COMPUTATIONAL METHODS FOR IDENTIFYING LANGUAGE VARIATION IN POLISH

Initial results presentation by Maria Irena Szawerna

WHAT IS THIS PROJECT ABOUT? - RECAP

1. Using computational methods to assess and identify the differences between two texts.
   1. Orthography, morphology, syntax
2. A 19th-century Polish memoir (manually annotated) vs. modern Polish corpora ~~(and maybe 17th/18th c.)~~.
3. Historical data, results relevant for potential preprocessing for using modern tools.
4. Potentially relevant for other nonstandard data and the processing thereof.
5. Comparison based on the performance of POS taggers and lemmatizers, statistical comparisons.

RESOURCES (TO BE CONTINUED)

1. NKJP (The National Corpus of the Polish Language)
2. Polish UD treebanks
3. *Morfeusz* 2 tagger
4. UD POS tagger
5. *Marmot* tagger
6. BERT for Polish
7. Stanza toolkit
8. ~~Korba (The Electronic Corpus of 17th and 18th century Polish)~~
9. Wspomnienia Juliusza Czermińskiego (Juliusz Czermiński's Memoir)
10. Polish Dependency Bank (PDB)

PLAN

1. ~~Weeks 3-6: consulting with the supervisor, gathering resources and background reading.~~
2. ~~Week 5: presenting the thesis plan.~~
3. ~~Weeks 6-8: deciding what annotation to use and carrying it out on a chunk of the data.~~
4. ~~(Additionally) Weeks 7-8: spring break.~~
5. ~~Week 9: training a BERT-based POS tagger.~~
6. ~~Weeks 10-12: testing the taggers and the lemmatizer.~~
7. ~~Weeks 13-14: error analysis, identifying methods for processing the error statistics into usable language variation information.~~
8. Weeks 15-22: finishing writing, spare time in case any of the stages before take longer than expected, ~~working with the NKJP programming access, if obtained.~~

PRELIMINARY RESULTS

Repository: https://github.com/Turtilla/swe-ma-thesis (private)

Model	Data	Accuracy
Stanza	PDB	90.89%
	memoir	83.49%
Morfeusz	PDB	97.77%
	memoir	91.01%

Table 1: Lemmatization accuracy per model and per test data type.

Error Type	Raw Freq.	Relative Freq. (%)
y	35	24.14%
proper name	30	20.69%
nie	19	13.10%
spelling	12	8.28%
surname	12	8.28%
capitalization	8	5.52%
abbreviation	8	5.52%
e	7	4.83%
ambiguous	3	2.07%
name	3	2.07%
unidentified	3	2.07%
problematic	2	1.38%
foreign	2	1.38%
archaic	1	0.69%

Table 2: Types of errors and their raw and relative frequencies among the historical tokens that were mislabelled by both Stanza and Morfeusz.

Model	Data	Accuracy	Precision	Recall	MCC
BERT	PDB	99.20%	99.20%	99.20%	99.08%
	memoir	94.50%	94.72%	94.50%	93.77%
Marmot	PDB	97.73%	97.75%	97.73%	97.38%
	memoir	90.61%	90.79%	90.61%	89.30%
Stanza	PDB	98.40%	98.41%	98.40%	98.16%
	memoir	93.31%	93.52%	93.31%	92.43%
UD Cloud	PDB	90.98%	91.17%	90.98%	89.59%
	memoir	83.41%	84.12%	83.41%	81.17%

Table 3: Evaluation measures (accuracy, precision (weighted), recall (weighted)), Matthew's Correlation Coefficient per model and per test data type. Although calculated, F1 is not given since it can be calculated from precision and recall. Per class precision and recall can be found in Appendix C

Error Type	**Raw Freq.**	**Relative Freq. (%)**
ambiguous	208	21.80%
capitalization	199	20.86%
y	109	11.43%
unidentified	62	6.50%
archaic	59	6.19%
UD	58	6.08&
surname	41	4.30%
e	41	4.30%
nie	28	2.94%
ending	24	2.56%
spelling	23	2.41%
proper name	21	2.20%
problematic	20	2.10%
digits	17	1.78%
foreign	13	1.36%
uncommon	12	1.26%
abbreviation	11	1.15%
impersonal	4	0.42%
name	2	0.21%
currency	1	0.11%
special	1	0.11%

Table 4: Types of errors and their raw and relative frequencies among the historical tokens that were mislabelled by at least two of four UPOS taggers (BERT, Marmot, Stanza, UD Cloud).

Model	Data	Accuracy	Precision	Recall	MCC
BERT	PDB	95.65%	95.13%	95.65%	95.47%
	memoir	89.39%	89.75%	89.39%	89.05%
Marmot	PDB	89.27%	88.95%	89.27%	88.83%
	memoir	80.22%	81.34%	80.22%	79.60%
Stanza	PDB	94.29%	94.25%	94.29%	94.05%
	memoir	87.68%	88.44%	87.68%	87.28%
Morfeusz	PDB	94.43%	95.36%	94.43%	94.20%
	memoir	84.26%	86.83%	84.26%	83.76%

Table 5: Evaluation measures (accuracy, precision (weighted), recall (weighted)), Matthew's Correlation Coefficient per model and per test data type. Although calculated, F1 is not given since it can be calculated from precision and recall.

Error Type	Raw Freq.	Relative Freq. (%)
ambiguous	199	38.20%
unidentified	65	12.48%
proper name	52	9.98%
y	39	7.49%
digits	25	4.80%
problematic	22	4.22%
nie	20	3.84%
spelling	18	3.46%
archaic	17	3.26%
foreign	16	3.07%
surname	12	2.30%
uncommon	10	1.92%
currency	8	1.54%
e	7	1.34%
gender	4	0.77%
vocative	3	0.58%
abbreviation	2	0.38%
name	2	0.38%

Table 6: Types of errors and their raw and relative frequencies among the historical tokens that were mislabelled by at least two of four XPOS taggers (BERT, Marmot, Stanza, Morfeusz).

NKJP

- PDB: Out of 7583 queries 44 (0.5802452855070552%) had no hits in NKJP.
- Historical: Out of 4302 queries 346 (8.042770804277708%) had no hits in NKJP.

N-GRAMS

- The results for XPOS are very large and not very informative.
- The results for UPOS show some variation but require more analysis on my end.

	test relative (	t relative (!
ADJ	10,0039	9,01225
ADP	10,4858	11,5302
ADV	3,24538	3,28602
AUX	2,49874	2,55687
CCONJ	3,26026	5,27902
DET	2,51956	4,19016
INTJ	0,02975	0
NOUN	24,9368	23,8577
NUM	0,78829	1,29302
PART	2,86462	2,00272
PRON	4,75355	4,90959
PROPN	3,31975	6,83453
PUNCT	16,7564	11,7052
SCONJ	2,04063	1,93467
SYM	0,0119	0
VERB	11,5656	10,9664
X	0,91918	0,64165

REFERENCES

1. Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media
2. 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.
3. Długosz-Kurczabowa, K. & Dubisz, S. (2006). *Gramatyka historyczna Języka Polskiego*. Wydawnictwa Uniwersytetu Warszawskiego.
4. 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).
5. 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).
6. 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.
7. Kieraś, W. & Woliński, M. (2017). Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego. *Język Polski*, XCVII(1), 75–83.
8. Kłeczek, D. (2021). Dkleczek/bert-base-polish-cased-v1 · hugging face. https://huggingface.co/dkleczek/bert-base-polish-cased-v1.
9. 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.
10. 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).
11. Mueller, T., Schmid, H., & Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 322–332). Seattle, Washington, USA: Association for Computational Linguistics.
12. Ossolineum (n.d.). Katalogi Ossolineum. https://katalogi.ossolineum.pl/ . Accessed: 03.04.2023.
13. 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.
14. Przepiórkowski, A., Bańko, M., Górski, R. L., & Lewandowska-Tomaszczyk, B., Eds. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN.
15. PWN (n.d.). Słownik języka polskiego. https://sjp.pwn.pl/ . Accessed: 04.04.2023.
16. PWN Editorial Team (n.d.). około - definicja, synonimy, przykłady użycia. Accessed: 05.04.2022.
17. Pęzik, P. (2012). Wyszukiwarka PELCRA dla danych NKJP. In A. Przepiórkowski, M. Bańko, R. L. Górski, & B. Lewandowska-Tomaszczyk (Eds.), *Narodowy Korpus Jezyka Polskiego* (pp. 253–273). Wydawnictwo PWN.
18. 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*.
19. 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.
20. Saloni, Z., Woliński, M., Wołosz, R., Gruszczyński, W., & Skowrońska, D. (2015). *Słownik gramatyczny języka polskiego*. Warsaw, 3rd edition.
21. The pandas development team (2020). pandas-dev/pandas: Pandas.
22. The University of Sheffield (n.d.). Universal dependencies POS tagger for pl / Polish. Accessed: 29.12.2022.
23. Universal Dependencies (n.d.a). UD for Polish. https://universaldependencies.org/pl/index.html . Accessed: 04.04.2023.
24. Universal Dependencies (n.d.b). Universal Dependencies. https://universaldependencies.org/treebanks/pl-comparison.html . Accessed: 04.04.2023.
25. Universal Dependencies (n.d.c). Universal POS tags. https://universaldependencies.org/u/pos/ . Accessed: 17.04.2023.
26. Waszczuk, J. (2012). Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of COLING 2012* (pp. 2789–2804). Mumbai, India: The COLING 2012 Organizing Committee.
27. Waszczuk, J., Kieraś, W., & Woliński, M. (2018). Morphosyntactic disambiguation and segmentation for historical Polish with graph-based conditional random fields. In *International Conference on Text, Speech, and Dialogue* (pp. 188–196).: Springer.
28. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Online: Association for Computational Linguistics.
29. 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.