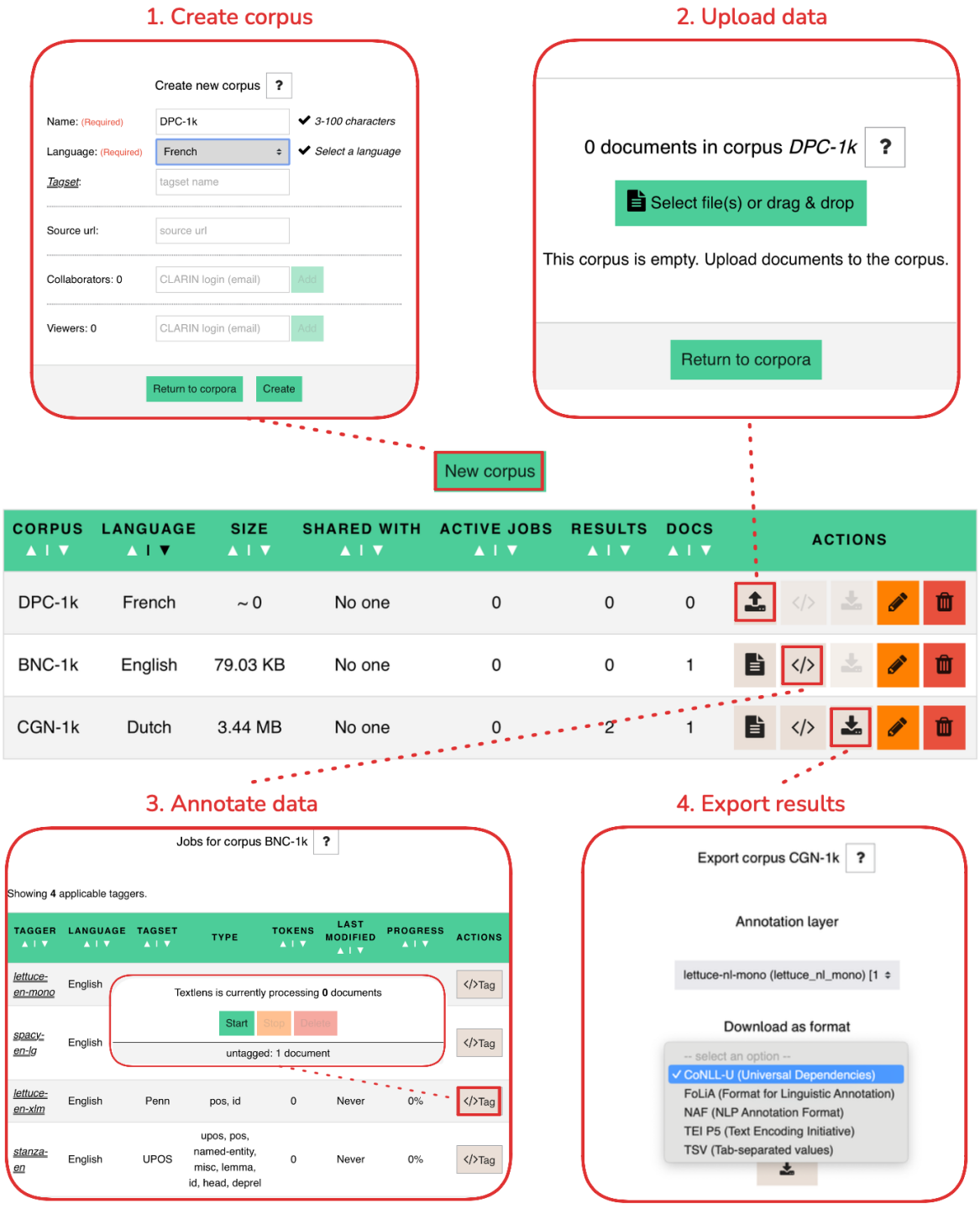


Figure 4: Appendix A - Textlens usage instructions.

ENGLISH	
Penn Treebank tag	UD tag
NN	NOUN
NNS	NOUN
NNP	PROPN
NNPS	PROPN
JJ	ADJ
JJR	ADJ
JJS	ADJ
RB	ADV
RBR	ADV
RBS	ADV
WRB	ADV
VBN	VERB
VBG	VERB
VB	VERB
VBP	VERB
VBZ	VERB
VBD	VERB
DT	DET
PDT	DET
WDT	DET
PRP	PRON
PRP$	PRON
WP	PRON
WP$	PRON
IN	ADP
CC	CCONJ
SYM	SYM
LS	SYM
CD	NUM
UH	INTJ
FW	FW
TO	ADP
RP	PART
EX	PRON
MD	VERB
POS	PART
,	PUNCT
.	PUNCT
:	PUNCT
;	PUNCT
?	PUNCT
!	PUNCT
[	PUNCT
]	PUNCT
(	PUNCT
)	PUNCT
{	PUNCT
}	PUNCT
\\	PUNCT
\	PUNCT
'	PUNCT
``	PUNCT
"	PUNCT
$	SYM
ADV	ADV
E	SYM
PROPN	PROPN
PUNCT	PUNCT

GERMAN	
STTS tag	UD tag
$(	PUNCT
$,	PUNCT
$.	PUNCT
ADJA	ADJ
ADJD	ADJ
ADV	ADV
APPO	ADP
APPR	ADP
APPRART	ADP
APZR	ADP
ART	DET
CARD	NUM
FM	FW
ITJ	INTJ
KOKOM	CCONJ
KON	CCONJ
KOUI	SCONJ
KOUS	SCONJ
NE	PROPN
NN	NOUN
PDAT	DET
PDS	PRON
PIAT	DET
PIS	PRON
PPER	PRON
PPOSAT	DET
PPOSS	PRON
PRELAT	DET
PRELS	PRON
PRF	PRON
PROAV	ADV
PTKA	PART
PTKANT	PART
PTKNEG	PART
PTKVZ	ADP
PTKZU	PART
PWAT	DET
PWAV	ADV
PWS	PRON
TRUNC	X
VAFIN	VERB
VAIMP	VERB
VAINF	VERB
VAPP	VERB
VMFIN	VERB
VMINF	VERB
VMPP	VERB
VVFIN	VERB
VVIMP	VERB
VVINF	VERB
VVIZU	VERB
VVPP	VERB
XY	SYM

FRENCH	
French TreeTagger	UD tag
NOM	NOUN
NAM	PROPN
ADJ	ADJ
ADV	ADV
VER:pper	VERB
VER:ppre	VERB
VER:infi	VERB
VER:impe	VERB
VER:pres	VERB
VER:subi	VERB
VER:impf	VERB
VER:simp	VERB
VER:subp	VERB
VER:cond	VERB
VER:futu	VERB
DET:ART	DET
PRO:PER	PRON
DET:POS	DET
PRO:POS	PRON
PRO:DEM	PRON
PRO:IND	PRON
PRO:REL	PRON
PRP	ADP
PRP:det	ADP
KON	CCONJ
ABR	SYM
SYM	SYM
NUM	NUM
INT	INTJ
FW	FW
PUN	PUNCT
PUN:cit	PUNCT
SENT	PUNCT
PROPN	PROPN
INTJ	INTJ
PRO	PRON
E	SYM

DUTCH	
CGN tag	UD tag
N(soort	NOUN
N(eigen	PROPN
SPEC(deeleigen	PROPN
ADJ(	ADJ
BW(	ADV
WW(	VERB
LID(	DET
VNW(	PRON
VZ(	ADP
VG(neven	CCONJ
SPEC(enof	CCONJ
VG(onder	SCONJ
SPEC(symb	SYM
SPEC(afk	SYM
TW(	NUM
TSW(	INTJ
SPEC(vreemd	FW
LET(	PUNCT
SPEC(meta	X
SPEC(afgebr	X
FW	FW
ADV	ADV
PROPN	PROPN
INTJ	INTJ
ADP	ADP
E	SYM
SYM	SYM

Figure 5: Appendix B - Language-specific PoS tags and conversions to UD tags.