Can Stanza be Used for Part-of-Speech Tagging Historical Polish?

Maria Irena Szawerna
Språkbanken Text
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg, Sweden
maria.szawerna@gu.se

Abstract

The goal of this paper is to evaluate the performance of Stanza, a part-of-speech (POS) tagger developed for modern Polish, on historical text to assess its possible use for automating the annotation of other historical texts. While the issue of the reliability of utilizing POS taggers on historical data has been previously discussed, most of the research focuses on languages whose grammar differs from Polish, meaning that their results need not be fully applicable in this case. The evaluation of Stanza is conducted on two sets of 10286 and 3270 manually annotated tokens from a piece of historical Polish writing (1899), and the errors are analyzed qualitatively and quantitatively. The results show a good performance of the tagger, especially when it comes to Universal Part-of-Speech (UPOS) tags, which is promising for utilizing the tagger for automatic annotation in larger projects, and pinpoint some common features of misclassified tokens.

1 Introduction and Background

Annotated data for historical or otherwise non-standard variants of language can be difficult or resource-consuming to obtain but is nevertheless necessary for certain linguistic inquiries. One of the possible methods of alleviating this issue is attempting to use tools developed for a contemporary standard language for automated annotation. However, the data in question differing from the standard may pose problems. Consider the example presented in Table 1, a sentence from a 19th-century Polish memoir: the differences between the original and the modern version of the same sentence pertain not only to spelling but also word order and vocabulary – but the extent to which these seemingly large differences affect the performance of modern tools is not clear. This paper aims to address this question and estimate what kinds of variation have the largest negative impact on tagging accuracy.

Table 1: A sentence from Juliusz Czerminski's memoir (Szawerna, 2023) in the original, with modernized spelling, modernized language, and in English.

Original sentence	Odjechał do Lwowa – nazajutrż miał wrucić i wrucił, ale w trumnie.
Modern spelling	Odjechał do Lwowa – nazajutrz miał wrócić i wrócił, ale w trumnie.
Modern language	Pojechał do Lwowa – miał wrócić dzień później, i wrócił, ale w trumnie.
English translation	He drove away to Lviv – he was supposed to return the day after and that he did, but in a coffin.

A considerable amount of research has already been conducted on the evaluation of various pre-trained part-of-speech (POS) taggers on historical texts to establish their effectiveness at annotating such texts. POS taggers trained on contemporary data tend to struggle with historical texts for a variety of reasons, such as out-of-vocabulary items, variation in spelling, capitalization, and punctuation, as well as differences in morphology and syntax and semantic shifts, but large performance improvements can be observed when relatively simple pre-processing methods such as spelling correction, spelling simplification, punctuation removal or normalization are used (Rayson et al., 2007; Scheible et al., 2011; Adesam and Bouma, 2016; Hupkes and Bod, 2016). A summary of the performance of various POS taggers when tested on historical data from various studies can be seen in Table 2. While taggers based on neural networks (NNs) have been shown to outperform other methods, much of the research predates those and is based on older architectures (Yang and Eisenstein, 2016; Adesam and Berdicevskis, 2021).

While most of the previously mentioned studies focus on languages from the Germanic family, this paper aims to evaluate a POS-tagger for modern Polish on historical texts. Given the differences be-

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 44–49

Table 2: Test results on raw and preprocessed data in other experiments (some results are for more than one tagger or data from various periods).

Paper	Language	Modern Test Set Accuracy	Historical Test Data Measures	Preprocessed[1] Test Data Accuracy
Rayson et al. (2007)	English	96%	Accuracy: 82–88.5%	89–93.2%
Scheible et al. (2011)	German	-	Accuracy: 69.6%	79.7%
Bollmann (2013)	German	-	Accuracy: 23–81.8%	83.4–95.6%
Hupkes and Bod (2016)	Dutch	96%	Accuracy: 60%	92%
Adesam and Bouma (2016)	Swedish	94.2%[2]	Accuracy: 45%	70%
Waszczuk et al. (2018)	Polish	-	Precision: 88.3–90.3% Recall: 88.3–90.3%	-
Szawerna (2023)	Polish	89.3–99.2%	Accuracy: 80.2–94.5%	-

tween Germanic and Slavic languages, other kinds of errors can appear in the tagger annotation. Moreover, the research mentioned in Table 2 was conducted on texts from not only various languages but also various periods. Waszczuk et al. (2018) evaluated the performance of a tagger on historical Polish data and reported quite high performance on texts from the 17th-20th-century, which is promising. However, the tool that they are reporting on, Morfeusz2, is a CRF-based tagger, which could mean that an NN-based tool could potentially perform even better. While the research presented by Szawerna (2023) includes various performance measures for several tools, the focus of that research was on identifying variation and not utilizing the tools for automated annotation; importantly, though, Szawerna (2023) does present a comparison of the performance of various tools, with Stanza performing better on historical data than Morfeusz2 which utilizes a combination of rule-based morphological analysis and CRF (conditional random fields) for tagging; Morfeusz2 did, however, outperform Stanza on modern texts (Kieraś and Woliński, 2017). While a fine-tuned BERT model did outperform Stanza, the latter is more of an out-of-the-box tool and is therefore more likely to be used in a pipeline, warranting the analysis of its performance on nonstandard data.

This paper builds upon the research presented in Szawerna (2023) and investigates the performance of a single tagger on a memoir from 1899 which also contains dialectical variation. Given the age of the data, the accuracy is expected to be around 90% accuracy[3], with Universal Part-of-Speech (UPOS) tagging performing better than tagging using language-specific (XPOS) tags. The tagger is expected to struggle with nonstandard spelling or capitalization, out-of-vocabulary items, and other previously mentioned issues.

2 Materials and Methods

The tagger used in this project is that provided by Stanza, a Natural Language Processing (NLP) toolkit featuring models for a large number of languages (Qi et al., 2020). The default model for Polish was trained and evaluated on the Polish Dependency Bank treebank (Wróblewska, 2018; Stanza, n.d.). It is also that corpus's test set that is used to exemplify the tool's performance on modern Polish in this paper, although it represents genres different from the historical texts. The main reasons for selecting this tagger are its ease of use and high reported accuracy on modern data.

The data used for testing the tagger in this project comes from the memoir of Juliusz Czerniński, who lived in the 19th century in the area corresponding to nowadays Eastern Poland and Western Ukraine. The original manuscript was composed in 1889, retyped on a typewriter, and recently digitized. No intentional alterations were made to e.g. seemingly misspelled tokens. This data was first presented by Szawerna (2023), where its divergence from modern Polish was asserted, especially when it comes to features typical for the dialects of that region (Kurzowa, 1983). According to Polański (2004), there was no singular universally accepted spelling convention around the time of the memoir's creation. Therefore, the text should

[1]The preprocessing methods varied between the experiments but often consisted of standardizing the spelling and punctuation.

[2]Here the tagger was trained on historical texts as well.

[3]Unfortunately Waszczuk et al. (2018) do not report accuracy as a measure.

not be considered to be representative of historical Polish of its time, both due to its dialectical features and spelling which is not representative of the bulk of the contemporaneous writing.

In its entirety, the data consists of 37,405 tokens. Out of those, the first 10286 tokens were manually annotated using Universal Dependencies' universal POS tags (UPOS tags). A subset of 3270 tokens was further annotated using XPOS tags. Both of these tagsets are utilized by Stanza. The only changes to the original text include the splitting of the "mobile inflection" as per the UD guidelines and removing any punctuation from inside numbers (Szawerna, 2023; Universal Dependencies, n.d.). This previously conducted manual annotation of the tokens has been reviewed, and a few corrections have been made.

Evaluation measures were calculated for both kinds of annotation. The results were also subjected to a qualitative analysis, the goal of which was to determine what kinds of errors are the most prevalent, which could give insights into what kinds of potential pre-processing could eliminate that problem. The misclassified examples were saved and manually annotated for the error type before being processed to obtain the relevant statistics.

3 Results

Stanza exhibits very good performance on modern Polish data and relatively good performance on historical data. Table 3 shows the accuracy achieved by the model on the respective datasets and tagsets.

Table 3: Stanza's accuracy per text type and tagset.

	Modern	Historical
UPOS	98.79%	94.15%
XPOS	94.76%	88.05%

A more detailed evaluation was obtained for the UPOS tagset. Figure 1 and Figure 2 visualize the per-class performance of the model for each dataset, with the counts for each class being normalized by the true positive count for that class (therefore, the values on the diagonal correspond to recall). It is worth pointing out that tags like *INTJ* and *SYM* were absent from the historical data altogether. What can be noted is that with the exception of many *SYM* and *INTJ* classes, the tagger shows more consistent performance on modern data than on historical. While for categories such as *ADJ*, *ADV*, *AUX*, *DET*, *NUM*, *SCONJ*, and X the results

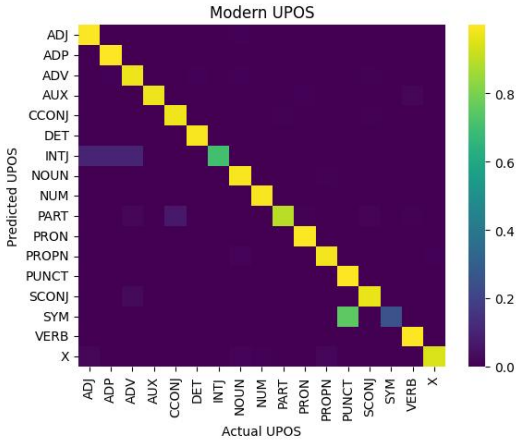

Figure 1: Normalized confusion matrix for UPOS tagging of the modern data.

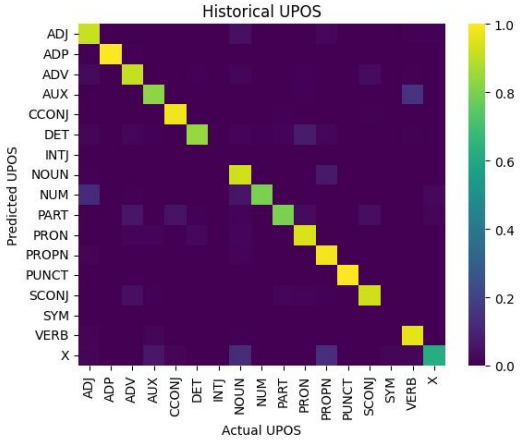

Figure 2: Normalized confusion matrix for UPOS tagging of the historical data.

on historical data are visibly lower, the overall performance on historical data is still rather good. The XPOS tagset is much larger, in the order of hundreds of tags, making a similar visual comparison uninformative, and a more detailed analysis is beyond the scope of this paper.

Another method of inspecting the tagger's performance is investigating the erroneously labeled tokens. Table 4 and Table 5 illustrate the frequency of specific kinds of errors present among the mistakes made by Stanza in the memoir, following the general annotation utilized by Szawerna (2023). While the exact proportions differ between the two tagsets, *spelling*, *ambiguous*, and *unidentified* type errors are the most common for both. Noticeably, UPOS tagging fails when it comes to tokens with unusual spelling, including capitalization, which seems to be relevant for identifying *PROPN* and the replacement of the *y* (/ɨ/) vowel with *e*, and

Error Type	Raw freq.	Relative freq.
spelling	293	48.67%
ambiguous	223	37.04%
unidentified	37	6.15%
vocabulary	35	5.81%
name	7	1.16%
abbreviation	4	0.66%
grammar	3	0.50%

Table 4: Frequency of errors by type for UPOS tagging.

Error Type	Raw freq.	Relative freq.
ambiguous	184	47.06%
unidentified	77	19.69%
spelling	51	13.04%
name	49	12.53%
vocabulary	22	5.63%
grammar	4	1.02%
abbreviation	4	1.02%

Table 5: Frequency of errors by type for XPOS tagging.

spelling the /j/ sound with *y*, which distort various inflectional endings. XPOS tagging struggles more with ambiguity (e.g. when more than one grammatical case uses the same ending), although the spelling variation not related to capitalization still has a non-negligible effect. One relevant type of ambiguous errors, present in both types of tagging, is that related to the sometimes questionable status of verb-derived nouns and adjectives. For example, the word *bombardowanie* 'bombing' is considered an established noun, but the tagger classifies it as a gerund (WSJP Editorial Team, 2014; nkj, n.d.), likely because of the form. Interestingly enough, among the annotated XPOS errors there are also several examples of the vocative case being ignored or the model defaulting to assigning the masculine grammatical gender to a pronoun despite the context implying that it should be feminine. There are also instances of verbs in the impersonal past form that are consistently misclassified.

4 Discussion

The results of the quantitative evaluation show a good performance of the tagger, exceeding most of the previously reported ones, including the results reported for the same data and tagger by Szawerna (2023),[4] possibly due to improvements that have been made to Stanza's model. On the other hand, Waszczuk et al. (2018) still achieve a better performance on XPOS tags using a CRF-based model. However, they use a more diverse and larger dataset which may consist of more standard Polish than the data investigated in this paper. Nevertheless, Stanza's performance on this test data is only around 4 (UPOS) and 7 (XPOS) percentage points below the accuracy it has shown on its own test set. Interestingly enough, the performance on the PDB test set is slightly higher than reported by Stanza (n.d.), possibly due to the corpus being pre-tokenized before being fed to the model.

A qualitative error analysis has approximated what the tagger struggles with when it comes to the test data. Previous studies have shown that variations in spelling, capitalization, punctuation, differences in morphology and syntax, and semantic shifts are some of the factors that make accurate tagging of historical texts using modern taggers difficult (Rayson et al., 2007; Scheible et al., 2011; Adesam and Bouma, 2016; Hupkes and Bod, 2016). In the case of Stanza, some of those issues, such as nonstandard capitalization, archaic vocabulary, and spelling have negatively impacted the tagger's performance. This is particularly prominent as far as UPOS tagging is concerned. As far as XPOS-tagging goes, issues pertaining to the inflectional morphology have been highlighted, such as confusing word endings or problems with words the class of which is ambiguous. Additionally, issues such as the possible underrepresentation of rarer classes in the training corpus could be noted, leading to biases concerning feminine pronouns and issues identifying the vocative case.

5 Conclusions and Future Work

Within this paper, a modern Polish POS tagger, Stanza, has been evaluated on historical and modern data, and some of the issues causing the drop in its performance on historical texts have been successfully identified. It has been shown that it can perform quite well on non-standard, historical Polish data from the late 19th century, and this can possibly be improved using some preprocessing methods, making it a promising candidate for at least assisting the annotation of historical texts, if not completely automating it. Many of the misclassified tokens were problematic due to issues previously identified in the literature in the field; however, some problems seemed to stem from the inflectionality of the language or be inherent to

[4] Other taggers used in that research achieve even higher scores.

the tagger itself. Potential biases stemming from the under-representation of certain classes in the training data for the tagger have also been shown.

In the future, it would be interesting to test the influence of various factors, such as e.g. punctuation or lowercasing, on the quality of tagging. Another possibility could be comparing the performance of multiple different taggers or tagging architectures on the same data, or testing the same tagger on data from different periods. Alternatively, one could juxtapose the results presented in this paper to those from tagging a very recent, nonstandard text, e.g. sourced from the web, to see to what extent the same issues are causing tagging problems. Finally, developing some methods for the pre-processing of texts from this period for subsequent tagging could also be quite useful. It would also be interesting to compare how the models for other languages included in Stanza perform on samples of historical texts from their respective languages.

As far as the data itself is concerned, it would be interesting to complete and review the annotation of the entire memoir, and see how the results of an analysis such as the one presented in this paper would change; this would also open up the opportunity for different kinds of research on the text.

Ethics Statement

Given the age of the data, its use does not pose an ethical challenge. The analysis of mistakes made by Stanza indicates some possible existing biases when it comes to assigning gender-marked XPOS tags to words the gender of which is ambiguous when the context is not taken into account. Simultaneously, it is worth pointing out that the re-using of existing tools should be encouraged, especially when it comes to resource-heavy tools (such as NN-based ones), as it potentially limits the negative environmental impact of training large models.

Limitations

It is also important to acknowledge the limitations of this research. Performing this kind of analysis on data from only one author and a relatively recent period does not fully address the question of whether it is possible to utilize the Stanza tool for POS tagging any Polish text, nor does the paper provide an answer as to what kind of performance would make a tagger sufficiently good for use in preprocessing pipelines for historical texts. In the latter case, the author is of the opinion that this decision should be made on a case-by-case basis, and depending on the downstream tasks. The data having been annotated by only one person makes it somewhat prone to errors. Moreover, no attempt at assessing the effect of pre-processing (spelling and punctuation normalization) has been presented, rendering a full comparison with some of the prior research impossible.

References

n.d. Polish NKJP part of speech tagset. Accessed: 08.11.2023.

Yvonne Adesam and Aleksandrs Berdicevskis. 2021. Part-of-speech tagging of Swedish texts in the neural era. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 200–209, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Yvonne Adesam and Gerlof Bouma. 2016. Old Swedish part-of-speech tagging between variation and external knowledge. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 32–42, Berlin, Germany. Association for Computational Linguistics.

Marcel Bollmann. 2013. POS tagging for historical texts with sparse training data. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 11–18, Sofia, Bulgaria. Association for Computational Linguistics.

Dieuwke Hupkes and Rens Bod. 2016. POS-tagging of Historical Dutch. In *LREC 2016: Tenth International Conference on Language Resources and Evaluation*, pages 77–82, Paris. European Language Resources Association (ELRA).

Witold Kieraś and Marcin Woliński. 2017. Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego. *Język Polski*.

Zofia Kurzowa. 1983. *Polszczyzna Lwowa i Kresów Południowo-Wschodnich do 1939 roku*. Państwowe Wydawnictwo Naukowe.

Edward Polański. 2004. Reformy ortografii polskiej - wczoraj, dziś i jutro. *Biuletyn Polskiego Towarzystwa Językoznawczego, Z. 60*, pages 29–46.

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

Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. 2007. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. Evaluating an 'off-the-shelf' POS-tagger on early Modern German text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 19–23, Portland, OR, USA. Association for Computational Linguistics.

Stanza. n.d. Model performance. Accessed: 06.11.2023.

Maria Irena Szawerna. 2023. *IŻ SWÓJ JĘZYK MAJĄ! An exploration of the computational methods for identifying language variation in Polish*. Master's thesis, University of Gothenburg.

Universal Dependencies. n.d. UD for Polish. Accessed: 08.11.2023.

Jakub Waszczuk, Witold Kieraś, and Marcin Woliński. 2018. Morphosyntactic disambiguation and segmentation for historical Polish with graph-based conditional random fields. In *International Conference on Text, Speech, and Dialogue*, pages 188–196. Springer.

Alina Wróblewska. 2018. Extended and enhanced Polish dependency bank in Universal Dependencies format. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 173–182, Brussels, Belgium. Association for Computational Linguistics.

WSJP Editorial Team. 2014. Bombardowanie. Accessed: 09.11.2023.

Yi Yang and Jacob Eisenstein. 2016. Part-of-Speech Tagging for Historical English. pages 1318–1328.

A Appendix

- GitHub repository (code and data)