Master Thesis Proposal

1. **Tentative title:** Computational methods for identifying language variation in Polish.
2. **Student's name:** Maria Irena Szawerna
   Contact details: gusszawma@student.gu.se or mariairenaszawerna@gmail.com, 0761917910
3. **Supervisor:** Aleksandrs Berdicevskis
4. **Examiner:** Asad Sayeed
5. **Short description and motivation:**

The goal of this thesis project is to explore computational ways of determining language variation using existing tools and resources (e.g. corpora/word lists, POS taggers, syntactic parsers, large language models, etc.; the extent of the inquiry will be determined at a later stage). Not only would some measure of how "different" a sample is from the standard be useful, but identifying in which areas and to what extent the two differ could be largely useful for areas of linguistics that deal with historical or nonstandard variations; while developing such a measure seems to be beyond the scope of this thesis, testing methods for the detection of linguistic variation could contribute to future research in this direction. For example, systematic detection of these differences could speed up comparative analyses of dialects, especially when there is much data to compare; understanding the ways in which a text diverges from the standard could lead to developing better pre-processing tools for it to undergo further processing (e.g. POS tagging), which then, in turn, could be used for other kinds of quantitative inquiries about the nature of the dialect. On the other hand, the identification of variation in language does not only concern historical linguistics, sociolinguistics, or dialectology. There do exist issues with NLU systems not handling nonstandard pronunciation, vocabulary, or syntax well. While this project does not intend to provide a solution for these issues, perhaps it could help with pinpointing the differences and therefore with the faster development of solutions to counteract the issues arising from those differences.

The project will utilize a previously unpublished, manually annotated (to the extent that it is needed) text in late 19th-century Polish, which is not standard. This data will then be used to test various tools trained on / developed for modern Polish (e.g. NKJP, Polish UD, Polish Parliamentary Corpus, various other corpora), and the results of that testing will hopefully reveal whether it can be used for variation identification: depending on the kind of access that I have to the various tools listed, I would, first and foremost, like to see what kind of variation (orthography, morphology, perhaps syntax) can be identified using various existing taggers and lemmatizers. Depending on the time and the kind of access that I gain to some of the resources, I would also like to determine in what ways the lexicon of the text in question differs from modern Polish (i.e. compare which words are not present in modern Polish resources). While I intend to detect the variation in the aforementioned 19th-century memoir, I want to compare the results obtained from this data with those from the pre-annotated corpus of 17th and 18th-century Polish, if time allows.

6. **Required knowledge and skills:**

a. Polish

b. Python (including various libraries)

c. Linguistics (including annotation, language variation)

d. Machine Learning (finetuning transformer models)

7. **A list of resources[1]:**

a. NKJP (The National Corpus of the Polish Language), hopefully including programming access (in progress)

b. Polish UD treebanks

c. *Morfeusz 2* tagger

d. UD POS tagger

e. *Marmot* tagger

f. BERT for Polish

i. And the corpora it is trained on

g. Stanza Lemmatizer

h. Korba (The Electronic Corpus of 17th and 18th century Polish)

i. *Wspomnienia Juliusza Czermińskiego* (The Memoirs of Juliusz Czermiński)

8. **Work plan:**

a. Weeks 3-6: consulting with the supervisor, gathering resources and background reading.

b. Week 5: presenting the thesis plan.

c. Weeks 6-8: deciding what annotation to use and carrying it out on a chunk of the data.

d. (Additionally) Weeks 7-8: spring break.

e. Week 9: training a BERT-based POS tagger.

f. Weeks 10-12: testing the taggers and the lemmatizer.

g. Weeks 13-14: error analysis, identifying methods for processing the error statistics into usable language variation information.

h. 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.

9. **References (WIP):**

a. 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.

b. 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.

c. Krystyna Długosz-Kurczabowa and Stanisław Dubisz. 2006. *Gramatyka Historyczna Języka Polskiego* [A Historical Grammar of the Polish Language]. Warszawa. Wydawnictwa Uniwersytetu Warszawskiego.

d. Dieuwke Hupkes and Rens Bod. 2016. https://hdl.handle.net/11245/1.535946 POS-tagging of Historical Dutch. In *LREC 2016: Tenth International*

[1] Not all will necessarily be used, as per point 5.

Conference on Language Resources and Evaluation, pages 77–82, Paris. European Language Resources Association (ELRA).

e. 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.

f. 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.

g. 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.

h. 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.

i. 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.

j. Yi Yang and Jacob Eisenstein. 2016. https://doi.org/10.18653/v1/N16-1157 Part-of-Speech Tagging for Historical English. Pages 1318–1328.