IŻ SWÓJ JĘZYK MAJĄ!
An exploration of the computational methods for identifying language variation in Polish

Maria Irena Szawerna

MASTER'S PROGRAMME IN LANGUAGE TECHNOLOGY
LT2402, SPRING 2023
SUPERVISOR: ALEKSANDRS BERDICEVSKIS
EXAMINER: ASAD SAYEED
OPPONENT: NADINA-MARIANA SUDITU

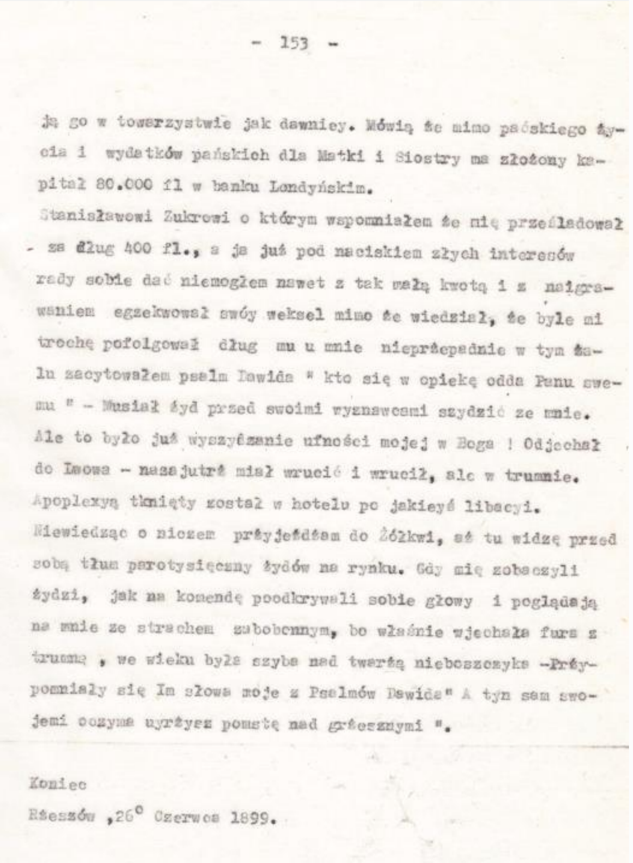
A niechaj narodowie wżdy postronni znają,
*Iż Polacy nie gęsi, **iż swój język mają!***

And may the other nations finally know
that Poles are not geese, **that they have their own language!**

Mikołaj Rej, 1562

Roadmap

- Data
  - Example
- Research Questions
- Background and Related Work
- Experiments
- Results
- Ethical Considerations
- Critiques and Limitations
- Future Work
- Conclusions



- 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 mię 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 pogląda ją na mnie ze strachem zabobonnym, bo właśnie wjechała fura z trumną , we wieku była szyba nad twarzą nieboszczyka -Przypomniały się Im słowa moje z Psalmów Dawida" A tym sam swojemi oczyma uyrzysz pomstę nad grzesznymi ".

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 Questions

1. Is it possible to identify language variation in terms of orthography, morphology, and syntax in a Polish text using tools and resources such as lemmatizers, POS-taggers, and modern corpora?
2. In what ways does the text in question, a 19th-century memoir by Juliusz Czermiński, differ from modern standard Polish?

Background and Related Work

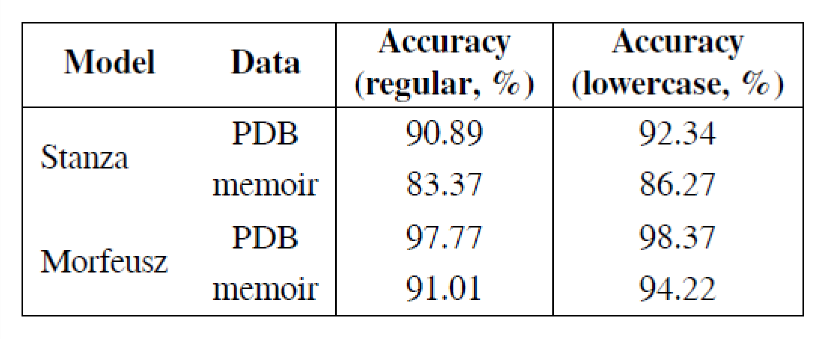
- Features and changes characteristic of 19th-century Polish and the *Kresy* dialects.
- Quantitative and corpus research in historical linguistics.
- Part-of-speech tagging of historical data.
- Methods for dealing with language variation in NLP.
- Detecting and modelling language variation and change.
- Appropriate models and tools (Polbert, Marmot, Morfeusz2+Concraft-pl, Cloud UD tagger, Stanza), resources (PDB-UD, National Corpus of Polish).

Experiments

- POS-tagging and lemmatization:
  - BERT, Marmot, Stanza, University of Sheffield UD Cloud tagger, Morfeusz2.
  - Error annotation for selected erroneous tags.
- N-gram (uni, bi, tri) statistics – approximation of syntactic variation.
- National Corpus of Polish vocabulary comparison.

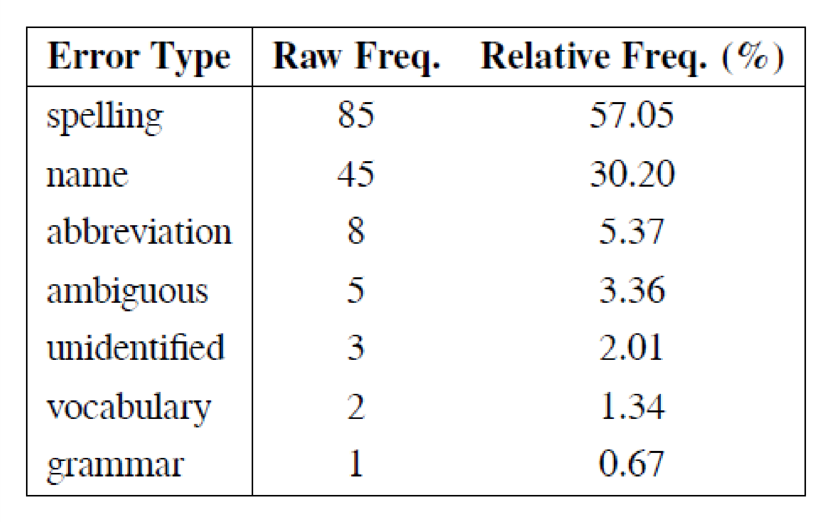
Tool	UPOS-tagging	XPOS-tagging	Lemmatization
BERT	Yes	Yes	-
Marmot	Yes	Yes	-
Stanza	Yes	Yes	Yes
Morfeusz	-	Yes	Yes
UD Cloud	Yes	-	-

Results: lemmatization



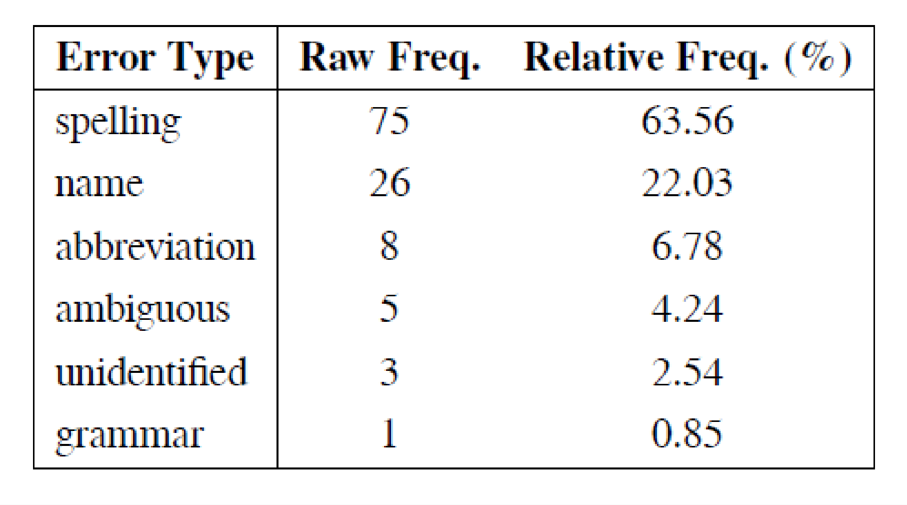
Model	Data	Accuracy (regular, %)	Accuracy (lowercase, %)
Stanza	PDB	90.89	92.34
	memoir	83.37	86.27
Morfeusz	PDB	97.77	98.37
	memoir	91.01	94.22

Original capitalization



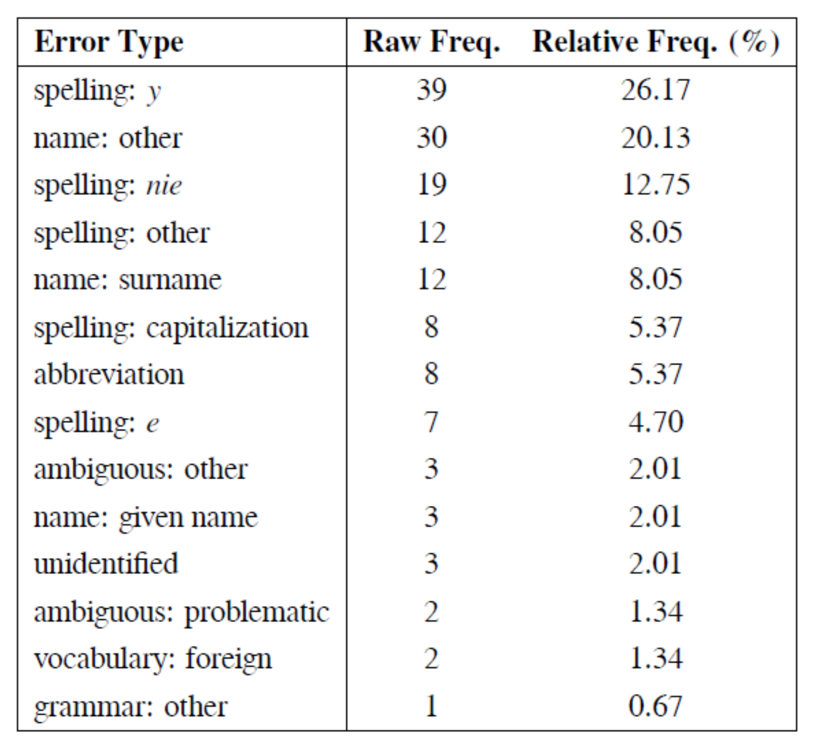
Error Type	Raw Freq.	Relative Freq. (%)
spelling	85	57.05
name	45	30.20
abbreviation	8	5.37
ambiguous	5	3.36
unidentified	3	2.01
vocabulary	2	1.34
grammar	1	0.67

Lowercased



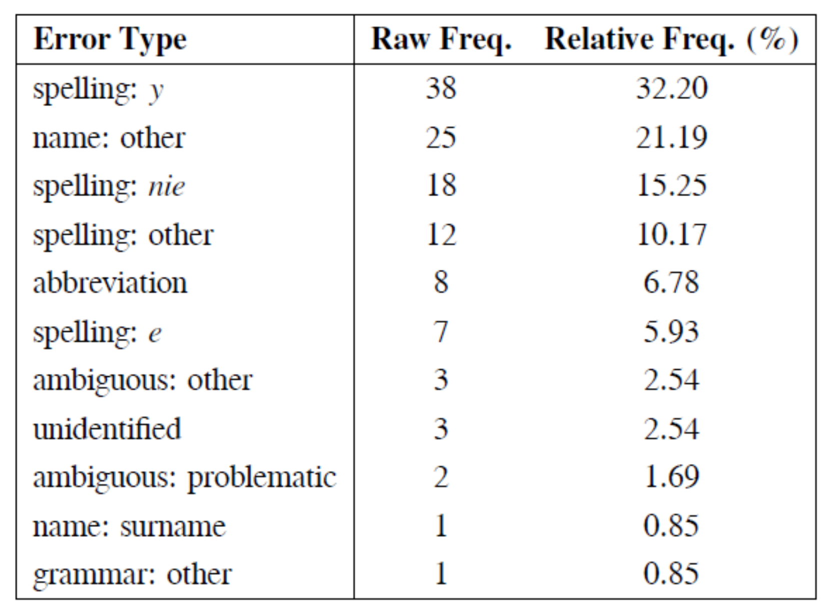
Error Type	Raw Freq.	Relative Freq. (%)
spelling	75	63.56
name	26	22.03
abbreviation	8	6.78
ambiguous	5	4.24
unidentified	3	2.54
grammar	1	0.85

Original capitalization



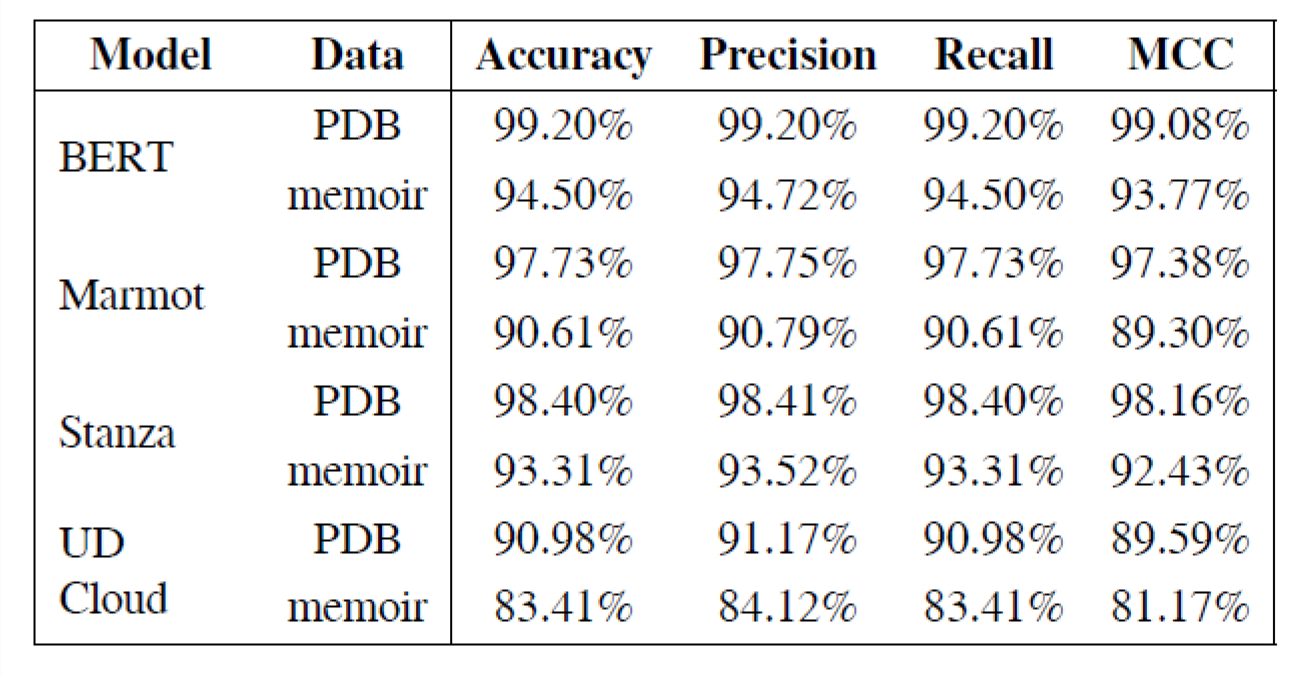
Error Type	Raw Freq.	Relative Freq. (%)
spelling: *y*	39	26.17
name: other	30	20.13
spelling: *nie*	19	12.75
spelling: other	12	8.05
name: surname	12	8.05
spelling: capitalization	8	5.37
abbreviation	8	5.37
spelling: *e*	7	4.70
ambiguous: other	3	2.01
name: given name	3	2.01
unidentified	3	2.01
ambiguous: problematic	2	1.34
vocabulary: foreign	2	1.34
grammar: other	1	0.67

Lowercased

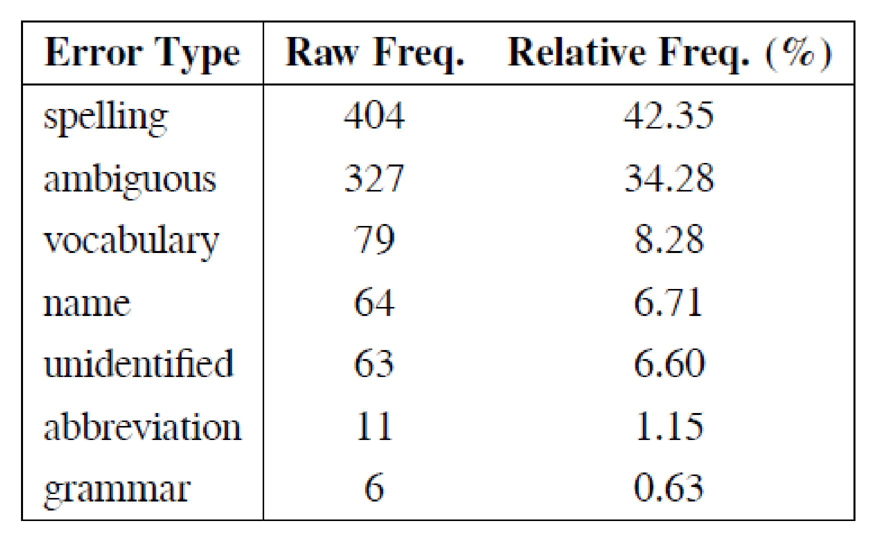


Error Type	Raw Freq.	Relative Freq. (%)
spelling: *y*	38	32.20
name: other	25	21.19
spelling: *nie*	18	15.25
spelling: other	12	10.17
abbreviation	8	6.78
spelling: *e*	7	5.93
ambiguous: other	3	2.54
unidentified	3	2.54
ambiguous: problematic	2	1.69
name: surname	1	0.85
grammar: other	1	0.85

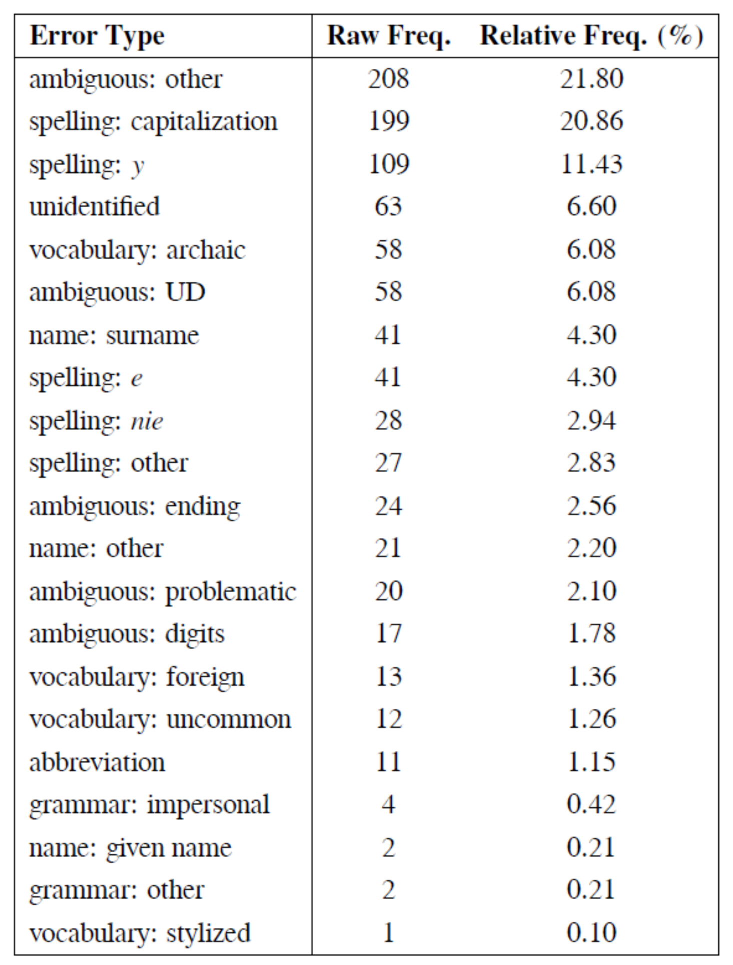
Results: UPOS tagging



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%

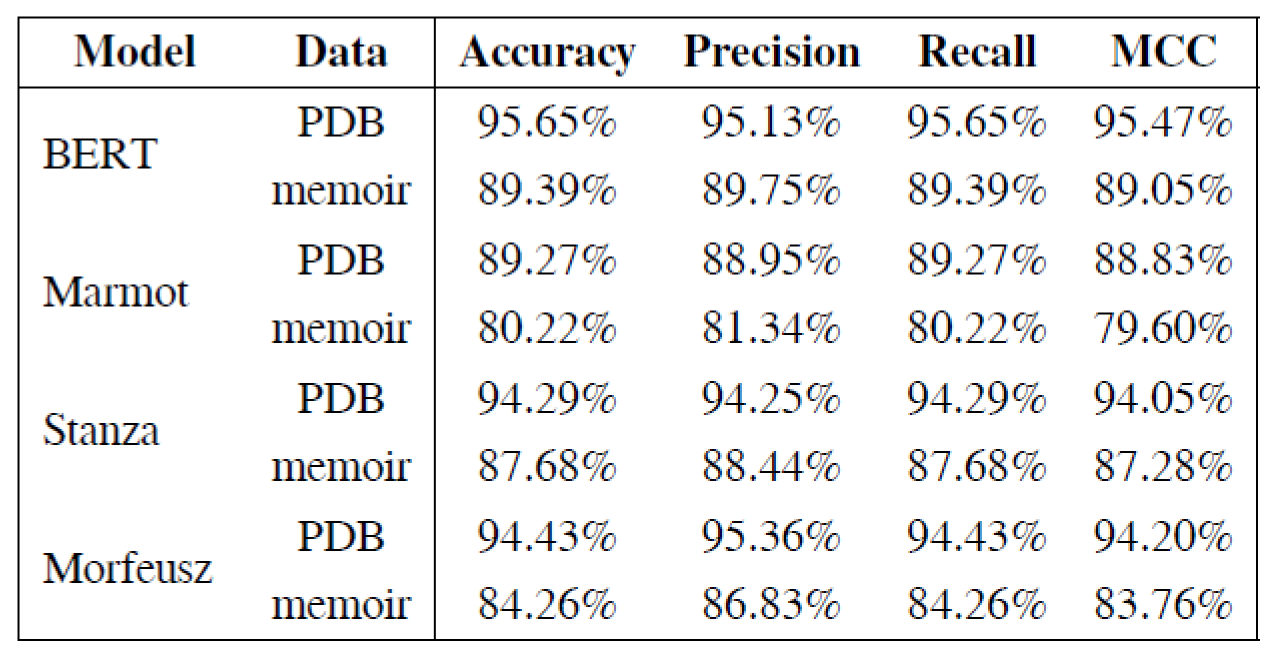


Error Type	Raw Freq.	Relative Freq. (%)
spelling	404	42.35
ambiguous	327	34.28
vocabulary	79	8.28
name	64	6.71
unidentified	63	6.60
abbreviation	11	1.15
grammar	6	0.63

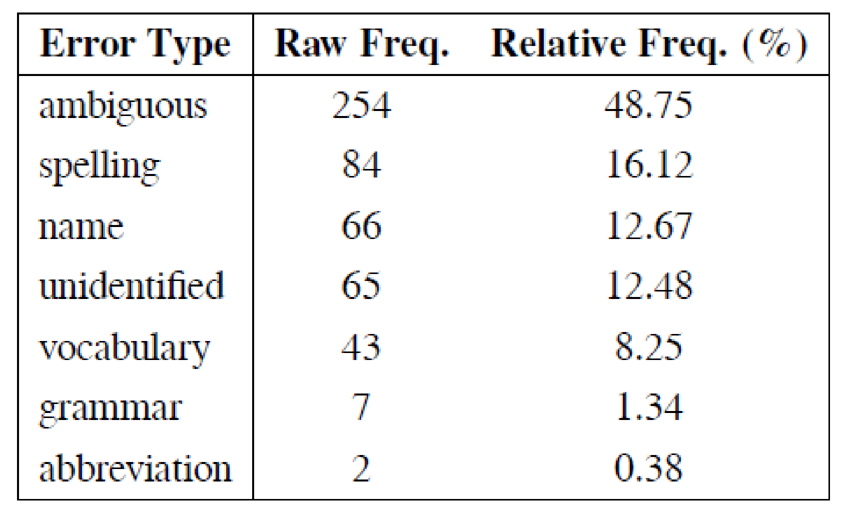


Error Type	Raw Freq.	Relative Freq. (%)
ambiguous: other	208	21.80
spelling: capitalization	199	20.86
spelling: *y*	109	11.43
unidentified	63	6.60
vocabulary: archaic	58	6.08
ambiguous: UD	58	6.08
name: surname	41	4.30
spelling: *e*	41	4.30
spelling: *nie*	28	2.94
spelling: other	27	2.83
ambiguous: ending	24	2.56
name: other	21	2.20
ambiguous: problematic	20	2.10
ambiguous: digits	17	1.78
vocabulary: foreign	13	1.36
vocabulary: uncommon	12	1.26
abbreviation	11	1.15
grammar: impersonal	4	0.42
name: given name	2	0.21
grammar: other	2	0.21
vocabulary: stylized	1	0.10

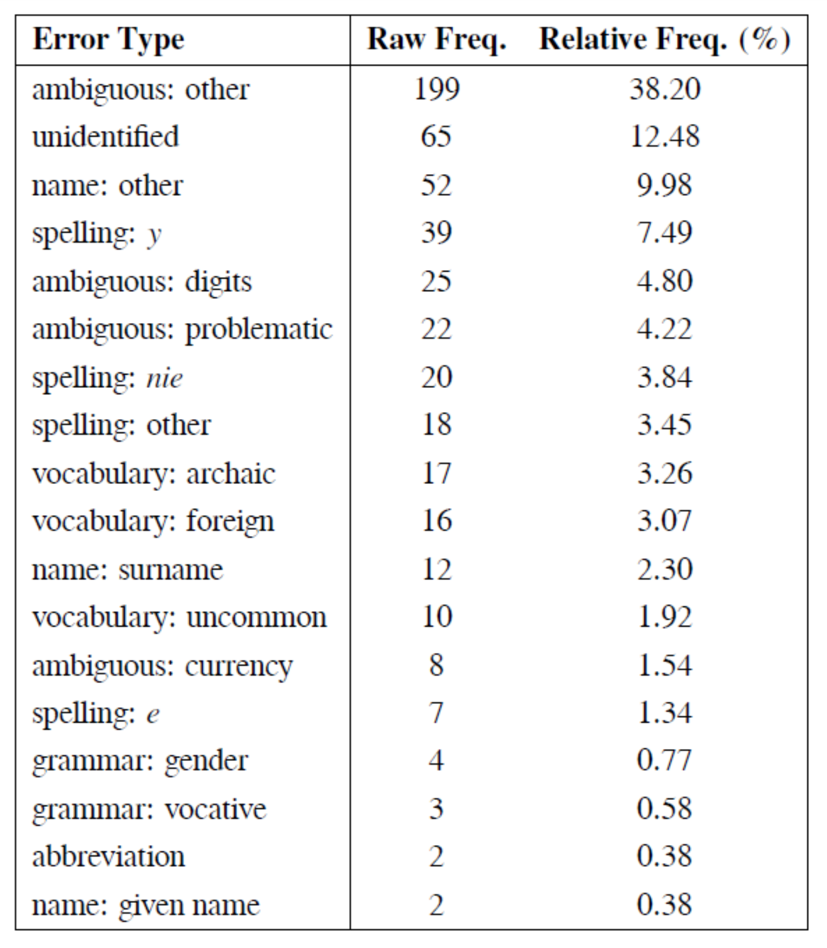
Results: XPOS tagging



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%

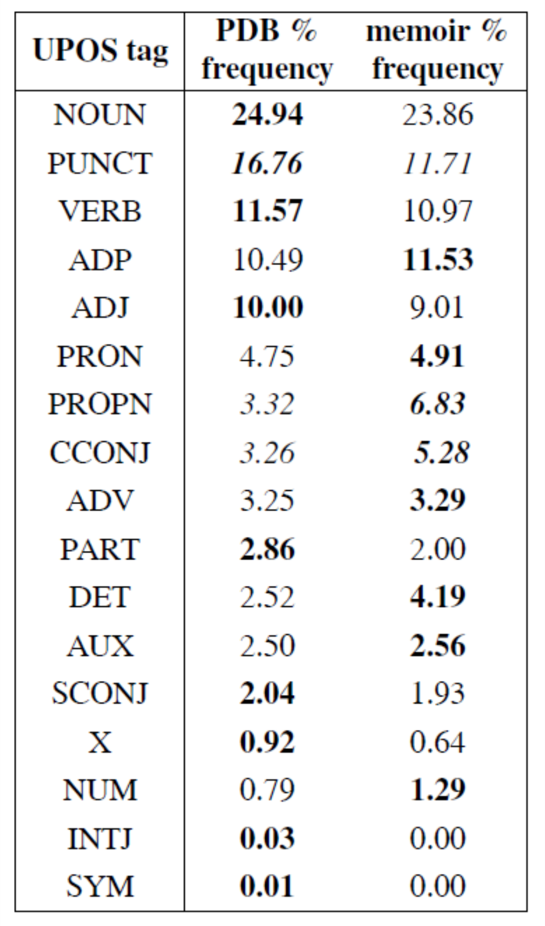


Error Type	Raw Freq.	Relative Freq. (%)
ambiguous	254	48.75
spelling	84	16.12
name	66	12.67
unidentified	65	12.48
vocabulary	43	8.25
grammar	7	1.34
abbreviation	2	0.38

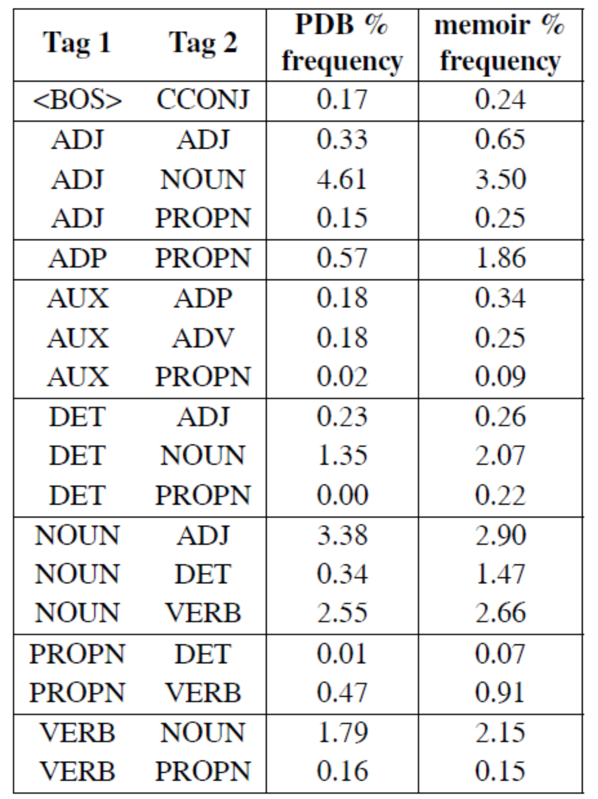


Error Type	Raw Freq.	Relative Freq. (%)
ambiguous: other	199	38.20
unidentified	65	12.48
name: other	52	9.98
spelling: *y*	39	7.49
ambiguous: digits	25	4.80
ambiguous: problematic	22	4.22
spelling: *nie*	20	3.84
spelling: other	18	3.45
vocabulary: archaic	17	3.26
vocabulary: foreign	16	3.07
name: surname	12	2.30
vocabulary: uncommon	10	1.92
ambiguous: currency	8	1.54
spelling: *e*	7	1.34
grammar: gender	4	0.77
grammar: vocative	3	0.58
abbreviation	2	0.38
name: given name	2	0.38

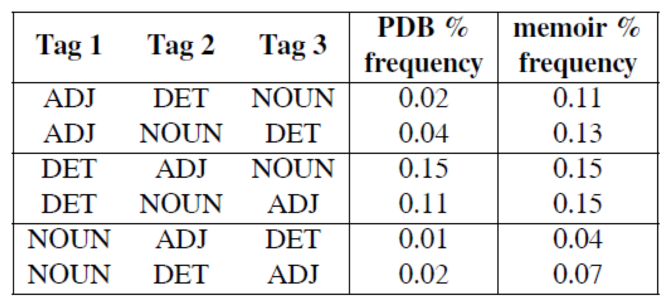
Results: n-gram statistics



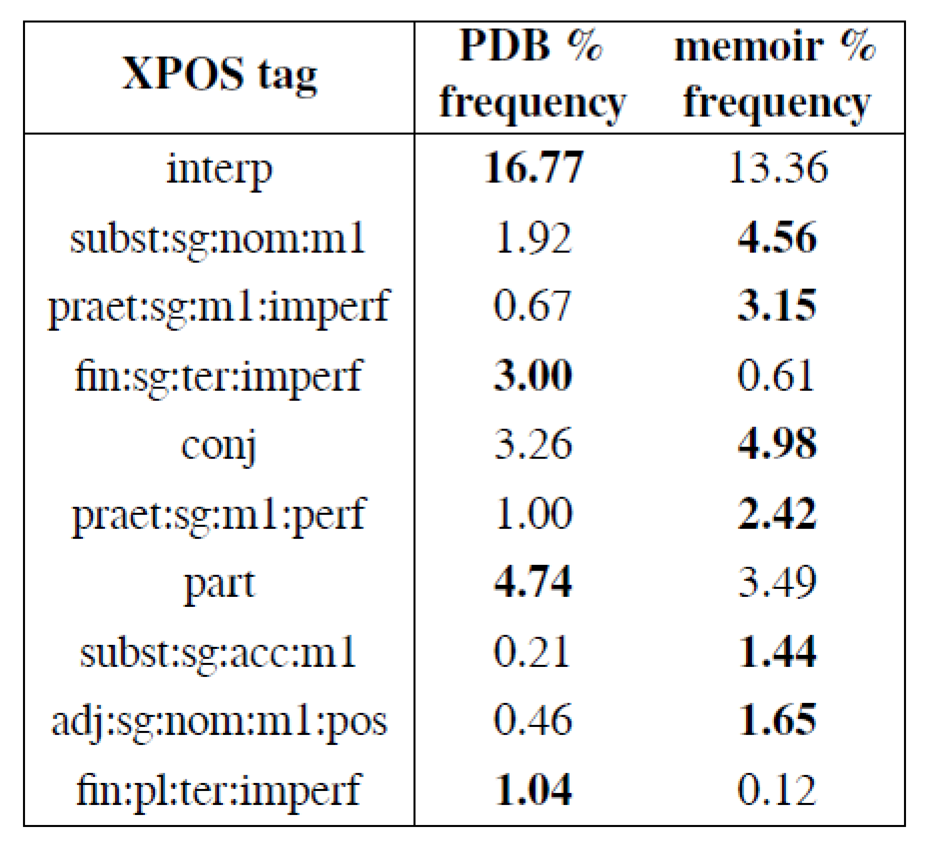
UPOS tag	PDB % frequency	memoir % frequency
NOUN	**24.94**	23.86
PUNCT	***16.76***	*11.71*
VERB	**11.57**	10.97
ADP	10.49	**11.53**
ADJ	**10.00**	9.01
PRON	4.75	**4.91**
PROPN	*3.32*	***6.83***
CCONJ	*3.26*	***5.28***
ADV	3.25	**3.29**
PART	**2.86**	2.00
DET	2.52	**4.19**
AUX	2.50	**2.56**
SCONJ	**2.04**	1.93
X	**0.92**	0.64
NUM	0.79	**1.29**
INTJ	**0.03**	0.00
SYM	**0.01**	0.00



Tag 1	Tag 2	PDB % frequency	memoir % frequency
<BOS>	CCONJ	0.17	0.24
ADJ	ADJ	0.33	0.65
ADJ	NOUN	4.61	3.50
ADJ	PROPN	0.15	0.25
ADP	PROPN	0.57	1.86
AUX	ADP	0.18	0.34
AUX	ADV	0.18	0.25
AUX	PROPN	0.02	0.09
DET	ADJ	0.23	0.26
DET	NOUN	1.35	2.07
DET	PROPN	0.00	0.22
NOUN	ADJ	3.38	2.90
NOUN	DET	0.34	1.47
NOUN	VERB	2.55	2.66
PROPN	DET	0.01	0.07
PROPN	VERB	0.47	0.91
VERB	NOUN	1.79	2.15
VERB	PROPN	0.16	0.15

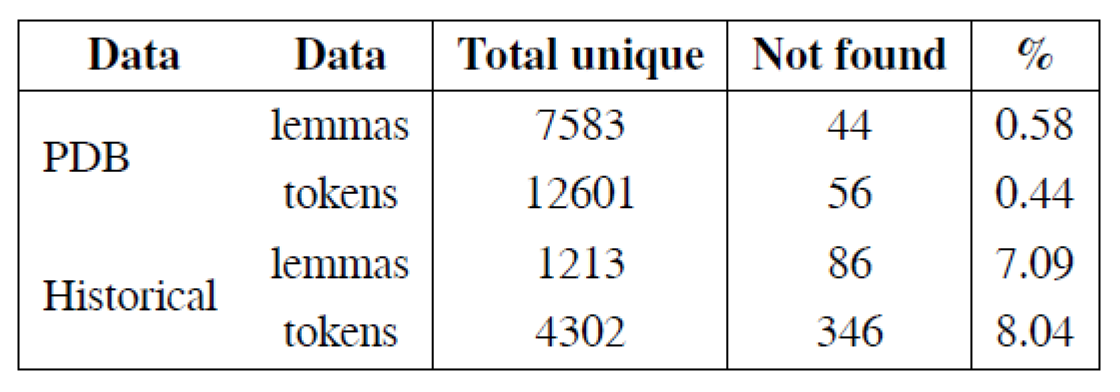


Tag 1	Tag 2	Tag 3	PDB % frequency	memoir % frequency
ADJ	DET	NOUN	0.02	0.11
ADJ	NOUN	DET	0.04	0.13
DET	ADJ	NOUN	0.15	0.15
DET	NOUN	ADJ	0.11	0.15
NOUN	ADJ	DET	0.01	0.04
NOUN	DET	ADJ	0.02	0.07



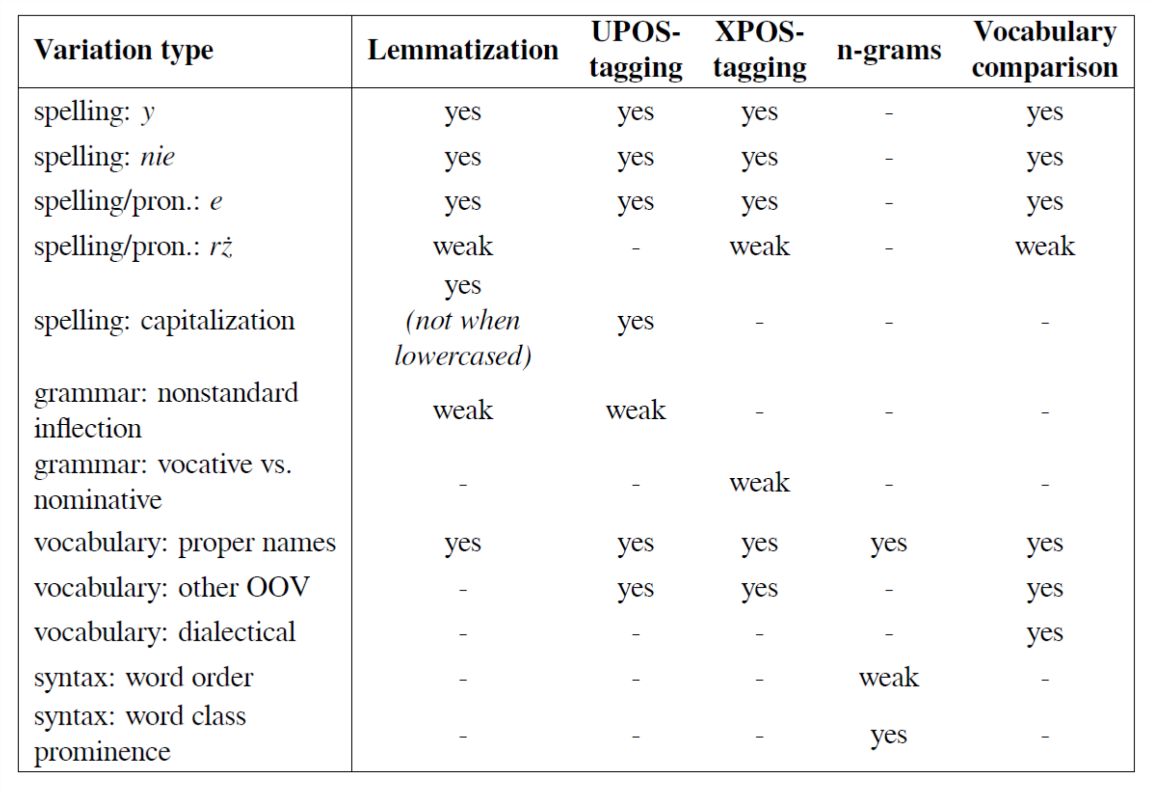
XPOS tag	PDB % frequency	memoir % frequency
interp	**16.77**	13.36
subst:sg:nom:m1	1.92	**4.56**
praet:sg:m1:imperf	0.67	**3.15**
fin:sg:ter:imperf	**3.00**	0.61
conj	3.26	**4.98**
praet:sg:m1:perf	1.00	**2.42**
part	**4.74**	3.49
subst:sg:acc:m1	0.21	**1.44**
adj:sg:nom:m1:pos	0.46	**1.65**
fin:pl:ter:imperf	**1.04**	0.12

National Corpus of Polish vocabulary comparison



Data	Data	Total unique	Not found	%
PDB	lemmas	7583	44	0.58
	tokens	12601	56	0.44
Historical	lemmas	1213	86	7.09
	tokens	4302	346	8.04

Results



Variation type	Lemmatization	UPOS-tagging	XPOS-tagging	n-grams	Vocabulary comparison
spelling: *y*	yes	yes	yes	-	yes
spelling: *nie*	yes	yes	yes	-	yes
spelling/pron.: *e*	yes	yes	yes	-	yes
spelling/pron.: *rż*	weak	-	weak	-	weak
spelling: capitalization	yes *(not when lowercased)*	yes	-	-	-
grammar: nonstandard inflection	weak	weak	-	-	-
grammar: vocative vs. nominative	-	-	weak	-	-
vocabulary: proper names	yes	yes	yes	yes	yes
vocabulary: other OOV	-	yes	yes	-	yes
vocabulary: dialectical	-	-	-	-	yes
syntax: word order	-	-	-	weak	-
syntax: word class prominence	-	-	-	yes	-

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   ...`
  - `...`

Critiques and limitations

- Potential transctiption errors, only some data used.
- Not representative of *Kresy* Polish, just one author.
- No comparison to older data.
- Potentially imperfect annotation.
- Potentially not ideal training setup for taggers.
- Subjective error annotation.
- Lacking n-gram result analysis.

Future work

- Completing the annotation of the data.
  - Adding the syntactic structure annotation.
- Comparison to more data.
  - More data from the same time and region.
  - Older data.
  - Contemporary non-standard data.
- Tagger or lemmatizer confidence.
- Cross-tool agreement.

Conclusions: back to Research Questions

1. Is it possible to identify language variation in terms of orthography, morphology, and syntax in a Polish text using tools and resources such as lemmatizers, POS-taggers, and modern corpora?
    1. Yes, with orthography (and pronunciation) being the most prominent ones.

2. In what ways does the text in question, a 19th-century memoir by Juliusz Czermiński, differ from modern standard Polish?
    1. See the table in the Results section.

Thesis and repository

- Both available at: https://github.com/Turtilla/swe-ma-thesis

UNIVERSITY OF
GOTHENBURG

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.
- Kieraś, W. & Woliński, M. (2017). Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego. *Język Polski*, XCVII(1), 75–83.
- Klemensiewicz, Z. (1976). *Historia Języka Polskiego*. Państwowe Wydawnictwo Naukowe.
- Kurzowa, Z. (1983). *Polszczyzna Lwowa i Kresów Południowo-Wschodnich do 1939 roku*. Państwowe Wydawnictwo Naukowe.
- Kłeczek, D. (2021). Dkleczek/bert-base-polish-cased-v1 · hugging face. https://huggingface.co/dkleczek/bert-base-polish-cased-v1.

- 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).
- 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.
- 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.
- Przepiórkowski, A., Bańko, M., Górski, R. L., & Lewandowska-Tomaszczyk, B., Eds. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN.
- PWN (n.d.). Słownik języka polskiego. https://sjp.pwn.pl/. Accessed: 04.04.2023.
- PWN Editorial Team (n.d.). około - definicja, synonimy, przykłady użycia. Accessed: 05.04.2022.
- 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.
- 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.
- Rej, M. (2015). *Wybór Pism*. Zakład Narodowy im. Ossolińskich.
- Saloni, Z., Woliński, M., Wołosz, R., Gruszczyński, W., & Skowrońska, D. (2015). *Słownik gramatyczny języka polskiego*. Warsaw, 3rd edition.
- 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.
- The University of Sheffield (n.d.). Universal dependencies POS tagger for pl / Polish. Accessed: 29.12.2022.
- Universal Dependencies (n.d.a). SubGender: sub-gender or animacy of masculine referents. https://universaldependencies.org/pl/feat/SubGender.html.
- Universal Dependencies (n.d.b). UD for Polish. https://universaldependencies.org/pl/index.html. Accessed: 04.04.2023.
- Universal Dependencies (n.d.c). Universal Dependencies. https://universaldependencies.org/treebanks/pl-comparison.html. Accessed: 04.04.2023.
- Universal Dependencies (n.d.d). Universal POS tags. https://universaldependencies.org/u/pos/. Accessed: 17.04.2023.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need.
- 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.
- 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.
- 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.
- 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.