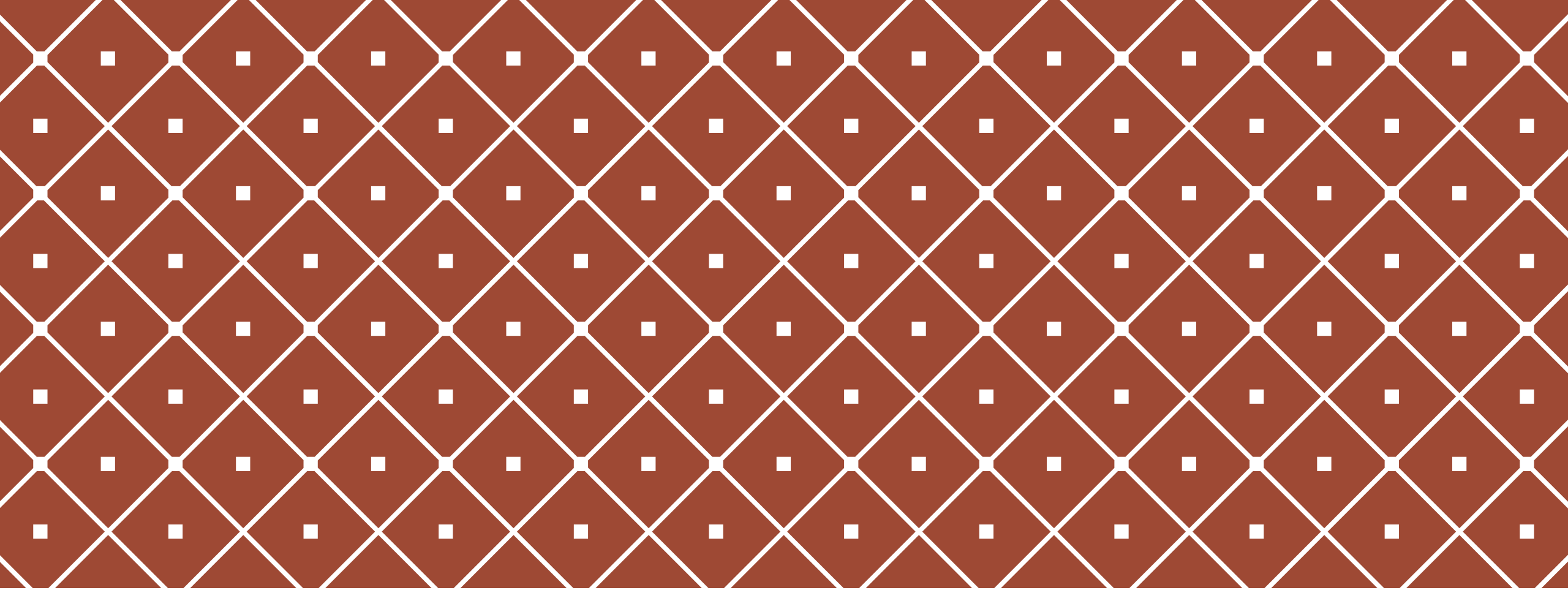

COMPUTATIONAL METHODS FOR IDENTIFYING LANGUAGE VARIATION IN POLISH

A thesis plan presentation by

Maria Irena Szawerna

WHAT IS THIS PROJECT ABOUT?

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.

KNOWLEDGE AND SKILLS

1. Polish
2. Python (including various libraries)
3. Linguistics (including annotation, language variation)
4. Machine Learning (finetuning transformer models)

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 Lemmatizer
8. Korba (The Electronic Corpus of 17th and 18th century Polish)
9. Wspomnienia Juliusza Czermińskiego (Juliusz Czermiński's Memoir)

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.

REFERENCES

1. Yvonne Adesam and Gerlof Bouma. 2016. https://doi.org/10.18653/v1/W16-2104 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.
2. Marcel Bollmann. 2013. https://aclanthology.org/W13-2302 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.
3. Krystyna Długosz-Kurczabowa and Stanisław Dubisz. 2006. *Gramatyka Historyczna Języka Polskiego* [A Historical Grammar of the Polish Language]. Warszawa. Wydawnictwa Uniwersytetu Warszawskiego.
4. Dieuwke Hupkes and Rens Bod. 2016. https://hdl.handle.net/11245/1.535946 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).
5. Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China. Association for Computational Linguistics.
6. Danuta Karwańska and Adam Przepiórkowski. 2011. On the Evaluation of Two Polish Taggers. In Goźdź-Roszkowski, Stanisław (red.), *Explorations across Languages and Corpora: PALC 2009*, pages 105–114, Frankfurt am Mein, Peter Lang.
7. 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.
8. Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. https://aclanthology.org/W11-1503 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.
9. 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.
10. Yi Yang and Jacob Eisenstein. 2016. https://doi.org/10.18653/v1/N16-1157 Part-of-Speech Tagging for Historical English. Pages 1318–1328.