

LeTTuce PoS-Tagger

A Sprout of Innovation in Multilingual NLP

Cynthia Van Hee, Pranaydeep Singh and Els Lefever
LT3, Department of Translation, Interpreting and Communication
Ghent University, Belgium
cynthia.vanhee, pranaydeep.singh, els.lefever@ugent.be

1 Introduction

In Natural Language Processing (NLP), a vast array of tools have emerged to tackle diverse tasks. As the user base grows and applications expand, comparative analyses are crucial to help users select the best tool for a specific dataset or language.

This research benchmarks both traditional part-of-speech (PoS) tagging tools and state-of-the-art models for English, French, German and Dutch across various text genres and domains. Moreover, we present a novel multilingual tool, **LeTTuce**¹ and evaluate its PoS tagging performance across different data genres and domains, encompassing social media, industry reviews (engineering, human resources, hotel, airline) and technical texts. For Dutch, we conduct an additional evaluation on historical correspondence writing².

2 Related research

PoS tagging errors are a key source of dependency parsing inaccuracies (Foster et al., 2011) and remain essential for tasks like syntactic parsing, information extraction, machine translation, sentiment analysis and corpus linguistics. Although PoS tagging has achieved high accuracy in many languages, challenges persist—especially with noisy social media data, morphologically rich and under-represented languages, creative usage, code-switching and historical texts.

Early taggers used rule-based and statistical models (e.g., HMMs, Maximum Entropy, CRFs). Deep learning introduced RNNs, LSTMs and Transformers like BERT, which improved context modeling and handling of morphological complexity and unknown words (e.g., Pota et al., 2019; Yang et al., 2018). These models also enabled better transfer learning, aiding PoS tagging and parsing in noisy, low-resource, historical and ancient language settings (e.g., Kim et al., 2017; Meftah & Semmar, 2018; Smidt et al., 2024; Szawerna, 2024). Recent work explores large language models (LLMs) for PoS tagging, leveraging their contextual knowledge to improve robustness across language varieties and low-resource and historical languages (e.g., Fang et al., 2025; Subedi et al., 2024).

This paper introduces LeTTuce, a multilingual PoS tagger and benchmarks it against spaCy (Honnibal & Montani, 2017), Stanza (Qi et al., 2020) and LeTs Preprocess (Van de Kauter et al., 2013). We compare traditional sequence labeling methods and neural architectures across languages, genres and domains to assess tagging performance and highlight remaining challenges.

3 Methodology

3.1 Data collection and annotation

The training data consist of newswire texts in four languages, sourced 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 can be found in Van de Kauter et al. (2013). Test data were collected via web crawling publicly available

¹All novel PoS models will be integrated in the CLARIN infrastructure (<https://www.clarin.eu/>), making them readily accessible to the research community.

²Unfortunately, comparable historical data were not available for the other languages

sources and via data processing agreements with industry partners. While access to proprietary partner data is restricted, the public dataset is available to researchers upon request. As shown in Table 1, the test corpus spans multiple genres and domains.

The data annotation was carried out by three trained linguists using language-specific tag sets: the Penn Treebank, the French TreeTagger, the Stuttgart-Tübingen tag set and the Dutch CGN corpus.

Genre	English	French	German	Dutch
Train				
Newspaper	1,552,940	337,143	2,608,975	1,326,444
Test				
Reviews	14,033	5646	-	16,961
Social media	4527	2896	-	4570
Technical	17,418	19,081	8707	16,517
Historical	-	-	-	1582

Table 1: Token counts per language and domain in the LeTTuce train and test sets.

3.2 Experimental Setup

The training corpus was used to fine-tune two pretrained models per language—a language-specific BERT and a cross-lingual XLM encoder (Conneau & Lample, 2019)—by adding a linear layer for token-level classification³. For Dutch, both fine-grained (fg) and coarse-grained (cg) PoS tags were available. While we tested both, only coarse-grained results are reported (Table 2) to maintain comparability with the other languages and models.

Initial experiments covered five language variants (EN, FR, DE, NL-fg, NL-cg), each evaluated with BERT and XLM across four domains (reviews, social media, technical, historical). In addition, we assessed cross-domain generalization by testing BERT and XLM models on combined-domain data for each language, as shown in Figure 1.

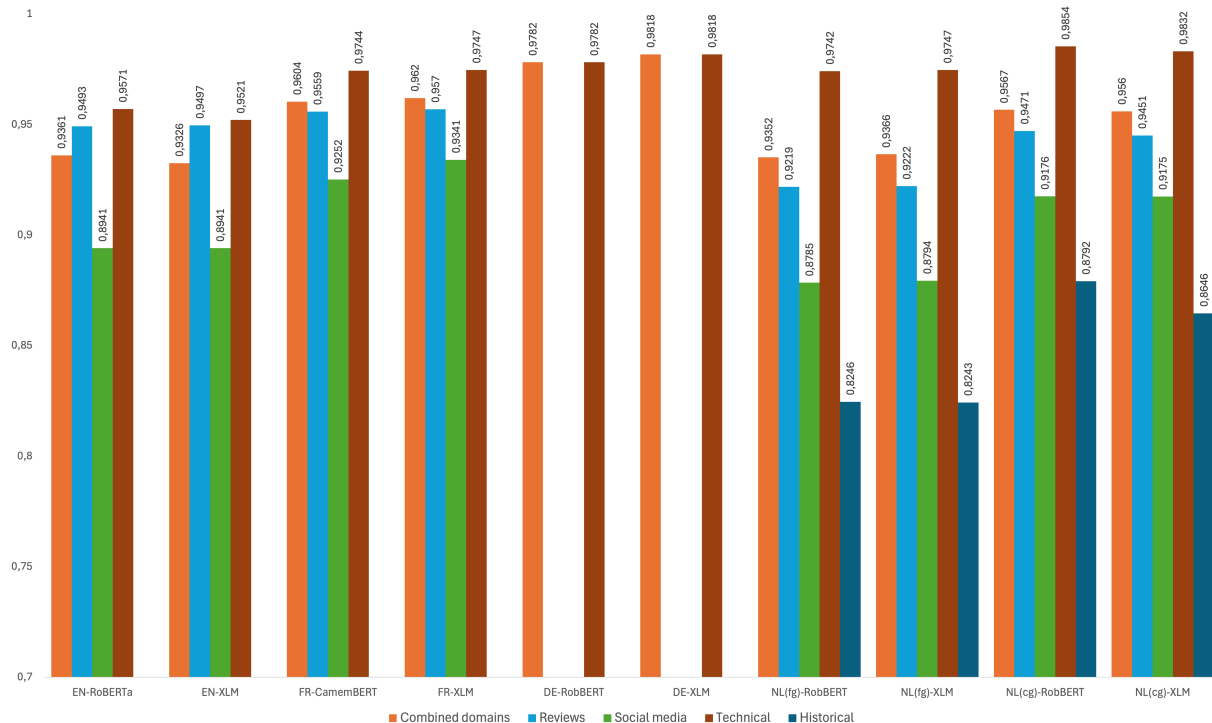


Figure 1: LeTTuce PoS tagging performance (F₁) obtained by various pretrained models across different data genres.

³The models are available via <https://github.com/lt3/Lettuce>.

4 Results

Figure 1 presents LeTTuce’s performance across four languages and domains, showing comparable results between the language-specific and XLM models, a trend that is consistent across all languages. Tagging performance is highest on technical texts, likely due to their consistent nature and standardized language. In contrast, performance on social media data is substantially lower, reflecting the challenges posed by non-standard orthography and informal language. The poorest results are observed on Dutch historical data, which can be attributed to spelling variation (e.g., “*litanien*”, “*z’er*”), capitalization (e.g., “*Eerwaarde*”, “*Woensdage*”), code switching and loanwords (e.g., “*jubilé*”, “*choses*”).

Language	PoS tagger	Reviews	Social media	Technical	Historical
English	spaCy	0.9296	0.8695	0.9343	-
	Stanza	0.9373	0.8763	0.9579	-
	LeTs Preprocess	0.9621	0.9093	0.9435	-
	LeTTuce	0.9497	0.8941	0.9521	-
French	spaCy	0.8489	0.8129	0.8712	-
	Stanza	0.9197	0.9009	0.9200	-
	LeTs Preprocess	0.9342	0.9048	0.9741	-
	LeTTuce	0.9570	0.9341	0.9747	-
German	spaCy	-	-	0.9559	-
	Stanza	-	-	0.9516	-
	LeTs Preprocess	-	-	0.9715	-
	LeTTuce	-	-	0.9818	-
Dutch	spaCy	0.9057	0.8467	0.9230	0.7799
	Stanza	0.9325	0.8873	0.9729	0.8374
	LeTs Preprocess	0.9453	0.8959	0.9726	0.9026
	LeTTuce	0.9451	0.9175	0.9832	0.8646

Table 2: Performance (F_1) comparison between state-of-the-art PoS taggers across languages and genres. High scores per domain are indicated in bold.

To facilitate interpretation of LeTTuce’s performance, Table 2 compares the part-of-speech tagging accuracy of four toolkits: spaCy, Stanza, LeTs Preprocess and LeTTuce (which leverages cross-lingual XLM encoders), evaluated on the same datasets to ensure fair benchmarking. This comparison not only highlights differences in tagging performance across toolkits, but also offers practical guidance for NLP practitioners in selecting the most suitable tool for their target language and domain.

Figure 2 illustrates consistent performance differences across all four taggers depending on the data genre. For English, French and Dutch, PoS tagging performance is best on technical texts, followed by review data, with noticeable drops in accuracy for social media and Dutch historical texts.

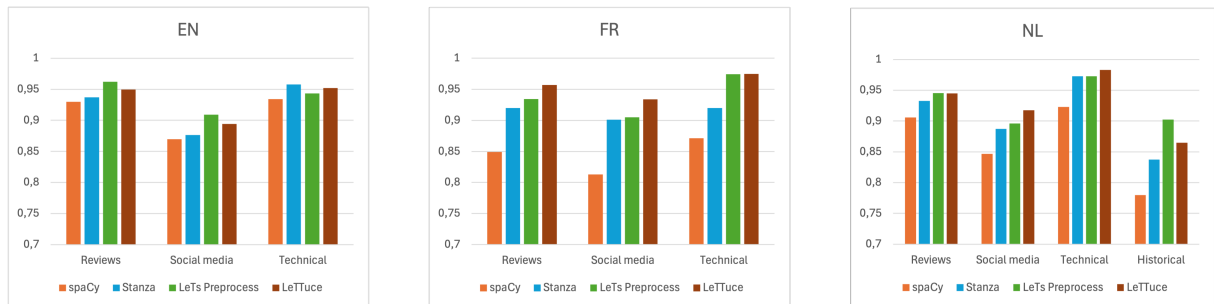


Figure 2: Tagging performance across domains for three languages and four PoS tagging toolkits.

Future work will extend this evaluation to include additional languages and text genres, aiming to provide a more comprehensive benchmark. Additionally, we plan to integrate lemmatization in order to develop a complete LeTTuce preprocessing toolkit.

References

- Brants, S., Dipper, S., Hansen, S., Lezius, W., & Smith, G. (2002). The tiger treebank. *Proceedings of the workshop on treebanks and linguistic theories*, 24–41. https://www.linguistics.ruhr-uni-bochum.de/~dipper/pub/tlt02_webVersion.pdf
- Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In *Proceedings of the 33rd international conference on neural information processing systems*. Curran Associates Inc.
- Fang, Z., Wu, L.-C., Kong, X., & Stewart, S. D. (2025, May). A comparative analysis of word segmentation, part-of-speech tagging, and named entity recognition for historical Chinese sources, 1900–1950. In M. Härmäläinen, E. Öhman, Y. Bizzoni, S. Miyagawa, & K. Alnajjar (Eds.), *Proceedings of the 5th international conference on natural language processing for digital humanities* (pp. 1–6). Association for Computational Linguistics. <https://aclanthology.org/2025.nlp4dh-1.1/>
- Foster, J., Çetinoglu, Ö., Wagner, J., Roux, J. L., Hogan, S., Nivre, J., Hogan, D., & van Genabith, J. (2011). #hardtoparse: POS tagging and parsing the twitterverse. *Analyzing Microtext, Papers from the 2011 AAAI Workshop, San Francisco, California, USA, August 8, 2011, WS-11-05*. <http://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/view/3912>
- Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing* [To appear].
- Kim, J.-K., Kim, Y.-B., Sarikaya, R., & Fosler-Lussier, E. (2017, September). Cross-lingual transfer learning for POS tagging without cross-lingual resources. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2832–2838). ACL. <https://doi.org/10.18653/v1/D17-1302>
- Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank (J. Hirschberg, Ed.). *Computational Linguistics*, 19(2), 313–330. <https://aclanthology.org/J93-2004>
- Meftah, S., & Semmar, N. (2018). A neural network model for part-of-speech tagging of social media texts. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the 11th international conference on language resources and evaluation*. European Language Resources Association (ELRA). <https://aclanthology.org/L18-1446>
- Paulussen, H., Macken, L., Vandeweghe, W., & Desmet, P. (2013). Dutch parallel corpus: A balanced parallel corpus for dutch-english and dutch-french. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for dutch : Results by the stevin-programme* (pp. 185–199). Springer. http://doi.org/10.1007/978-3-642-30910-6_11
- Pota, M., Marulli, F., Esposito, M., De Pietro, G., & Fujita, H. (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. <https://www.sciencedirect.com/science/article/pii/S0950705118305392>
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the ACL: System Demonstrations*. <https://nlp.stanford.edu/pubs/qi2020stanza.pdf>
- Smidt, G. R., Lefever, E., & de Graef, K. (2024, May). At the crossroad of cuneiform and NLP: Challenges for fine-grained part-of-speech tagging. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 1745–1755). ELRA; ICCL. <https://aclanthology.org/2024.lrec-main.154/>
- Subedi, B., Regmi, S., Bal, B. K., & Acharya, P. (2024, May). 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 N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 6974–6979). ELRA; ICCL. <https://aclanthology.org/2024.lrec-main.611/>

- Szawerna, M. I. (2024, March). Can stanza be used for part-of-speech tagging historical Polish? In N. Falk, S. Papi, & M. Zhang (Eds.), *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: Student research workshop* (pp. 44–49). Association for Computational Linguistics. <https://aclanthology.org/2024.eacl-srw.4/>
- van Noord, G., Bouma, G., Van Eynde, F., de Kok, D., van der Linde, J., Schuurman, I., Sang, E. T. K., & Vandeghinste, V. (2013). Large scale syntactic annotation of written dutch: Lassy. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for dutch: Results by the stevin programme* (pp. 147–164). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30910-6_9
- Van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L., & Hoste, V. (2013). LeTs pre-process: the multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal*, 3, 103–120.
- Yang, L., Zhang, M., Liu, Y., Sun, M., Yu, N., & Fu, G. (2018). Joint POS Tagging and Dependence Parsing With Transition-Based Neural Networks. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 26(8), 1352–1358. <https://doi.org/10.1109/TASLP.2017.2788181>