Evaluating the Stanza NLP toolkit's performance on historical Polish

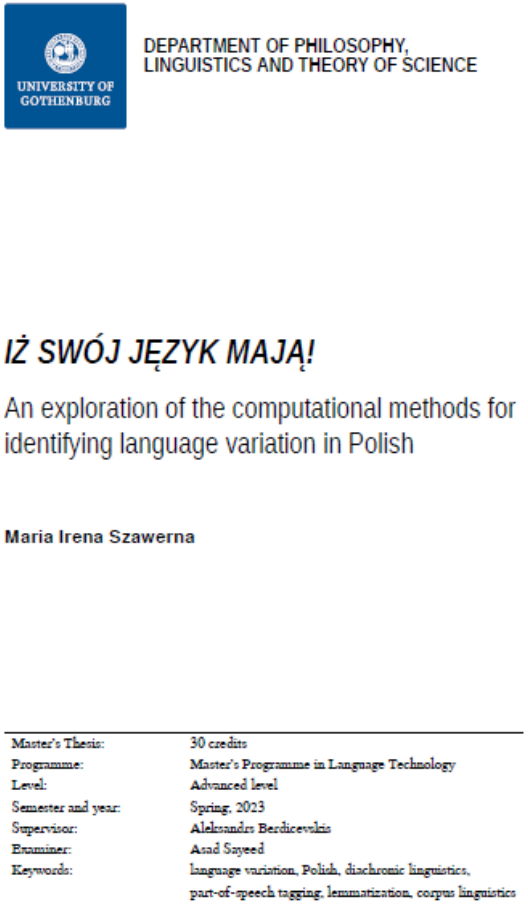
Maria Irena Szawerna

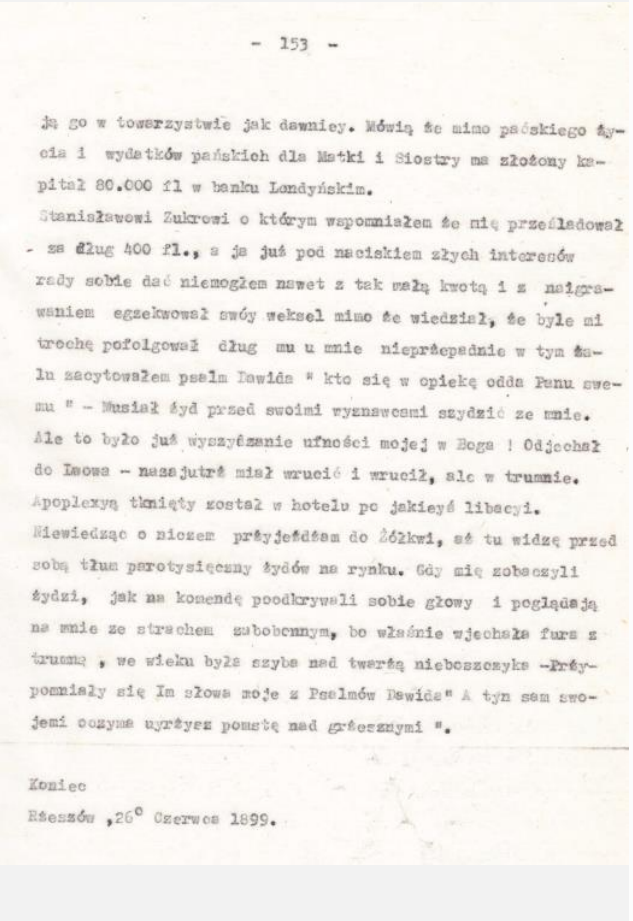
Roadmap

- Research Context
  - Original project
  - Related work
- Data
  - Example
- Research Question
- Experiment
- Results
- Ethical Considerations
- Future Work
- Conclusions

Related Research

- MA thesis project at the University of Gothenburg
- History of Polish
- Quantitative and corpus research in historical linguistics
- Part-of-speech tagging of historical data
- Methods for dealing with language variation in NLP
- Detecting and modelling language variation and change





- 153 -

ją go w towarzystwie jak dawniey. Mówią że mimo pańskiego życia i wydatków pańskich dla Matki i Siostry ma złożony kapitał 80.000 fl w banku Londyńskim.
Stanisławowi Zukrowi o którym wspomniałem że nią prześladował - za dług 400 fl., a ja już pod naciskiem złych interesów rady sobie dać niemogłem nawet z tak małą kwotą i z naigrawaniem egzekwował swóy weksel mimo że wiedział, że byle mi trochę pofolgował dług mu u mnie nieprzepadnie w tym żalu zacytowałem psalm Dawida " kto się w opiekę odda Panu swemu " - Musiał żyd przed swoimi wyznawcami szydzić ze mnie. Ale to było już wyszydzanie ufności mojej w Boga ! Odjechał do Lwowa - nazajutrz miał wrucić i wrucił, ale w trumnie. Apoplexyą tknięty został w hotelu po jakieyś libacyi. Niewiedząc o niczem przyjeżdżam do Żółkwi, aż tu widzę przed sobą tłum parotysięczny żydów na rynku. Gdy mię zobaczyli żydzi, jak na komendę poodkrywali sobie głowy i poglądają na mnie ze strachem zabobonnym, bo właśnie wjechała furs z trumną , we wieku byłs szyba nad twerią nieboszczyka -Przypomniały się Im słowa moje z Psalmów Dawida" A tyn sam swojemi oczyma wyrżysz pomstę nad grześznymi ".

Koniec
Rzeszów ,26° Czerwca 1899.

Data

- 1899 memoir.
- Copied over from a manuscript.
- Visible variation in e.g. spelling, still intelligible for a native speaker.
- Manual UD-style annotation (with pre-annotation).
  - Total: 37 405 tokens.
  - UPOS-annotated: 10 286 tokens.
  - XPOS-annotated, lemmatized: 3271 tokens.

Data – example

Original:

Odjechał do Lwowa – nazajutrż miał wrucić i wrucił, ale w trumnie.
Apoplexyą tknięty został w hotelu po jakieyś libacyi.

Modernized spelling:

Odjechał do Lwowa – nazajutrz miał wrócić i wrócił, ale w trumnie.
Apopleksją tknięty został w hotelu po jakiejś libacji.

Heavily modernized language:

Pojechał do Lwowa – miał wrócić dzień później, i wrócił, ale w trumnie.
Dostał udaru w hotelu po jakiejś imprezie.

English:

He drove away to Lviv – and he was supposed to return the day after and that he did, but in a coffin.
He had suffered a stroke at a hotel after some party.

Research Question

How well does the Stanza NLP toolkit perform on a sample of 19th-century Polish and what errors does it tend to make?

Experiments

- XPOS- and UPOS-tagging, lemmatization
- Error annotation
- Tools and resources:
  - Stanza NLP toolkit
  - Jupyter Notebook
  - PDB-UD

Results: lemmatization

	Accuracy (original)	Accuracy (lowercase)
PDB-UD	90.89%	92.34%
Historical	83.58%	86.55%

error	raw	relative
unidentified	215	40.04%
stanza	94	17.50%
spelling	94	17.50%
name	61	11.36%
ambiguous	36	6.70%
vocabulary	20	3.72%
grammar	9	1.68%
abbreviation	8	1.49%

error	raw	relative
unidentified	212	48.18%
spelling	96	21.82%
name	60	13.64%
ambiguous	35	7.95%
vocabulary	20	4.55%
grammar	9	2.05%
abbreviation	8	1.82%

Results: UPOS-tagging

	Accuracy
PDB-UD	98.40%
Historical	93.31%

error	raw	relative
spelling	301	43.75%
ambiguous	244	35.47%
name	55	7.99%
unknown	52	7.56%
vocabulary	29	4.22%
abbreviation	6	0.87%
grammar	1	0.15%

Results: XPOS-tagging

	Accuracy
PDB-UD	94.29%
Historical	87.71%

error	raw	relative
ambiguous	196	48.76%
spelling	61	15.17%
name	55	13.68%
unknown	54	13.43%
vocabulary	20	4.98%
abbreviation	5	1.24%
annotation	4	1.00%
numeral	4	1.00%
grammar	3	0.75%

Results: trends in errors

- Spelling: *y* (*suchey* instead of *suchej*)
- Spelling: *nie* (*niemają* instead of *nie mają*)
- Spelling/pronunciation: *e* (*małem* instead of *małym*)
- Spelling/pronunciation: *rż* (*warżenia* instead of *warzenia*)
- Spelling: capitalization (*Dziedzica* instead of *dziedzica*)

Results: trends in errors

- Grammar: nonstandard inflection (*człowiecze* instead of *człowieku*)
- Grammar: vocative vs. nominative (*Asińdźka* instead of *Asińdźko*)
- Grammar: impersonal verb forms
- Vocabulary: proper names
- Vocabulary: other OOV items
- Ambiguity: numerals
- Ambiguity: verb-derived nouns and adjectives
- Miscellaneous errors.

Ethical considerations

- Old data.
- Not expecially computationally heavy.
- Explores utilizing tools for underrepresented dialects or languages.
- Gender annotation on pronouns – gender bias?
  - `# sent_id = train-s2896`
  - `# text = - Ty nie wiesz?` (ENG: Do you not know?)
  - `# orig_file_sentence = 200-2-000093_morph_5.47-s#7092`
  - `...`
  - `2  Ty  ty  PRON  ppron12:sg:nom:m1:sec  ...`
  - `...`

Future work

- Comparison to more data
  - More data from the same time and region
  - Older data
  - Contemporary non-standard data
- Research on pre-processing methods

Conclusions: back to Research Question

- How well does the Stanza NLP toolkit perform on a sample of 19th-century Polish and what errors does it tend to make?
  - Significantly worse performance
  - Errors related to dialectical and diachronic variation
  - Miscellaneous errors
- Stanza is not fully reliable as an annotation tool for nonstandard data, but can be used for preannotation

Thesis and conference repository

- Thesis and code available at: https://github.com/Turtilla/swe-ma-thesis
- Presentation and code available at: https://github.com/Turtilla/WSMF-presentation

Thank you for your attention!

Bibliography

- Adesam, Y. & Bouma, G. (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* (pp. 32–42). Berlin, Germany: Association for Computational Linguistics.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (pp. 610–623). New York, NY, USA: Association for Computing Machinery.
- Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media. Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Online: Association for Computational Linguistics.
- Bollmann, M. (2013). POS tagging for historical texts with sparse training data. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (pp. 11–18). Sofia, Bulgaria: Association for Computational Linguistics.
- Dipper, S. & Waldenberger, S. (2017). Investigating diatopic variation in a historical corpus. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)* (pp. 36–45). Valencia, Spain: Association for Computational Linguistics.
- Donoso, G. & Sánchez, D. (2017). Dialectometric analysis of language variation in Twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)* (pp. 16–25). Valencia, Spain: Association for Computational Linguistics.
- Dorn, R. (2019). Dialect-specific models for automatic speech recognition of African American Vernacular English. In *Proceedings of the Student Research Workshop Associated with RANLP 2019* (pp. 16–20). Varna, Bulgaria: INCOMA Ltd.
- Dunaj, B. (2019). "Historia języka polskiego" Zenona Klemensiewicza a potrzeba nowej syntezy. *LingVaria*, 14.
- Długosz-Kurczabowa, K. & Dubisz, S. (2006). *Gramatyka historyczna Języka Polskiego*. Wydawnictwa Uniwersytetu Warszawskiego.
- Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19.
- Estarrona, A., Etxeberria, I., Etxepare, R., Padilla-Moyano, M., & Soraluze, A. (2020). Dealing with dialectal variation in the construction of the Basque historical corpus. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects* (pp. 79–89). Barcelona, Spain Online): International Committee on Computational Linguistics (ICCL).

- Garcia, M. & García Salido, M. (2019). A method to automatically identify diachronic variation in collocations. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change* (pp. 71–80). Florence, Italy: Association for Computational Linguistics.
- Garimella, A., Amarnath, A., Kumar, K., Yalla, A. P., N, A., Chhaya, N., & Srinivasan, B. V. (2021). He is very intelligent, she is very beautiful? On Mitigating Social Biases in Language Modelling and Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 4534–4545). Online: Association for Computational Linguistics.
- Garrette, D. & Alpert-Abrams, H. (2016). An unsupervised model of orthographic variation for historical document transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 467–472). San Diego, California: Association for Computational Linguistics.
- Gruszczyński, W., Adamiec, D., Bronikowska, R., & Wieczorek, A. (2020). ELEKTRONICZNY KORPUS TEKSTÓW POLSKICH Z XVII I XVIII W. – PROBLEMY TEORETYCZNE I WARSZTATOWE. (pp. 32–51).
- Hämäläinen, M., Partanen, N., & Alnajjar, K. (2021). Lemmatization of historical old literary Finnish texts in modern orthography. In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale* (pp. 189–198). Lille, France: ATALA.
- Hovy, D. (2018). The social and the neural network: How to make natural language processing about people again. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media* (pp. 42–49). New Orleans, Louisiana, USA: Association for Computational Linguistics.
- Hovy, D. & Purschke, C. (2018). Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4383–4394). Brussels, Belgium: Association for Computational Linguistics.
- Hovy, D. & Spruit, S. L. (2016). The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 591–598). Berlin, Germany: Association for Computational Linguistics.
- Hupkes, D. & Bod, R. (2016). POS-tagging of Historical Dutch. In *LREC 2016: Tenth International Conference on Language Resources and Evaluation* (pp. 77–82). Paris: European Language Resources Association (ELRA).

- Jenset, G. B. & McGillivray, B. (2017). *Quantitative Historical Linguistics: A Corpus Framework*. Oxford University Press.
- Johannessen, J., Kåsen, A., Hagen, K., Nøklestad, A., & Priestley, J. (2020). Comparing methods for measuring dialect similarity in Norwegian. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 5343–5350). Marseille, France: European Language Resources Association.
- Johannsen, A., Hovy, D., & Søgaard, A. (2015). Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning* (pp. 103–112). Beijing, China: Association for Computational Linguistics.
- Jurgens, D., Tsvetkov, Y., & Jurafsky, D. (2017). Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 51–57). Vancouver, Canada: Association for Computational Linguistics.
- Jurman, G., Riccadonna, S., & Furlanello, C. (2012). A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. *PLOS ONE*, 7(8), 1–8.
- Klemensiewicz, Z. (1976). *Historia Języka Polskiego*. Państwowe Wydawnictwo Naukowe.
- Kurzowa, Z. (1983). *Polszczyzna Lwowa i Kresów Południowo-Wschodnich do 1939 roku*. Państwowe Wydawnictwo Naukowe.
- Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Baltimore, Maryland: Association for Computational Linguistics.
- McEnery, T., Baker, P., & Burnard, L. (2000). Corpus resources and minority language engineering. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)* Athens, Greece: European Language Resources Association (ELRA).
- McGillivray, B. & Jenset, G. B. (2023). Quantifying the quantitative (re-)turn in historical linguistics. *Palgrave Communications*, 10(1), 1–6.
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56 – 61).
- Ossolineum (n.d.). Katalogi Ossolineum. https://katalogi.ossolineum.pl/. Accessed: 03.04.2023.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Peirsman, Y., Geeraerts, D., & Speelman, D. (2010). The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16(4), 469–491.
- Ponti, E. M., O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E., & Korhonen, A. (2019). Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3), 559–601.

- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (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*.
- Rayson, P., Archer, D., Baron, A., Culpeper, J., & Smith, N. (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora.
- Regnault, M., Prévost, S., & Villemonte de la Clergerie, E. (2019). Challenges of language change and variation: towards an extended treebank of medieval French. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)* (pp. 144–150). Paris, France: Association for Computational Linguistics.
- Sánchez-Marco, C., Boleda, G., & Padró, L. (2011). Extending the tool, or how to annotate historical language varieties. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 1–9). Portland, OR, USA: Association for Computational Linguistics.
- Scheible, S., Whitt, R. J., Durrell, M., & Bennett, P. (2011). Evaluating an 'off-the-shelf' POStagger on early Modern German text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 19–23). Portland, OR, USA: Association for Computational Linguistics.
- Soria, C., Russo, I., Quochi, V., Hicks, D., Gurrutxaga, A., Sarhimaa, A., & Tuomisto, M. (2016). Fostering digital representation of EU regional and minority languages: the digital language diversity project. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 3256–3260). Portorož, Slovenia: European Language Resources Association (ELRA).
- The pandas development team (2020). pandas-dev/pandas: Pandas.
- Universal Dependencies (n.d.b). UD for Polish. https://universaldependencies.org/pl/index.html. Accessed: 04.04.2023.
- Waszczuk, J., Kieraś, W., & Woliński, M. (2018). Morphosyntactic disambiguation and segmentation for historical polish with graph-based conditional random fields. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, Speech, and Dialogue* (pp. 188–196). Cham: Springer International Publishing.
- Wróblewska, A. (2018). Extended and enhanced Polish dependency bank in Universal Dependencies format. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)* (pp. 173–182). Brussels, Belgium: Association for Computational Linguistics.
- Zampieri, M., Malmasi, S., & Dras, M. (2016). Modeling language change in historical corpora: The case of Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4098–4104). Portorož, Slovenia: European Language Resources Association (ELRA).
- Zampieri, M., Nakov, P., & Scherrer, Y. (2020). Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26, 595 – 612.