DEPARTMENT OF PHILOSOPHY, LINGUISTICS AND THEORY OF SCIENCE

IŻ SWÓJ JĘZYK MAJĄ!

An exploration of the computational methods for identifying language variation in Polish

Maria Irena Szawerna

Master's Thesis:	30 credits
Programme:	Master's Programme in Language Technology
Level:	Advanced level
Semester and year:	Spring, 2023
Supervisor:	Aleksandrs Berdicevskis
Examiner:	Asad Sayeed
Keywords:	language variation, Polish, diachronic linguistics, part-of-speech tagging, lemmatization, corpus linguistics

Abstract

Computational approaches to language variation continue to contribute in a relevant way to various fields, including Natural Language Processing (NLP) and linguistics. Being able to accommodate variation within natural language increases the robustness of NLP models and their usefulness in real-life applications; simultaneously, detecting and describing variation and trends that govern it is one of the main goals of sociolinguistics and historical linguistics, meaning that some of the advances in NLP can contribute to these fields as well. As one of the current trends in historical linguistics appears to be quantitative and corpus research, the need for annotated historical data is becoming more and more apparent.

Within this thesis, a selection of tools and methods are tested for their ability to detect variation between a manually annotated sample of non-standard historical Polish and corpora of modern Polish and tools based on them. The experiments include part-of-speech tagging with two tagsets, lemmatization, vocabulary comparisons, and an n-gram analysis. The results reveal what kinds of variation each approach can discover in the text and to what extent. Since the majority of the presented methods require the data to be annotated, they would be time- and resource-consuming if applied to larger corpora; nevertheless, they do reveal certain trends in variation as well as information on what kind of preprocessing may be needed for this sort of data to be successfully automatically annotated, which could enable the creation of a larger corpus, facilitating further research. Additionally, a comparison of tagging and lemmatizing performance of various tools on modern Polish is presented, and the annotated historical text itself constitutes a relevant contribution as well.

Preface

This thesis would not have taken shape without a number of wonderful people, and I would like to take some time to thank them all.

I would like to thank my supervisor, Sasha Berdicevskis, for believing in the initial idea I had and for guiding me through many of the decisions that shaped the thesis into what it is today, as well as for showing me that I need not forsake my love of historical linguistics even in this program.

I will forever be grateful to the rest of the teachers and staff of the MLT program, who have taught me so many of the skills that I put to the test while working on this topic, but also many more that I will hopefully utilize in the future. I am also happy to have met all the wonderful people that I studied with throughout these two years. I cannot wait to see what future lies ahead of us!

I want to extend my deepest thanks to my family members who were involved in procuring the historical data: my grandfather, Piotr Kociatkiewicz, who meticulously transcribed our mutual ancestor's memoir from scans provided by my aunt Anna Chodorowska. It was my grandfather's question about whether he should standardize the spelling of the text that sparked my interest in it as a linguistic resource — and without that question and the endeavor to transcribe the memoir, this thesis would not have been written.

I am also deeply grateful for the time and help given to me by dr hab. Piotr Pęzik, who not only made it possible for me to access the National Corpus of Polish (NKJP) but also took his time to explain to me its ins and outs.

I want to thank all of my friends and family whose support and faith in me carried me through the ups and downs: my father, Michał Szawerna, for his eagerness to discuss all things linguistics with me; my grandmother, Ewa Wajda-Kociatkiewicz and uncle Mariusz Chodorowski for their faith in me; my partner, Tobias Ström, for his patience and for believing in me at times when I did not; my friends — Tomek Kos, Kasia Kuśmierczyk, Jakub Wróbel, Piotrek Szulc, Sandra Kierasińska, Freja Edvardsson, and so many others — for their unwavering support and friendship, despite some of us being hundreds of kilometers apart.

Finally, I hope that my mother, Justyna Kociatkiewicz, would be proud to see how far I have made it — and that she would be happy to see that her love for our family history and heritage persists in this thesis.

Contents

List of Figures

List of Tables

1 Introduction

In his iconic quote, referenced in the title of this thesis, the famous 16th-century Polish poet Mikołaj Rej declared his conviction that Polish people have their own language:

> *A niechaj narodowie wżdy postronni znają,*
> *Iż P o l a c y nie gęsi, iż swój język mają!*[1]
>
> Rej (2015)[2]

While one would be hard-pressed to find someone trying to dispute the existence of the Polish language, one perhaps should consider certain alterations to the quote itself: the average Polish user no longer speaks in the same fashion as Mikołaj Rej did when he wrote his poetry, and variation occurs not only diachronically, but also regionally. Perhaps it would be more correct to say then that Poles *swe języki mają!*[3]

Language variation plays an essential role in natural language processing: natural language as used by speakers is ever-changing, and NLP tools have to, to some extent, account for even the synchronic variation. As Zampieri et al. (2020) or Dorn (2019) point out, developing methods to handle language variation is also relevant for adapting the existing tools to minority languages or dialects. While there are methods that allow for the utilization of raw data, Ponti et al. (2019) remarks that other methods still rely on annotated data which is difficult to come by for less popular languages. Simultaneously, as Jenset & McGillivray (2017) note, historical linguistics on its own is a data-driven field, and access to data, as well as methods for processing it, are very important; at the same time, they highlight the usefulness of annotated corpora. This annotation may be especially useful when it comes to languages with a richer morphology, such as Slavic languages, as it may enable e.g. searching for all the inflectional forms of a given word (Pęzik, 2012). Thus, identifying the kinds of variation that occur between two languages or dialects not only constitutes a contribution to the body of knowledge about those languages on its own but also opens up possibilities for adapting existing tools for major languages to be used for the partial automatizing of data annotation.

1.1 Research Questions

The inquiries conducted within this thesis are based on a 19th-century memoir by Juliusz Czerniński, who came from the area of nowadays Eastern Poland and Western Ukraine. Given its age as well as the sociopolitical and geographical context, the text is expected to differ from standard modern Polish, and to bear features typical for the historical dialect of that area. The data itself and the annotation process (conducted by the author of the thesis) are outlined in more detail in subsection 3.1 and subsection 3.2. With the previously mentioned issues of language variation both in NLP and in historical linguistics, and the research described in section 2, the hope is for this text to yield some insights concerning possibilities for the identification of language variation in Polish using computational tools and resources.

[1] 'And may the other nations finally know that Poles are not geese, that they have their own language!'
[2] While the reference is to a 2015 edition of the poet's collected works, the poem itself was written in 1562.
[3] 'have their own languages!'

Consequently, the aims of this thesis can be described as two sides of the same coin, as it seeks to simultaneously answer the following two 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?*

1.2 Motivation

As mentioned at the start of this section, investigating language variation is not only its own field of linguistics but is also relevant for NLP — and methods and discoveries within these fields can inform each other. It is interesting to see how tools intended for working with modern languages can be used to identify ways in which those differ from their historical counterparts. These differences can help inform the pre-processing of the texts or ways in which the tools need to be adapted to enable a more reliable data annotation, which, in turn, can be used for further qualitative, corpus-based inquiries into some historical form of a language.

Simultaneously, language change is not only a thing of the past, but a continuous process that may lead to modern tools becoming outdated in the future; furthermore, language varies also based on factors other than time, such as geography or social class, and this variation can also prove problematic to NLP applications, e.g. when it comes to language recognition or processing of historical data (Jurgens et al., 2017; Regnault et al., 2019; Zampieri et al., 2020). While methods such as cross-linguistic transfer learning do exist, they cannot be applied to all the tools equally, especially non-neural ones or ones that are not available for fine-tuning.

While the major focus of this thesis is on exploring historical and dialectical language variation in Polish alongside the methods that can be employed for such investigations, hopefully, it can yield insights into and spark some discussion about related topics, such as the handling of linguistic variation in NLP, computational methods in historical linguistics, as well as resources for historical linguistics and the annotation thereof.

1.3 Contributions

Within this thesis, a variety of ways for discovering historical linguistic variation using tools intended for modern languages is tested. The methods are reviewed with regard to the kind of results that they yield and the amount of annotation or preparation needed. While the majority of the methods do require the data to be annotated, a number of observations (described in more detail in section 4) can be made based on the results, such as the influence of nonstandard capitalization on tagging and lemmatizing, prevalent spelling differences reflecting a different spelling standard or different pronunciation, issues that the existing tools have with various kinds of proper names; additionally, certain gender bias is revealed to exist as far as the XPOS tagset and the tools used to annotate with it, which can hopefully contribute to the discussion concerning the tagsets used for annotating data in Polish and certain biases that are possibly present in the training data as well.

Simultaneously, due to the nature of the experiments, a second set of observations emerges, as a comparison of the tagging and lemmatizing performance of multiple tools on modern data is conducted, which can inform the choice of a tool for other research or real-life application. Finally, the data that has been manually annotated by the author of this project and that is utilized within it constitutes a non-negligible contribution to the body of annotated historical data for Polish.

1.4 Scope

This thesis encompasses attempting to identify language variation in a specific text with the use of tools such POS taggers and lemmatizers, resources such as corpora, and methods such as n-gram count analysis. These are used to approximate variation in terms of spelling, vocabulary, and syntax, although certain observations concerning the pronunciation (the consistent shift from /a/ to /ɛ/) are also made. However, due to the data in question coming from a single author, the conclusions cannot be extended to the language of the time and region at large.

2 Background and Related Work

2.1 The history of Polish (19th-century)

When discussing a historical variant of a language it is important to situate it in the context of what is known about the history of that language in general. Historical grammars and descriptions are available for many of the major languages, and Polish is no exception — although recently some concerns have been raised; Dunaj (2019) states that some of the most thorough descriptions of the history of Polish may be outdated given more recent research, and therefore in need of being updated. Nevertheless, much of the general information from these sources remains relevant and can be supported by more recent publications.

In what is considered to be one of the core texts of the study of historical Polish, Klemensiewicz (1976) describes the development of that language from mid-12th century. Throughout the book, the author adopts the following periodization of the language[4]: Old Polish (until 1500s), Middle Polish (from 1500s to 1780s), New Polish (from 1780s to 1939), with no information provided about the period between 1939 and the modern times; however, it is worth pointing out that the book was written in 1976, and the author claims that the changes induced by the post-WW2 events are still in motion and refrains from characterizing them. Długosz-Kurczabowa & Dubisz (2006) adopt a similar division, but they include a subdivision of both the Middle Polish and New Polish periods into parts 1 and 2, introducing a more fine-grained distinction.

While Klemensiewicz (1976) provides an extremely thorough characterization of each of those periods, including the socio-historical context, only the New Polish period appears to be relevant to the topic of this thesis, and therefore it is this period whose characterization will be discussed in more detail. The duration of the New Polish period is characterized by many political and social changes, including a long-term occupation by a number of various countries; nevertheless, that period of unrest was preceded by some attempts at describing the contemporaneous Polish language. Although the time of partitioning of Poland was characterized by certain repressions when it came to e.g. receiving education in Polish, the language was not pushed out of use, partly due to different areas being subject to different rules, and partly due to strong resistance and a "national spirit." Although this situation only worsened as time progressed, the author attributes the survival of the Polish language and identity, especially in East Poland, to family and traditional upbringing. It is relevant to point out that even this seemingly difficult time featured the creation of e.g. dictionaries of Polish.

Klemensiewicz (1976) characterizes the changes that were taking place in the New Polish period as the following:

1. Phonology: final loss of what the author spells as *ȧ, ė, ȯ* (close versions of /a/, /ɛ/, /o/, a remnant of the long-short distinction in the Polish vowel system); reduction of /ija/ and /ɨja/ to /ja/ in borrowings; depalatalization of syllable-final soft consonants; adoption of the pronunciation /ɕr/ over /ɕz̨/ and /ʑr/, /jz̨/ over /ʑz̨/; variation in the pronunciation of sibilant

[4] This periodization appears to be motivated by historical events such as the turning point between the Middle Ages and the Renaissance (1500s), the Enlightenment and the partitioning of Poland (late 1700s), the start of the Second World War (1939).

sounds; establishment of modern-like stress patterns (penultimate syllable, including clitics; separate stress in the constituent words of compound phrases).

2. Inflection of nouns: continued variation in the masculine singular genitive and dative forms; adoption of the plural accusative as the plural nominative form in the masculine paradigm, with a following pejoration of these forms in some cases; loss of the *-a* ending for plural nominative forms Latin borrowings; gradual loss of the *-ów* ending in the plural genitive masculine nouns whose root ends in a soft consonant; plural genitive endings taking over plural accusative, the original endings persisting only in stylized texts; *-ę* replacing *-ą* in the majority of feminine nouns in singular accusative; variation between *-e*, *-i*, and *-y* as nominative, accusative, and vocative forms of plural feminine nouns; loss of the "masculine" ending *-ów* used in feminine and neuter plural genitive nouns; temporary loss of the nasal /ɛ̃/ in favor of /ɛ/ in nominative and accusative singular forms of neuter nouns; loss of the *-y* and *-i* endings for the plural instrumental forms of neuter and masculine nouns in favor of *-ami* and *-mi*.

3. Inflection of pronouns and adjectives: gradual loss of the *-ę* ending in singular accusative feminine pronouns in favor of *-ą*; continued variation in the endings (*-ym, -im, -em, -ymi, -imi*) for the instrumental and locative singular masculine and neuter forms as well as the plural instrumental forms of pronouns and adjectives.

4. Inflection of numerals: generalization of the *-u* ending in numerals; two parallel forms of the dative form of *dwa* 'two', replacement of *-ą* with *-u* in the instrumental case of numerals.

5. Verbal inflection: replacement of *-m* with *-my* in the first person plural in the present tense; the loss of the pluperfect tense; preference for attaching the conditional marker to the l-participle; preference for constructing the future tense by combining an auxiliary verb *być* 'to be' with the l-participle instead of the infinitive; development of the *-szy* ending of the past participle forms into *-łszy* or *-wszy*; attempts of reviving the dual forms.

6. Word formation: increase of the popularity of zero-derivation; fall of the popularity of derivational suffixes such as *-ak, -nik, -ły, -ec*; preference for derivational suffixes *-arz, -acz, -dło* in narrow fields (e.g. technical, artistic); prominence of derivational suffixes *-ik, -ina, -isko, -ość, -stwo, -i, -y, -ić*; change of meaning of the derivational suffix *-ek*; prevalence or acronyms.

7. Syntax: preference for the nominal subject complement to be in the instrumental case, while for an adjectival one to be in the nominative; replacement of the genitive by the accusative in object position; tendency to inflect *jeden* 'one' in double-digit numerals; preference for rection over agreement in double-digit numerals ending with *dwa* 'two'; decrease in popularity of the preposition *ku* 'towards, for, in order to'; rise in popularity of a new form of the final clause consisting of a conjunction followed by the infinitive form of the verb if the subject of both clauses is the same; loss of a variety of conjunctions and particles; variation between the pronouns *co* 'what, which, who' and *który* 'which, who'.

8. Lexicon: increase in specialized vocabulary; raised awareness of synonyms and near-synonyms and their usefulness; loss of a variety of words; a variety of neologisms or new derived words; largely negative attitudes to foreign borrowings, although those are numerous (including borrowings from Latin, French, German, Russian, Ruthenian, English).

9. Orthography and orthoepy: multiple attempts at standardization, one reform made in 1918; little adherence to normalized spelling in the press and other sources; final reform of the period in 1936; attempts at normalizing pronunciation in the early 1900s.

Other sources, such as Długosz-Kurczabowa & Dubisz (2006), opt for separate descriptions of the evolution of different aspects of the language; nevertheless, they do provide summaries of changes happening at certain points in time. They divide the New Polish period into two parts, with the breaking point between the two around 1900. They additionally list the following changes for the period preceding that tipping point: preference for the accusative in objects; preference for rection over agreement in numerals; differentiation of the passive from subject-less expressions; increase in use of conjunctions in complex sentences; loss of the impersonal forms with the passive participle; loss of the *accusativus cum infinitivo* construction; preference for formal rather than semantic agreement; preference for infinitives over l-participles in some final clauses; strengthening of the subordinate status of participle clauses; nominalization; increase in number of periphrastic expressions; decrease in the use of modals; shortening of sentences, decrease in popularity of very complex sentences; finer definition of connective words or phrases; division into a colloquial and formal norm.

While both sources overlap significantly in terms of the changes that they mention, they provide slightly different perspectives and levels of generalization when it comes to the changes characterizing this time period.

2.2 *Kresy* (Borderlands) Polish

A fair amount of work has gone into characterizing specific dialects of Polish, including the ones present in former eastern Poland (lands that were no longer part of Poland after the Second World War). Since that is the area of origin of the data discussed in this thesis, this research is worth mentioning. Kurzowa (1983) presents an extensive discussion of the topic. She characterizes the area as ethnically diverse, with the main distinction being between the ethnically Polish and Ukrainian populations, a divide which was at times aggravated by historical events. Some of the oldest textual evidence of the dialect in question comes from the 15th century, in a text which displays features similar to those of Polish from the region of Lesser Poland with clear East Slavic influences; these two languages are essential for the development of "Borderlands Polish[5]." It is worth pointing out that, as time progressed, so did this dialect evolve, with regional and social variation, to a large extent dependent on the social dynamics of the Polish and Ukrainian populations in a given region.

Kurzowa (1983) characterizes the Borderlands dialect of Polish in the following fashion:

1. Phonetics and phonology: intensified and prolonged pronunciation of stressed syllables resulting in a differentiation of pronunciation of vowels in stressed and unstressed syllables (/ɛ/, /o/ raised to /i/, /ɨ/, /u/, /a/ raised to /ɛ/ in unstressed syllables, /i/, /ɨ/ lowered to /e/ in stressed syllables); use of /ɨ/ in place of /i/ in some words; use of /i/ instead of /ɛ/ in

[5]The Polish term *Kresy (Wschodnie)*, which describes the area in question, is often translated as '(Eastern) Borderlands'; therefore, the name for this dialect used throughout this thesis will be 'Borderlands Polish', so as to make it more salient than 'Kresy Polish'.

some words; denasalization of nasal vowels; ellipsis or insertion of some vowels; preference for /ɛ/ over /o/ in verb roots; labialization of back vowels; stress on syllables preceding the penultimate one in past tense verb forms; lack of palatalization of labial consonants, replaced by a separate phone or /ɲ/; replacement of palatalized /t/ and /d/ with /ɕ/ and /d͡ʑ/; palatalization of all all velar consonants before /i/ and /ɛ/; presence of both palatalized and unpalatalized /l/, variation in pronunciation of *ł*; presence of both bilabial and labiodental voiced fricatives; distinction between voiced and voiceless *h*; evolution of /r̪/ into /rzˬ/ or /rʂ/, or even just /r/, as opposed to just /zˬ/ or /ʂ/; variation in sibilant sounds; insertion of /m/ before labial consonants and /n/ before other consonants; lack of devoicing of some consonant clusters; affrication of certain consonant clusters; voicing of /s/ following /r/ or /n/; assimilation, simplification, dissimilation in a selection of words; cross-word boundary voicing.

2. Nominal inflection: variation in grammatical gender relative to Polish; lack of distinction between masculine and non-masculine nouns in the plural; a larger number of masculine nouns ending in *-o* in the nominative; lack of inflectional distinction between nominative and vocative for masculine nouns; larger prominence of *-a* as a possible inflectional ending for the genitive case for masculine nouns; prominence of the *-owi* ending in the dative case for masculine nouns; prominence of *-a* as an ending for foreign masculine nouns in plural nominative; prominence of *-am* as an ending for plural dative for masculine nouns; tendency to end all feminine nouns with *-a*; loss of the archaic dual form *ręce* ‘two hands, two arms’; loss of accusative *-ę* endings in favor of *-y* or *-i*; *-och* as an ending for numerals in genitive and locative; tendency to select short forms of personal pronouns.

3. Verbal inflection: changes to various verb roots by analogy to other verbs; expression of the past tense by a combination of a personal pronoun and the l-participle; noticeable mobility of endings expressing person; preference for *-y, -i, -ym*, or *-im* instead of *-ę, -em* in the first person singular forms of present tense verbs; preference for *-m* instead of *-my* in first person plural forms of present tense verbs; homonymy between first person singular and third person singular or first person plural present tense verb forms; *je* as an alternative third person singular form of *być* ‘to be’.

4. Word formation: presence of the *-ko* suffix to denote diminutives, nouns being bearers of features of other nouns or adjectives, nouns denoting a person carrying out some action; the diminutive suffix *-yk* as an alternative to *-ek*; the suffix *-czuk* to denote diminutives and demonyms; the suffix *-aka* to denote a person carrying out an action, in Ukrainian borrowings; suffixes *-yło, -ajło* in verb-derived and other nouns; popularity of suffixes *-usio, -osia, -unio, -unia, -cio, -cia*, sometimes preceeded by the infix *-uń-* as diminutives of varying degrees; rare use of Ukrainian suffixes *-aga, -yga, -acha, -un*; the suffix *-ka* attached to foreign loadwords; *-ny* or *-nny* as suffixes forming adjectives; suffixes *-en* and *-szy* used in pronouns; tendency to derive nouns from verbs with *-ość* as a suffix; tendency to use suffixes like *-ka, -enie, -anie, -ęcie* to denote actions that already have other established names; use of different prefixes in verbs to denote an already established meaning.

5. Syntax: the disappearance of the masculine/non-masculine distinction in plural past tense verbs; use of the accusative in constructions where usually genitive is used; ACI; infinitives in final clauses; constructions with *wziąć* ‘to take’ resembling an auxiliary; nonstandard use of prepositions; use of the active participle adjective as a participle clause; negation as an intensifier in sentences; multifunctionality of the conjunction *że* ‘that, because’.

What can be noticed is that while there are parallels between the development of Polish and its eastern dialects, there are some significant differences as well, particularly when it comes to the phonology, sometimes resulting in changes to morphology, alongside some noticeable syntactic differences. From the perspective of this thesis, it is important to be aware both of the changes that characterized the time period in which the memoir in discussion was written, as well as the dialect typical for its area of origin, as both historical and dialectical features may appear to be present in the data in question. However, certain areas are more likely to be reflected in the following experiments; therefore, sections pertaining to orthography (and phonology, if the pronunciation is reflected in spelling), lexicon, and syntax are perhaps the most relevant. For instance, as shown in section 4, the /a/ to /ɛ/ change, the /r̪/ to /rz̪/ change, or competing spelling standards are recognizable in the results.

2.3 Computational approaches to language variation

As mentioned in the introduction, language variation is a relevant topic within the field of NLP, as it is ever-present and can, from the practical perspective, lead to the decrease in performance of some tools (Dorn, 2019; Zampieri et al., 2020). Additionally, language variation is inherent to human languages, and therefore being able to model and process it appears to be a natural challenge for the field. Research tackling language variation from the NLP perspective can be divided into historical and modern approaches. The former area focuses on various challenges centered around diachronic variation or other kinds of variation in historical texts, while the latter is concerned with various types of synchronic variation, including factors such as gender, age, geographical area, etc., and more discussion of dialectical variation is present in the context of synchronic studies. The discussion of historical variation is somewhat more relevant to this thesis, but some of the studies presented below tackle both issues or continue to be relevant to the topic. Simultaneously, from a more linguistic perspective, various computational methods have been and continue to be utilized in a variety of ways in research in the field of historical linguistics. While the field itself is quite old, when it comes to utilizing computational tools it is more relevant to look at more recent resources, as they are more likely to reflect the current state of affairs in the field.

While some of the following subsections have served as a direct inspiration for the experiments conducted as a part of this thesis, other ones are mentioned to present the sheer variety of different approaches to investigating and utilizing historical and dialectical data using computational methods, and to inform potential future work.

2.3.1 Computational Historical Linguistics and Corpus Linguistics

One prominent and conventional way of utilizing computational methods in historical linguistics, which is simultaneously highly relevant to this thesis, is for quantitative, corpus-based inquiries. As McGillivray & Jenset (2023) note, corpus-based research and quantitative research in this field have been on the rise in recent years. Jenset & McGillivray (2017) also present a framework for working with historical data quantitatively. They highlight that these methods are often pioneered by people interested in the new technology that is available and in ways to apply it to historical linguistics. It is therefore difficult to define the field of computational historical linguistics as anything other than the overlap between the methods and tools utilized in a variety of computational fields and historical data. Jenset & McGillivray (2017) argue strongly in favor of well-annotated corpora, highlighting the importance of not only thorough content annotation but also the inclu-

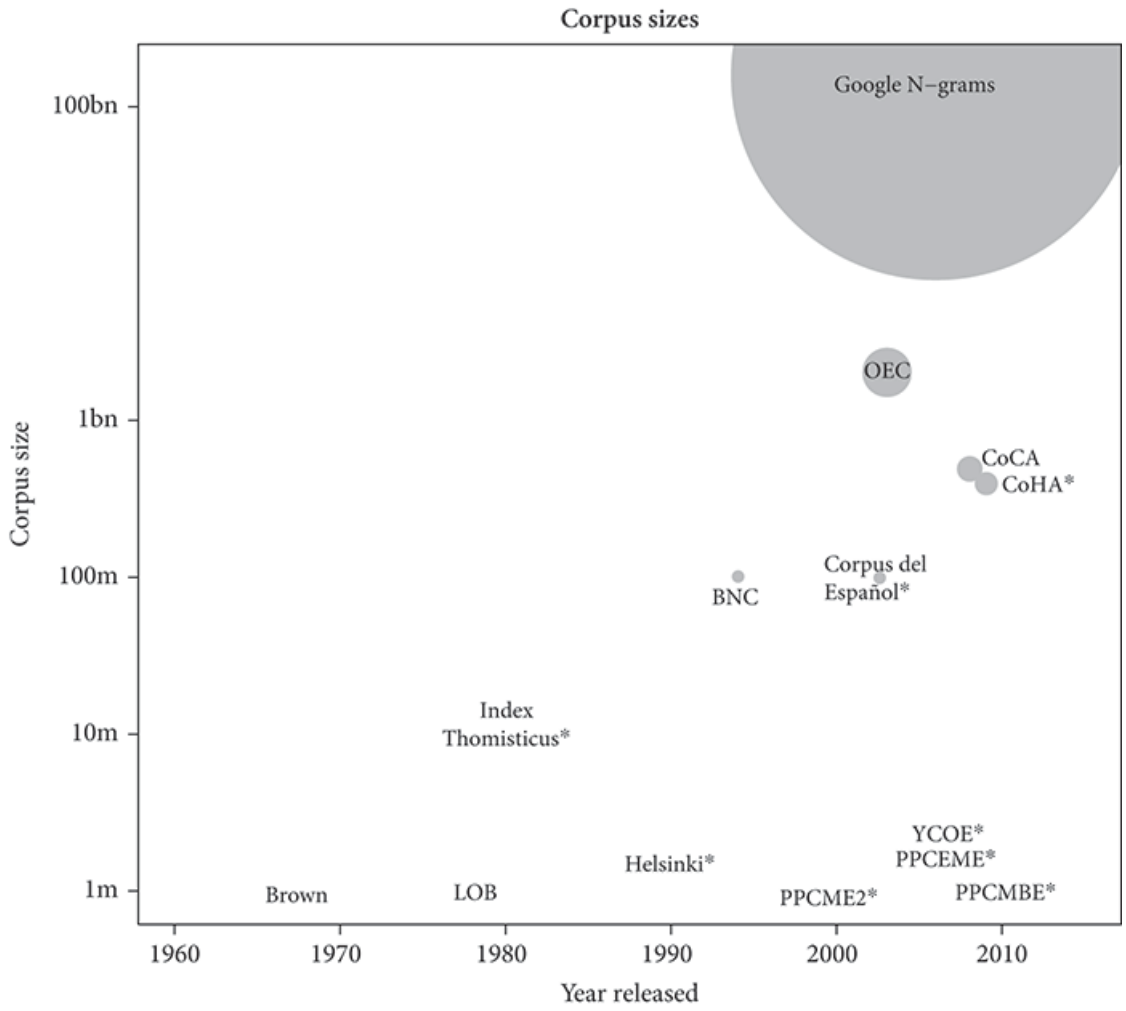


Figure 1: Selected historical and modern, predominantly English corpora by year and size, image from Jenset & McGillivray (2017).

sion of metadata. They note that in the case of manual annotation, strict guidelines should be enforced, and mention automated annotation as a promising field — although historical data does present various issues for such an approach, especially when the tools used are not well-adapted. Figure 1 illustrates the size disparity between selected, mostly English, modern and historical corpora, highlighting the possible need for more historical data in this format. A more extensive historical corpus can be, for example used to automatically detect and track collocations and their development, as well as other changes (Garcia & García Salido, 2019).

2.3.2 Syntactic variation

Another interesting computational approach, implemented by Johannsen et al. (2015), pertains to assessing variation in syntax. The authors gather data from speakers from varying backgrounds, tag it using state-of-the-art dependency parsers and POS taggers, and extract subtrees representing relations between different POS tags present in the data. The authors conduct an analysis of a selection of most prominent relations and compare the differences between groups of speakers based on e.g. their age and gender. While the authors of this paper focus on the variation between users of a language based on age and gender, this approach could likely be implemented for diachronic studies as well. According to the authors, this method allows for a larger amount of

data to be analyzed than traditional sociolinguistic methods. A simplified version of this method, using only part-of-speech tags, is utilized in subsection 3.8; however, that simplification removes the possibility for a comparison of the results to those from the paper. It also makes it impossible to observe long-distance dependency relations.

2.3.3 Part-of-speech tagging of historical data

One more area that could be said to be balancing between NLP and (computational) historical linguistics is part-of-speech tagging of historical data. Research of this kind is conducted for a variety of reasons: on the one hand, evaluating the way in which taggers adapted to modern data perform on historical data can help improve the tools and pre-processing procedures themselves, and on the other hand, the ability to accurately tag data using automated tools, as mentioned in subsubsection 2.3.1, is highly relevant for the creation of corpora.

Paper	**Language**	**Modern Text Accuracy (%)**	**Historical Test Data Accuracy (%)**	**Preprocessed Historical Test Data Accuracy (%)**
Rayson et al. (2007)	English	96	82–88.5%	89–93.2%
Scheible et al. (2011)	German	-	69.6%	79.7%
Bollmann (2013)	German	-	23–81.8%	83.4–95.6%
Hupkes & Bod (2016)	Dutch	96	60%	92%
Adesam & Bouma (2016)	Swedish	94.2[6]	45%	70%

Table 1: Test results on modern, historical, and preprocessed historical data in other experiments. Note: these experiments used different kinds of taggers, tagsets, pre-processing methods, and data, which means that their results are not fully comparable.

Table 1 presents a variety of studies where various taggers, predominantly trained on modern data, were tested on historical data, with and without preprocessing. Not included in the table is Waszczuk et al. (2018), as the measures that they do not provide accuracy as a measure, but instead use precision and recall (both around 88.3% for baroque texts and 90.3% for texts from 1830–1918). They also do not utilize any preprocessing procedures. However, this study is highly relevant to the topic of the thesis as not only does it tackle Polish, but also tests one of the taggers tested in the subsequent sections.

2.3.4 Data normalization vs. variation

A number of papers concerning historical NLP and the application of modern tools to historical data highlight the improvements that the normalization of e.g. spelling or punctuation can yield to the performance of various tools and that a requirement for normalization is imposed by many available tools (Rayson et al., 2007; Scheible et al., 2011; Bollmann, 2013; Hupkes & Bod, 2016; Adesam & Bouma, 2016; Garrette & Alpert-Abrams, 2016; Estarrona et al., 2020; Hämäläinen et al., 2021). However, Dipper & Waldenberger (2017) point out that the mapping of different variants in historical data to their modern counterparts that occurs in such preprocessing can be

[6]Here the tagger was trained on historical data as well.

very informative as far as language variation and change is concerned. Using the normalized forms as a way to establish which historical word-forms are equivalent, the authors conduct a diatopic mapping of language variation in historical German. Additionally, they group and analyze the so-called "rewrite rules" for normalization by area and conduct an analysis to reveal what kinds of variation it is mitigating, such as morphological, phonological, or graphemic variation, indicating that these are the kinds of differences that can be inferred from the word-forms themselves. Findings presented by Eisenstein (2015) support the claim that orthographic variation can be motivated phonologically, while simultaneously showing that the extent to which a new variant becomes widespread can depend on e.g. the word form or a specific meaning.

2.3.5 Tool adaptation

Another solution when it comes to using existing tools for nonstandard data is adapting the tools themselves instead of normalizing the data. For instance, as far as syntactic annotation of historical data is concerned, Regnault et al. (2019) highlight that there still is a need for adapting the existing tools if high performance is to be achieved. In their research they adapt a metagrammar in order to be able to automatically annotate Old and Middle French texts. Alternatively, Sánchez-Marco et al. (2011) argue that extending a tool's dictionary and retraining it with a small corpus leads to very good results on nonstandard data, as illustrated by their experiments on historical Spanish; additionally, they conduct an error analysis of the remaining problematic tokens in order to establish the ways in which their method improved the tool's performance.

2.3.6 Modelling language change and dialectical variation

In another approach, Zampieri et al. (2016) attempt to model language change in historical Portuguese using word and POS n-grams as features for SVM classification of the source texts in terms of the date of publication; while they conclude that the latter are not as informative, they conclude that they can provide some insights in later analysis. They also note that the larger the word n-grams, the worse the performance, and that the opposite is true for POS n-grams. Looking at lexical variation, Peirsman et al. (2010) attempt to use computational methods to retrieve cross-lectal synonyms and identify lectal markers with Belgian and Netherlandic Dutch corpora and resources as data; while their methods are successful at identifying some differences, they are still hindered by polysemy and the colloquial status of some of the words. In an investigation into the modern dialects of Spain, Donoso & Sánchez (2017) approximate the difference in dialects using cosine similarity and the Jensen-Shannon metric; however, they still rely on a lexical database to select which concepts to target in their comparison. While Hovy & Purschke (2018) utilize Doc2vec to obtain document representations for German and, together with geographical information, form clusters corresponding to dialect areas, this method does not appear to report what kind of variation is identified.

2.4 Computational background

Although their use in this thesis is discussed in more detail in section 3, relevant algorithms and architectures are introduced in this subsection. It is important to note that this thesis does not attempt to improve any of these, but instead tests how they can be utilized in similar inquiries.

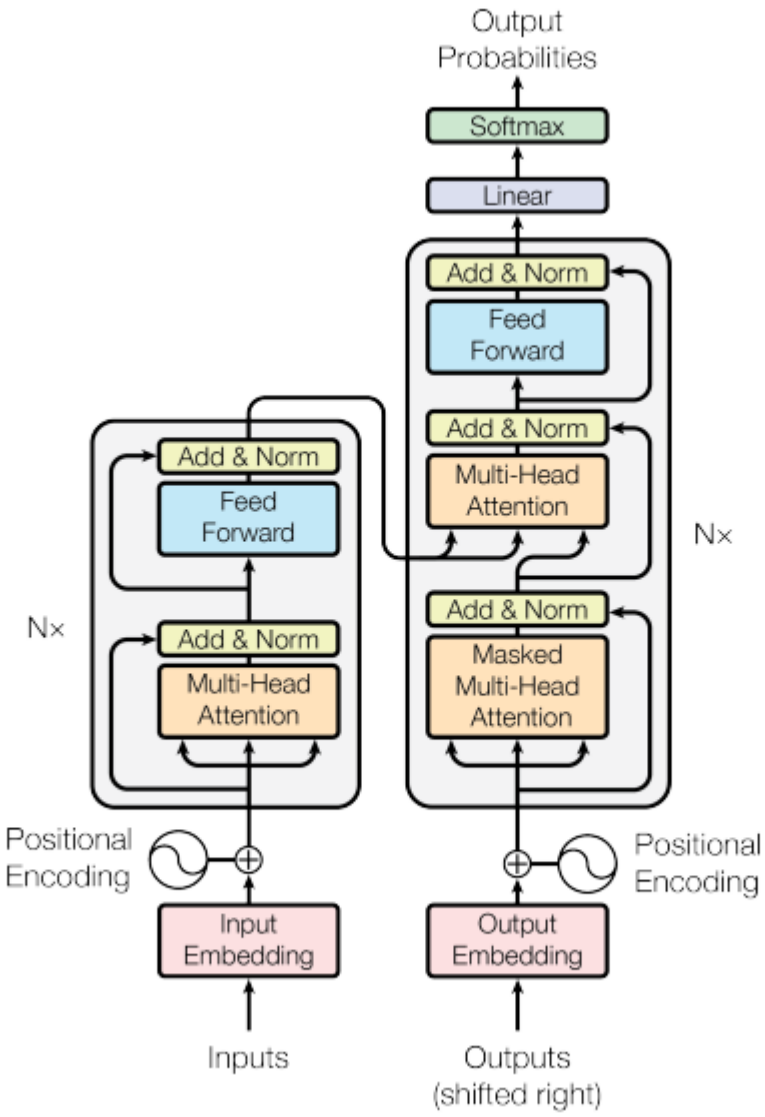


Figure 2: The architecture of a transformer model, image from Vaswani et al. (2017).

Within this thesis, a selection of pre-trained tools is used. These include Morfeusz2 and Concraft-pl, as described in Waszczuk (2012) and Kieraś & Woliński (2017), the Stanza neural pipeline, outlined in Qi et al. (2020), and University of Sheffield's UPOS tagger (The University of Sheffield, nd). In addition, two pre-existing tagger architectures are trained on selected data. These include a pipeline provided by Wolf et al. (2020) for utilizing any of the HuggingFace BERT-like models (in the case of this thesis it was BERT for Polish) (Kłeczek, 2021). These harness the power of the so-called transformer neural model architecture (as depicted in Figure 2), first introduced by Vaswani et al. (2017), which employs attention to determine which elements should weigh in on the result; this allows the model to better take into account which elements of the sentence can indicate the appropriate tag, in addition to the information from the word representation. The other trainable tagger architecture is Marmot, an improved Conditional Random Fields-based tagger, using what authors call "pruned CRFs." This approach consists of "[creating] increasingly complex lattices and to [filtering] candidate states at every step to prevent a polynomial increase in lattice size" (Mueller et al., 2013). The method for this is presented in Figure 3. It is worth noting that neither

of these algorithms has been implemented from scratch, and openly available implementations are used in this thesis, and the use of all of these tools is specified in section 3.

```
1: function GetSumLattice(sentence, τ⃗)
2:     gold-tags ← getTags(sentence)
3:     candidates ← getAllCandidates(sentence)
4:     lattice ← ZeroOrderLattice(candidates)
5:     for i = 1 → n do
6:         candidates ← lattice.prune(τ_{i−1})
7:         if gold-tags ∉ candidates then
8:             return lattice
9:         end if
10:        if i > 1 then
11:            candidates ← mergeStates(candidates)
12:        end if
13:        candidates ← addTransitions(candidates)
14:        lattice ← SequenceLattice(candidates, i)
15:    end for
16:    return lattice
17: end function
```

Figure 3: A mock-algorithm for CRF pruning, image from Mueller et al. (2013).

In addition, this thesis makes use of a number of existing libraries for Python 3 and their implementations of various algorithms and measures, such as the evaluation measure calculations provided by Pedregosa et al. (2011) or data structures from The pandas development team (2020).

3 Experimental Setup

This section contains a description of both the data and the experiments that were conducted. The entirety of the code, alongside unannotated and annotated data, is provided in Appendix A. The results of the experiments are presented and discussed in section 4. A description of the data and the annotation process is provided in subsection 3.1 and subsection 3.2. Generally, the experiments can be divided into three categories: tagger and lemmatizer testing (subsection 3.3, subsection 3.4, subsection 3.5, subsection 3.6, subsection 3.7, with the error analysis detailed in subsection 3.10), n-gram statistics (subsection 3.8), and investigations into the National Corpus of Polish (subsection 3.9).

3.1 Data

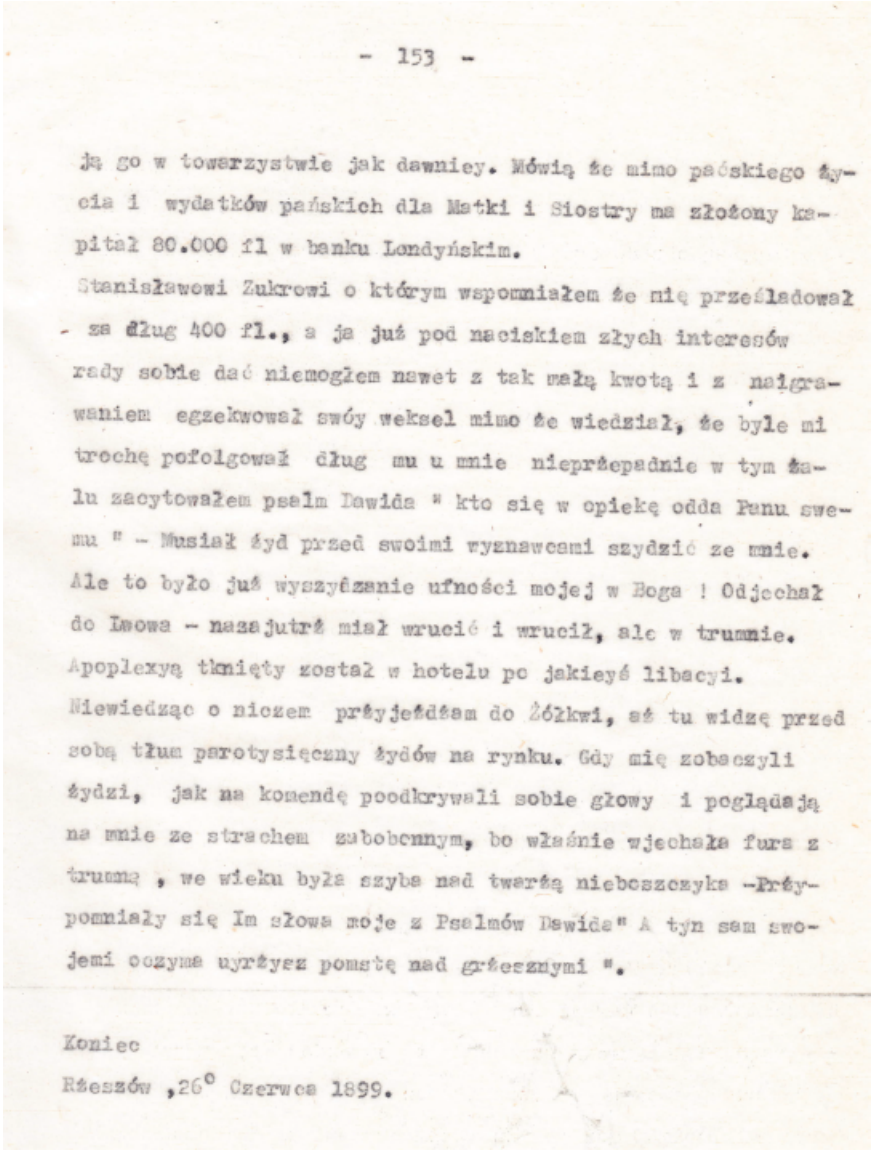
The data used in the experiments originates from a memoir penned by Juliusz Czermiński in 1899 in Rzeszów. The original manuscript is preserved in the collection of Zakład Narodowy im. Ossolińskich (also known as Ossolineum) with the signature 15374/II, according to the library's catalog, but cannot be accessed digitally (Ossolineum, nd). At some point in the past, typewriter copies of the manuscript (possibly made from another copy and not the original manuscript) have been made and distributed among the author's descendants. In recent years, one of them, Piotr Kociatkiewicz, undertook the effort of copying over the text into a Word file, and it is this digitalized data that was used throughout the thesis. Unfortunately, due to the time constraints and the physical difficulty of accessing the manuscripts, no assessment of the quality of the transcription could be made.

As mentioned before, the data originates from one author and belongs to the genre of memoir. The author was a native of an area that encompasses nowadays south-eastern Poland and western Ukraine but was not independent at that time, and instead a part of the Austro-Hungarian Empire following the partitions of the Polish–Lithuanian Commonwealth in the late 18th century. From what can be gathered from the contents of the memoir, the author considered himself to be Polish and wrote in an idiolect closely resembling the Polish language. However, due to the text's age and region of origin, it is likely that it diverges from standard modern Polish with regard to spelling (from which pronunciation may be inferred), grammar, and vocabulary. This assumption is strengthened by the fact that following the periodization of the history of Polish as outlined by Długosz-Kurczabowa & Dubisz (2006), the text could be classified as an example of writing in "early" New Polish (npol. 1.), which diverges from Modern or "late" New Polish (npol. 2.).

Both the relative understandability of the text to a native speaker of Polish and the potential for it to differ from standard Polish make it a good candidate for inquiries into how potential differences between a historical text and a modern standard could be identified computationally. It is, nevertheless, worth keeping in mind that this text need not be representative of Polish in general at the time of its writing and this thesis should be regarded as an investigation of this particular memoir and its author's language, not of late 19th century Polish at large.

The entirety of the memoir consists of 37 405 tokens, according to Microsoft Word's word count functionality. Out of those, the first 360 sentences, corresponding to 10 286 tokens, were manually annotated with UPOS (universal part of speech) tags, and the first 115 sentences, corresponding to 3271 tokens, were additionally annotated with XPOS (language-specific part of speech) tags

and lemmas following the tagsets used by Wróblewska (2018) in PDB (Polish Dependency Bank), which itself is the largest UD corpus available for Polish, and therefore conforms to the UD format. This decision was made due to the accessibility and universality of these tagsets, and because the results could then be compared to PDB's test set; additionally, some other tools, while not trained on PDB, seemed to be using the same tagset. The details of the annotation process are discussed in subsection 3.2. While this means that only roughly more than a quarter of the text was annotated with the UPOS tags and less than a tenth with the XPOS tags and lemmas, the annotation of the entire text was deemed to be beyond the scope of this thesis, especially given the complicated nature of the XPOS tags and the fact that the annotation had to be of high quality. Additionally, small test samples are not unheard of when it comes to tagger-related experiments using historical data (Bollmann, 2013; Hupkes & Bod, 2016; Rayson et al., 2007).



- 153 -

ją go w towarzystwie jak dawniey. Mówią że mimo pańskiego ży-
cia i wydatków pańskich dla Matki i Siostry ma złożony ka-
pitał 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 naigra-
waniem egzekwował swóy weksel mimo że wiedział, że byle mi
trochę pofolgował dług mu u mnie nieprzepadnie w tym ża-
lu zacytowałem psalm Dawida " kto się w opiekę odda Panu swe-
mu " - 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 fura z
trumną , we wieku była szyba nad twarzą nieboszczyka -Przy-
pomniały się Im słowa moje z Psalmów Dawida" A tyn sam swo-
jemi oczyma wyrżysz pomstę nad grzesznymi ".

Koniec
Rzeszów ,26° Czerwca 1899.

Figure 4: The final page of the typewriter transcription of the original text of the memoir, from which the digital copy was made. Private collection. Image courtesy of Anna Chodorowska.

Simultaneously, most of the procedures described in the subsequent subsections were also conducted on the PDB (Polish Dependency Bank) corpus, so that the results could be compared to those obtained on modern data; additionally, some of the taggers were locally trained on the PDB

train set. As mentioned before, PDB is the largest UD-style treebank available for Polish. It features 22 152 sentences consisting of a total of 347 377 tokens, annotated according to the UD guidelines in the CoNLL-U format. Aside from the lemmas, UPOS, and XPOS tags, which are utilized in this project, the treebank also features an annotation of syntactic relations between words (Wróblewska, 2018; Universal Dependencies, ndc).

Another corpus utilized in this thesis is the National Corpus of Polish, a large collection of Polish texts that the authors claim to be balanced and representative of the language. While the entirety of the corpus is not available to be downloaded due to copyright issues, there do exist search engines for it, and a small subcorpus is available for downloading. This subcorpus is manually annotated with a tagset closely resembling the UD XPOS tags (Przepiórkowski et al., 2012). Within this project, only one of the search engines is directly utilized, although other tools may rely on data from this corpus (Pęzik, 2012).

In the initial stages of the project, there was an idea to compare the results obtained from the discussed data with results from running the same experiments on a subset of the Korba Corpus, also known as The Electronic Corpus of 17th and 18th c. Polish Texts (up to 1772) (Gruszczyński et al., 2020). Although code allowing for the extraction of the desired data from the corpus files was developed for the needs of this thesis, it was later discovered that not only does the corpus not include UPOS tags, but its XPOS-like tags differ in small but relevant ways from the ones used in PDB. Finding a way to unify these tagsets was deemed to be beyond the scope of this thesis and the Korba Corpus was not used in later experiments.

3.2 Data Annotation

The process of data annotation occurred in a number of steps. First, the data was converted from a `.docx` file to a `.txt` file and segmented so that every line corresponded to a paragraph or a section in the original text. This served as a basis for the first major step in the annotation, namely the manual annotation of a selected subsection of the text with UPOS tags. Subsequently, Python code in the form of a Jupyter Notebook that allowed for the pre-tagging using the Morfeusz morphological analysis tool (Kieraś & Woliński, 2017) in tandem with Concraft-pl (Waszczuk, 2012; Waszczuk et al., 2018), a morphosyntactic tagger which relies on Morfeusz's analyses was developed (these two tools are discussed in more detail in subsection 3.6). This was used for pre-annotating the subset of the data that was intended to be annotated with XPOS tags and lemmas, as those were the types of annotation provided by Morfeusz and Concraft-pl. The results, along with the UPOS tags, were outputted into a `.conllu` file which adhered to the standards of that format. This pre-annotation was then manually reviewed and corrected wherever necessary.

As mentioned in subsection 3.1, the tagset used for this task was the same as the one used in the Polish Dependency Bank, the largest of the UD-standard treebanks for Polish (Wróblewska, 2018). That was also the corpus that was consulted in problematic cases; whenever necessary, an online dictionary of the Polish language was consulted as well (PWN, nd).

Each type of tagging (lemma, UPOS, XPOS) was characterized by its own difficulties. When it comes to manual lemmatization — the task that would appear to be the easiest, at least to a native speaker — the issue was deciding what lemmas to enter for words that were spelled in an unconventional way. A number of the words in the text were spelled together in ways that are not permitted by standard Polish; other words were simply spelled using a different spelling

convention or in a way that possibly reflected pronunciation. A decision was made to preserve these peculiarities, while simultaneously trying to present the word in its base form, in an attempt to infer the idiolectal base form. For example, *oyca* '(of the) father' was lemmatized to *oyciec* instead of *ojciec* 'father', which would have been the modern spelling. This was done in order to preserve the original spelling of the words and reflect how the author would have likely written the base form of the word. In addition, in one of the experiments which consisted of comparing the vocabulary of the text with that of a modern Polish corpus (discussed in more detail in subsection 3.9), preserving the original spelling was essential, as one of the goals of the comparison was to determine if words and lemmas with that spelling occur in the corpus. It is, nevertheless, important to note that this decision may have negatively impacted the lemmatization performance of some tools during evaluation.

UPOS tags, which not only reflect the approximate word class but sometimes also the role of a word in the sentence, have proven to generally be rather straightforward to assign. Nevertheless, there were some instances of words that could be classified as more than one class without a straightforward way to differentiate between those two supposed meanings. One such example is the word *około* 'around,' which could be classified as either an adposition or a particle in the treebanks - and for which the Dictionary of the Polish Language provided two practically identical definitions, that did not allow for an easy distinction between the two (PWN Editorial Team, nd). Another problematic category was the rule that verbs normally treated as auxiliaries should be classified as regular verbs in purely existential sentences (Universal Dependencies, ndb).

Finally, the XPOS tags required the most attention during the review of the preannotation, predominantly due to the fact that oftentimes they include a lot of information about the features of the token, such as gender, number, aspect, etc. Consequently, there was not much room left for token-level ambiguity, but issues stemming from syntactic ambiguity persisted. Another major issue throughout the annotation process was determining whether a verb-derived word should be classified as a gerund/participle or as a noun or adjective. For example, *bombardowanie* 'bombing' could be treated either as a noun or as a gerund of the verb *bombardować* 'to bomb.' If the word was attested for in PDB, it was tagged in the same fashion as in the corpus. Otherwise, the decisive factor was the presence of the derived form as an independent word in a dictionary.

3.3 Experiment 1: BERT XPOS and UPOS-tagging

The first experiment consisted of fine-tuning a BERT model for a token classification task. Being able to fine-tune one's own model was beneficial, as one was in full control of the data and tagset used in the process. However, that cannot be said for the data utilized in the training of the original BERT model.

The fine-tuning and evaluation were conducted using the code and instructions provided in the Transformers library for Python in `transformers/examples/legacy/token-classification/`, with minor changes meant to adapt the procedure to the provided data (Wolf et al., 2020). Preprocessing of both the training, evaluation, and testing data was modified to include another script, `preproc_bert.py`, which removed the lines required by the CoNLL-U format for Polish for non-split tokens (i.e. situations where an element that the UD requires to be described separately is attached to another word, for instance *słyszałem* 'I heard' is required to be split into *słyszał*, the l-participle of the verb meaning 'to hear' and *em*, the "mobile inflection" indicating the person). These lines do not feature any annotation but indicate the range

of original word. Therefore, they were irrelevant for the tagging, and could actually be disruptive if left in the text. No other major changes were made to the settings of the fine-tuning, as the goal of this experiment was not to find the best hyperparameters for the task, and the suggested hyperparameters were assumed to be acceptable.

The model used as a basis for the fine-tuning was `bert-base-polish-cased-v1` by Kłeczek (2021). While both the cased and uncased versions of the model perform well on different evaluation tasks, according to the author the cased model features improvements over the uncased one. Additionally, due to the historical data featuring unconventional capitalization and many proper names, it was deemed relevant to maintain the capitalization.

A total of two models were fine-tuned, one for UPOS-tagging, and one for XPOS-tagging. They were trained, evaluated, and tested on the PDB data. Subsequently, both of the models were tested on the historical data, which was pre-processed in the same fashion as the PDB data. The results were automatically saved in `.txt` files. Although this process did output a selection of evaluation measures, for the sake of comparability, those were recalculated in a Jupyter Notebook file using functions from `functions.py`, a Python file containing functions used across several different experiments, both for the modern and historical test set, with the evaluation measures' implementation from Scikit-learn and various pandas elements (Pedregosa et al., 2011; The pandas development team, 2020; McKinney, 2010). A number of `.xlsx` files containing all of the annotations and only the erroneous ones were created for later analysis.

3.4 Experiment 2: Marmot XPOS and UPOS-tagging

The next experiment similarly consisted of training a tagger architecture on PDB data. In this case, the tagger was a CRF-based framework called Marmot (Mueller et al., 2013). Although Marmot does have pre-trained models for Polish, their tagsets did not appear to be compatible with the one used in this thesis. Therefore, a new model was trained on the PDB training set, and tested on both the PDB test set and the historical data. Just as in subsection 3.3, this data had to be preprocessed using `preproc_bert.py`. Marmot can be trained to tag both UPOS and XPOS simultaneously, so only one model was trained.

Marmot does not output any evaluation measures, so the results were imported into a Jupyter Notebook and the necessary measures were calculated there. Same as before, the results were also output in the form of two `.xlsx` files, one for all the results and one including just the mistakes made by the tagger.

3.5 Experiment 3: Stanza XPOS-tagging, UPOS-tagging, and lemmatization

Another tagging service that was used to annotate the historical data was that provided by Stanza (Manning et al., 2014; Qi et al., 2020). Stanza's neural pipeline provides all three desired functionalities: lemmatization, XPOS annotation, and UPOS annotation. The default package for the Polish language in Stanza is based on PDB, which was extremely convenient as the tagsets were certain to match if no changes were introduced while constructing the package. In order to obtain the annotations, the Stanza pipeline was run in a Jupyter Notebook environment on both the modern and historical data. Measures compatible with those from other experiments were output for every category, and `.xlsx` files containing all of the annotation and the errors were produced for

later comparison. In addition, measures and results were produced for the comparison of lowercased gold standard and lowercased output of lemmatization, as it has been observed that Stanza returns all the lemmas in lowercase.

3.6 Experiment 4: Morfeusz XPOS-tagging and lemmatization

Unlike the previously tested taggers, this one relies on two complementary tools. Morfeusz is a morphological analyzer, which provides both the possible morphological analyses of input tokens and their lemmas. This is done on the basis of provided linguistic data. In the case of this experiment, and for the sake of compatibility with the other tool, the SGJP data, based on a Grammatical Dictionary of Polish was selected (Saloni et al., 2015; Kieraś & Woliński, 2017). In order to disambiguate Morfeusz's predictions, another tool, Concraft-pl is used. Similarly to Marmot in subsection 3.4, it is a CRF-based tool. The pre-trained model that is provided and that is compatible with the aforementioned version of Morfeusz has been trained on data from the National Corpus of Polish (Przepiórkowski et al., 2012; Waszczuk, 2012; Waszczuk et al., 2018). While these tools have not been based on or trained on PDB, the tagset that their use appears to be compatible with the XPOS tagset of PDB. This divergence in terms of source data is naturally something that should be taken into account when comparing the results from different taggers. Nevertheless, the matching tagsets make these tools a relevant addition to the project.

In this experiment, the pipeline designed for the pre-annotation of the historical data using Morfeusz and Concraft-pl, organized within a Jupyter Notebook file, was modified to obtain the predictions based on the input from the `.conllu`, not `.txt` file, as before. Naturally, the desired output format was also changed, as the results of the tagger experiments were not saved in `.conllu` files. Since this combination of tools does not have the option to introduce custom tokenization, an algorithm matching misparsed tokens to their gold standard counterparts had to be developed. Once again, as far as lemmatization was concerned, measures were obtained both with the original capitalization and with both the gold standard and the predictions being lowercased, so that a comparison between different tools' performance when it comes to lemmatization could be made, as the other one investigated within this thesis has proven to return results solely in lowercase. Similarly to the experiments with other taggers, the list of all tokens and annotations was saved for both XPOS tags and lemmas in the form of an `.xlsx` file, along with similar spreadsheets containing only the errors that the tools have made.

3.7 Experiment 5: UD Cloud UPOS-tagging

The final tagger used in this project is a maximum entropy model based on the GATE Learning Framework plugin hosted by the University of Sheffield (The University of Sheffield, nd). This tagger was trained on the Polish corpora available in UD, including PDB. This tagger, called the "UD tagger" in this thesis, only provides UPOS tags for the input tokens. One of its severe limitations is the daily quota that makes it impossible for large amounts of data to be processed at once, which meant that the PDB test data had to be tagged in batches. Despite these shortcomings, the tagger was still deemed worthy of inclusion, as it is also based, at least in part, on PDB, and provides tagging that is compatible with the tagset used in this project, and simultaneously the architecture of this tagger differs from the other ones used.

The tagging process was conducted via the provided API, from a Jupyter Notebook file. As men-

tioned before, the way the data was fed to the tagger was dictated by the practical constraints of a daily limit imposed on the tagger. Therefore batches of sentences were fed to it, and, in the case of the PDB test set, the whole set had to be split for the tagging process. Analogously to the case of Morfeusz and Concraft-pl (see subsection 3.6), this tagger tokenizes the text on its own, meaning that some elements may end up misparsed in comparison to the manual tokenization, and a similar algorithm was used to single out these elements and assure that the output list from the tokenizer was of the same length as the input, with 1:1 correspondence between the elements. Same as with the other taggers, both the output and only the erroneous tags were saved to `.xlsx` files.

3.8 Experiment 6: n-gram statistics

Following Johannsen et al. (2015)'s method of approximating a language user's syntax by means of obtaining "treelet" statistics, i.e. statistics of subtrees in the UD-style dependency relation trees, an attempt was made at developing a similar method, albeit perhaps less informative when it comes to long-distance relations. While Johannessen et al. (2020) did use automatic pre-annotation, Regnault et al. (2019) suggest that in case of historical data such tools might need adapting. Due to a manual revision of pre-annotation using non-adapted tools being deemed too time-consuming for this project, the closest subunits that could be obtained for the data were n-grams over the XPOS and UPOS tags. As previously mentioned, this method may hopefully reveal insights into close-distance dependencies or trends in word order but fails at taking into account long-distance relations. Nevertheless, differences between the sentences in PDB's test set and the historical data in question in terms of these statistics could hint at larger syntactic differences.

The statistics were obtained from the manually annotated `.conllu` files and the PDB test set with Python code within a Jupyter Notebook. N-grams were constructed with functions using NLTK tools and displayed and saved to `.xlsx` files using pandas (Bird et al., 2009; The pandas development team, 2020; McKinney, 2010). For both kinds of data and POS annotation unigrams, bigrams, and trigrams were constructed and counted, with both the raw numbers and the relative frequencies provided. The relative frequencies of given n-grams were also compared side by side for the modern and historical data.

3.9 Experiment 7: National Corpus of Polish vocabulary comparison

The final investigation into the historical data itself consisted of assessing in what ways its vocabulary differs from that of modern Polish. To this end, an API access to the National Corpus of Polish was used (Pęzik, 2012). Access to the most recent version of the API was provided upon request by Dr. Pęzik.

Vocabulary comparisons were conducted in a Jupyter Notebook file for both PDB's test set and the historical data, on tokens as well as lemmas. All of the search settings were kept except for a narrowing of the search range in terms of time, balance, and limit. The limit variable determined how many detailed hits were returned out of all the matches. This information was not very relevant, as the goal of the search was only to determine whether or not the word occurs in a given subsection of the corpus. Therefore this was reduced from the default setting of 20 to 1. The balance setting determines whether the search is conducted in a balanced subcorpus or across all the texts. Since the search was intended to verify whether a word is attested for in the corpus, it was found that a larger selection of text would be preferable, despite it being unbalanced. The time

range for the source texts was narrowed down to 1945 until 2023. While the memoir was written at a time that is considered to be the tipping point between the first and second period of New Polish, the experiments conducted within this thesis were intended to compare it with modern Polish (Długosz-Kurczabowa & Dubisz, 2006). Therefore, some other, more recent cutoff point had to be selected, keeping in mind the fact that the selection of texts in the corpus should still be representative of all of the language. The decision was made for 1945 to be that cutoff point, as that date marked the end of the Second World War and the most recent change to Poland's borders.[7]

The results were retrieved both for tokens in their original form and for the lemmas; in the latter case, a special search syntax was used, where `lemma**` is intended to return all the possible inflectional forms for the word. Raw numbers and a relative proportion of the words not found in the corpus were obtained, and all of the unique tokens or lemmas for the historical data were saved in `.xlsx` files alongside their counts in the aforementioned subcorpus. The code also printed out a list of the items that were not found in the National Corpus of Polish. Finally, a qualitative inspection of the items that were not found in the corpus was conducted, alongside a comparison of the historical ones with a lexicon of words typical for Borderlands Polish provided in Kurzowa (1983).

3.10 Tagging and lemmatization error annotation

Since a major part of the experiments consisted of testing various taggers on modern and historical data, it was decided that a joint error analysis should be conducted. Since the goal of the thesis is not to verify which tagger is the best or what kind of errors particular types of taggers tend to make, this kind of analysis was considered warranted, as it would reveal which tokens were consistently problematic across the different tools, and not what types of errors a particular tagger tended to struggle with.

The error analysis was conducted in three steps. First, a Jupyter Notebook file was created in which the lists of the results for different lemmatizers (Stanza, Morfeusz), XPOS taggers (Stanza, Morfeusz, BERT, Marmot), and UPOS taggers (Stanza, BERT, Marmot, UD) were combined and tokens for which at least two tools provided a wrong tag were extracted. For the lemmatizers that meant that all of the tools had to make a mistake while tagging the data, while for the POS taggers only half of them had to be wrong. This eliminated errors that stemmed purely from a certain tagger's tendency to misclassify some tokens unless that tendency was shared by more than one tool. Those errors were then saved into three separate `.xlsx` files. In the second stage, the errors were manually annotated depending on the subjectively perceived possible cause of the error. Each category (lemmatization, XPOS, UPOS) received a slightly different set of error types due to apparent differences in what tokens were problematic, but a number of error types persisted cross-categorically. Finally, annotated `.xlsx` files were loaded into a second Jupyter Notebook, and counts and relative frequencies for each error type were calculated.

[7] Alternatively, 1989, the year that Poland's communist government fell and the country underwent a transformation, was considered, but was deemed too recent

4 Results and Discussion

4.1 Lemmatization

As outlined in subsection 3.5 and subsection 3.6, the two lemmatization tools that were used in this thesis were Stanza and Morfeusz. The only evaluation measure that was obtained for lemmatization was accuracy, as lemmatization differs from other classification problems in terms of the sheer number of possible classes, and therefore other measures were considered superfluous. Additionally, the measures, results, and error lists were returned for the original gold standard and predictions as well as for lowercased gold standard and predictions. This was done because Stanza appeared to return only lowercase predictions, and it was deemed interesting to compare how lowercasing would affect the lemmatization performance and variation detection. Table 2 depicts the accuracy per model and type of test data. What is immediately visible is Morfeusz's better performance on both modern and historical data. As far as the original results and gold standard are concerned, many of the mismatches between the gold standard lemma and the assigned lemma in the case of Stanza were due to Stanza returning all lemmas in lowercase by default. Lowercasing both of the lists (not the input tokens, only the predictions and the standard) has proven to increase the performance of both of the tools, on both kinds of input (modern and historical), although to a slightly different extent. As far as the historical data is concerned, both of the lemmatizers improve by around 3 percentage points, meaning that the lowercase output is not the only cause of Stanza's inferior lemmatization performance.

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

Table 2: Lemmatization accuracy per model and per test data type.

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

Table 3: General types of errors and their raw and relative frequencies among the historical tokens that were mislabelled by both Stanza and Morfeusz for the unaltered output.

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

Table 4: Detailed types of errors and their raw and relative frequencies among the historical tokens that were mislabelled by both Stanza and Morfeusz for the unaltered output.

Another noticeable difference is the significantly worse tagging performance of the tools on the historical data when compared to the modern data. The qualitative error analysis conducted on tokens that were mislabelled by both of the lemmatizers[8] revealed characteristics of the mislabelled tokens that could be identified by a human annotator. The error statistics can be seen in Table 3 and Table 4 for the unaltered data and in Table 5 and Table 6 for the lowercased outputs and gold standards, while explanations and examples of the specific error types can be found in Table 19 and Table 20 in Appendix B. A number of general categories can be distinguished from among the errors. The most frequent one, **spelling**, encompasses *y*, *e*, *nie*, capitalization, other spelling peculiarities. The second most prominent category is that of **name**, which includes surnames, given names, and other proper names. These two categories appear to be rather text-specific, as the potentially unique vocabulary and a specific non-standard way of spelling are not features of the Polish language in general, but of the writing of this particular author. While there are other general (and detailed) categories, they account for a much smaller selection of the errors present in the text, and the bulk of the issues appear to be connected to the text's peculiarities. Simultaneously, the problematic tokens hint at there being a need for a more uniform way of determining what lemma to choose for verb-derived nouns and adjectives or words that have more than one acceptable spelling. What is worth pointing out is that while certain spelling decisions made by the author could be explained by there having been various competing spelling conventions, the substitution of *a* for *e* in a number of instances (perhaps more prominent in subsection 4.2) appears to be characteristic of the Borderlands dialects of Polish, as described in subsection 2.2, proving that the variation in

[8]This allowed for the elimination of errors caused by tagger-specific issues and made it possible to focus on instances where it was more likely that it was the token itself that was problematic; for instance, this eliminated nearly all of the instances where Stanza returned a lowercase lemma where it was not expected to do that.

the text is not only diachronic but likely also regional. It is also worth noting that some rare kinds of errors can be quite interesting too, as is the case with the one instance of the **grammar** error. In this case an unusual inflectional form *człowiecze* 'man' is selected for the vocative of *człowiek*, with the modern form being *człowieku*[9]. This difference is not listed in any of the previously discussed sources for historical Polish for the relevant time period, nor does it appear to be strictly dialectical; the form is attested for in the National Corpus of Polish, though.

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

Table 5: General types of errors and their raw and relative frequencies among the historical tokens that were mislabelled by both Stanza and Morfeusz for the lowercased output.

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

Table 6: Detailed types of errors and their raw and relative frequencies among the historical tokens that were mislabelled by both Stanza and Morfeusz for the lowercased output.

While comparing the results for the original and lowercased comparisons, one can note that the improvement in tool performance is due to some errors from the categories of spelling, name, and vocabulary disappearing. In the first and the last category it appears that capitalization still had some role to play — a foreign word or an unusually spelled word were capitalized and therefore one of the taggers returned a capitalized lemma as well. The most major change occured in the

[9]The form *człowiecze* is perhaps considered more poetic than archaic, but it is certainly not an everyday standard form.

category of name, though, with surnames and given names being the most affected, but with some of place names also now being "correctly" lemmatized. This hints to the casing of the output in comparison with the casing of the standard being quite relevant, and having the potential to lower the apparent tagging performance. Nevertheless, the gold standard lemmas were capitalized in the case of names since that is the UD standard which was being followed. Additionally, were the lemmatized forms to be used in some downstream tasks, the removal of capitalization could have negatively impacted those, as it does carry some meaning, especially for various kinds of proper names — so while for this task lowercasing helped eliminate the instances that were not truly problematic or signified no variation other than the text including many names, overall the phenomenon of a lemmatizer only returning lowercase lemmas could be deemed disadvantageous.

While a major drawback of this method of discovering a historical text's peculiarities is that there needs to be a gold standard to compare the overall performance of the lemmatizers to, it does reveal some interesting insights into the kinds of tokens that appear to be nonstandard for Polish and typical for a given text, including a plethora of proper names and unusual spelling, as well as singular instances of unconventional inflection. What could be done in the case of texts with no gold standard is simply reviewing the entire output.

4.2 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%

Table 7: UPOS-tagging evaluation measures (accuracy, precision (macroaveraged and weighted), recall (macroaveraged and weighted)), Matthews Correlation Coefficient per model and per test data type. Although calculated, F1 is not given since it can be derived from precision and recall. Per class precision and recall can be found in Appendix C

As detailed in subsection 3.3, subsection 3.4, subsection 3.5, and subsection 3.7, the four taggers that either utilized or could be trained to utilize the UD tagset were BERT, Marmot, Stanza, and the UD Cloud tagger. Unlike in the case of lemmatization, when it comes to tagging, there is a specific number of classes in question, which allowed for the extension of the evaluation metrics from just accuracy to accuracy, weighted precision, weighted recall, and Matthews Correlation Coefficient (MCC), which are presented in Table 7. MCC is featured in this comparison as it is considered to be appropriate for multiclass classification problems with unbalanced classes, and to be superior to accuracy in such cases (Jurman et al., 2012). Since the UD tagset is not large, precision and recall were also calculated for each class for a deeper insight into which classes are the most problematic

(Universal Dependencies, ndd). These results can be found in Appendix C. Based on the general evaluation measures BERT performed best on both the modern and the historical dataset, while the UD Cloud tagger has the worst tagging performance on both of the test sets. Although not by a lot, Stanza's neural pipeline outperforms Marmot. For all of the taggers, the historical dataset achieves a consistently lower score than the modern one, indicating issues that are not specific to the taggers themselves.

Similarly as in the case of lemmatization, manual annotation of errors made by the taggers has revealed certain recurring features, as outlined in Table 8 and Table 9; this time, however, the token in question did not have to be misclassified by all of the taggers, and instead only two of them had to have made a mistake while tagging a given token, as that was deemed more likely to still remove the tagger-specific issues while preserving more information on the possible problematic features of the data than if all of the taggers had to misclassify the token. A more detailed definition of the types of errors along with examples can be found in Table 21 and Table 22 in Appendix B. It is worth pointing out that the kinds of changes that were made in the case of lemmatization, namely the lowercasing of the output and gold standard should make no difference here, and therefore that procedure was not carried out.

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

Table 8: General 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).

Similarly to the errors made during lemmatization, the errors can be divided into overarching categories, with **spelling** and **ambiguous** being the most prominent ones. Unlike in the case of lemmatization, **name** errors do not constitute a large part of all the errors. **Unidentified** errors rise in importance relative to the other task. The **ambiguous** general class, which has seen the largest increase, includes errors pertaining to ambiguous word forms or endings, as well as noun- and adjective-like words formed by derivation which could be considered e.g. a participle form of a verb and errors stemming from UD guidelines concerning VERB and AUX tags. Clearly, the word form itself (or the inflectional ending itself) should not be the only factor determining a word's class; this may hint at an additional level of difficulty in terms of unusual word order if such information is utilized by the tagger. What is worth noting in the detailed error division is the high number of **capitalization** errors. It appears that capitalization is factored in as a feature when it comes to tagging words as PROPN — and, as a result, regular words written unexpectedly with a capital letter at the start are often misclassified as such. One infrequent, but interesting type of error that surfaces in this task is that related to **impersonal** verb forms, such as *cięto* 'was being cut'. While this form is not uncommon in modern Polish, it appears to be problematic for some

of the taggers.

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

Table 9: Detailed 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).

Class-specific measures yield a deeper insight into which classes in particular are more problematic. For BERT (as can be seen in Table 25), all of the classes in the modern text, with the exception of INTJ and SYM perform very well. The two aforementioned classes could be problematic due to their scarcity. They are also not present at all in the historical data. As for the memoir, there is a noticeable drop in tagging performance for ADV, AUX, PART, PROPN, SCONJ, and X — which are also less numerous than some of the other classes. Additionally, there is possibly some confusion between AUX and PART when it comes to the classification of the token *to* 'it' or 'is'. Naturally, the numerous new proper nouns found in the historical text are also problematic, not to mention the issue of nonstandard capitalization.

Marmot (Table 26) similarly struggles with SYM and INTJ in the modern data, but it also scores relatively lower on AUX, PART, PROPN, and X on the PDB test set. When it comes to the historical data, a considerable drop in tagging performance can be noticed for the ADJ category,

alongside ADV, AUX, PART, PROPN, SCONJ, and X. Aside from the ADJ category (which is more complicated due to place name-derived adjectived being capitalized by the author and surnames being misclassified as adjectives), the same categories seem to be problematic for Marmot as for BERT, although in this case some of these issues are already noticeable on the modern test set.

Stanza's results (which can be found in Table 27) show high precision, but low recall on the problematic INTJ and SYM classes. Otherwise, the tagger performs quite well across all of the categories in the modern data, with the possible exception of relatively low recall on PART. When it comes to the historical data, a noticeable drop in tagging performance can be observed for ADJ, AUX, PROPN, SCONJ, and X, which again mostly overlaps with the categories that were problematic for the aforementioned taggers.

The UD Cloud tagger (Table 28) shows a much more varied tagging performance across the classes on both the modern and historical test data. As far as the PDB test set is concerned, it struggles with ADJ, ADV, AUX, NUM, PROPN, and X classes in particular. As for the memoir, a large drop in tagging performance can be observed for most classes. Noticeably, the tagger performs well on ADP, CCONJ, and PUNCT, and has high precision, but very low recall on DET. While the issues that this tagger has with the modern data do partly overlap with the classes that were problematic for other taggers, this tendency is not that clear in the historical data due to the overall bad performance of the tagger. However, the issues visible on the PDB test set may hint at these classes being more problematic in both modern and historical data, and perhaps the errors made while classifying those are more visible due to the classes' lower frequencies.

Although there is an overlap in terms of what features can lead to a token being misclassified, they seem to be of different importance when it comes to assigning UPOS tags compared to lemmatization. UPOS tagging does share the same issue as lemmatization when it comes to the need for some gold standard to compare the tagging to. While it does hint at certain text-specific issues, such as nonstandard spelling or unusual vocabulary, many of the errors appears to stem from the ambiguity of some tokens. Aside from the presence of the impersonal verb forms, which are not strictly speaking nonstandard, this experiment does not reveal any new kinds of variation.

4.3 XPOS-tagging

In accordance with what was described in subsection 3.3, subsection 3.4, subsection 3.5, and subsection 3.6, the four tools used for XPOS-tagging experiments were BERT, Marmot, Stanza's neural pipeline and the combination of the morphological analyzer Morfeusz and a morphosyntactic tagger Concraft-pl. All of these tools were previously used for either lemmatization or UPOS tagging. While the same general evaluation measures as in UPOS-tagging were employed in this task, a decision was made to leave out the tag-specific measures due to the sheer number of possible classes. The possible labels for the fine-tuned BERT model, extracted from the training and test set of PDB along with the historical data, consist of 898 different classes.

The overall tagging performance of the tools in the XPOS-tagging task was worse than in the UPOS-tagging one. This could be attributed to a larger number of classes that require finer distinctions to be made. Same as before, all of the tools perform noticeably worse on historical data compared to modern data. The best-performing tool on both test sets is BERT and the worst is Marmot. What is worth pointing out in this case though is that Morfeusz and Concraft-pl's CRF

architecture does outperform Stanza's neural pipeline slightly, but only on modern data.

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 10: XPOS-tagging evaluation measures (accuracy, precision (macroaveraged and weighted), recall (macroaveraged and 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.

A manual error analysis and annotation have revealed a number of trends concerning the mistakes that the taggers make, as presented in Table 11 and Table 12. Definitions and examples of the errors can be found in Table 23 and Table 24 in Appendix B. A similar trend can be noticed in terms of the kinds of errors as in the UPOS-tagging task. This time, the most numerous error type is the **ambiguous** errors (both in the general and detailed error classification), where one word form corresponds to multiple possible tags (e.g. in some declension paradigms certain cases have the same form). Once again, the prevalence of these errors hints either at the taggers not being able to properly utilize the contextual information that is necessary for the disambiguation of the class either due to their architecture or to the text's unusual word order; and again, the latter cannot be fully concluded simply from these results. While **spelling** and **name** errors are still relatively prominent, the first category has become proportionally noticeably less common, while the latter one's relative frequency has doubled; the same can be said about the **unidentified** errors.

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

Table 11: General 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).

One relatively prominent error category, namely digits, appears to stem from the tools utilizing different strategies when it comes to numbers written as digits. Some of them classify them as *dig*, while others attempt to provide the tag as if the number were written out with letters. Another tagset-related issue is that of currency mistakes; the XPOS tagset generally divides masculine nouns into three subgenders: m1, which includes animate human nouns, m2, which includes animate non-human nouns, and m3, which includes inanimate nouns. These are supposed to reflect which form the determiner *który* 'which' takes when referring to that noun (Universal Dependencies, nda). According to this, however, there are less obvious words that belong to the m2 category, such as currency names, which did seem to pose a problem for the taggers. In general, some mistakes were made when distinguishing between the m1 and m3 categories for other tokens as well, but they were not classified separately.

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

Table 12: Detailed 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).

Finally, although not numerous, two kinds of errors are worth mentioning in the context of the data, namely the **gender** and **vocative** ones. Although the **gender** category partly overlaps with the aforementioned masculine subgender distinction, a number of the instances in this error category stem from the fact that not all personal pronouns in Polish reflect gender (the first and second person in both singular and plural do not have an overt marking of the gender on the pronoun itself). However, since some do, all of the personal pronouns are annotated for gender. This is

surprising, as for some of them the form does not indicate this feature, and it could only be deduced from the context, which is not always sufficient; it appears that the majority of such ambiguous cases in PDB are treated as masculine. This leads to a number of personal pronouns being misclassified as masculine, when they should, in fact, be feminine, as they are uttered by a female character and followed by feminine verb or adjective forms. While this is still a tagger error (there do exist instances in the PDB corpus where such a pronoun was classified as feminine), it highlights an important issue with the tagset and possibly the tagging convention. As for the **vocative** category, when the word can be interpreted otherwise, the taggers tend to not utilize the tags for the vocative case; one item classified as this kind of error actually supports the claim that the nominative and vocative could be the same in Borderlands Polish, as described in subsection 2.2. Both of these issues hint at the fact that the training data used for the taggers may be lacking in dialogues where both first and second-person personal pronouns and nouns in the vocative case would be more present. Additionally, the tendency to select masculine over feminine for the pronouns may hint at a potential gender bias in the data.

The issues that plague the XPOS taggers resemble those characteristic of UPOS tagging, with some novel types of errors being more indicative of the issues with the tagset or the training data, such as the **gender** and **vocative** ones. Once again, while this method does hint at unusual vocabulary and spelling practices being present in the text, and allows for the noticing of patterns within those trends, it requires prior manual annotation, which makes it more difficult to utilize it at a larger scale.

4.4 N-gram statistics

The constructed unigram comparisons for UPOS and XPOS tags offer insight into the composition of the text, without delving into the word order or sentence structure. Due to there being over 600 different XPOS tags in the combined test sets (and even more when it comes to their combinations in bi- and trigrams), the entirety of the comparison cannot be represented within this thesis, and instead a subsection of the most popular tags will be discussed. `.xlsx` files with full comparisons, sorted by the relative frequency of an n-gram in the PDB data can be found in the thesis repository, as outlined in Appendix A.

When it comes to the UPOS unigram frequencies, as presented in Table 13, for the majority of the tags the difference is not large. The exception is punctuation, which is noticeably less numerous in the memoir. This could be caused by unusual punctuation employed by the author or differences in terms of the structure of the sentences in both of the sets. Longer sentences without many commas or very few utterances in quotation marks could account for a part of this discrepancy. On the other hand, the memoir features proportionally more PROPN items than the PDB test set. As detailed in the other subsections of this section, these tokens do account for a non-negligible part of what sets the historical text apart from modern ones, both in terms of vocabulary differences and issues with lemmatization and tagging. While the prevalence of PROPN is more likely due to the genre or the topic of the memoir, higher relative frequencies of CCONJ and DET in the memoir seem to indicate something related more to the language or the style of the text. The frequency of CCONJ may indicate, along with the lower relative number of punctuation, that the author prefers long sentences the components of which are connected using coordinating conjunctions, instead of splitting them into shorter sentences. As for DET, it is important to note that while Polish does not have a proper article system, DET is used to denote demonstrative pronouns and possessive pronouns. While this could partly be explained by the topic and the fact that within a coherent

text like the memoir the author may, for example, refer more to previously mentioned entities (which would warrant using demonstrative pronouns), it could also, to an extent, be simply typical for the author's language. Another interesting general observation that can be made is that lexical categories, with the exception of PROPN and NUM, are more prominent in PDB than in the memoir, while the memoir features proportionally more function words. These differences hint at the possibility of a different distribution of certain constructions, but that could be a consequence of genre differences, not just diachronic or regional variation.

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

Table 13: The UPOS unigram % frequencies for the modern and historical test data. The higher relative frequency is indicated in bold, and the most prominent differences are in italics.

As mentioned before, bigram and trigram frequency lists are too large to be included in the thesis, possibly even in the appendix. They can be accessed in the form of `.xlsx` files in the repository included in Appendix A. Nevertheless, a discussion of them is warranted, as, unlike unigrams, they may hold clues to variation in word order and syntax. Since the relative frequencies provided for bigrams situate them in the general collection of all the bigrams, it is not possible to compare them directly within one subcategory. That is because, for example, the frequencies of all the bigrams starting with <BOS>[10] in the historical data do not add up to the same fraction as in the modern data. Therefore, while the bigram <BOS> SCONJ may have the relative frequency in the modern text of 0.15% and 0.10% in the modern text, 0.10% could constitute half of all the

[10] A placeholder or padding token marking the beginning of sentence.

instances of bigrams starting with <BOS> for the historical text, while 0.15% could instead only be a fourth of all such bigrams in the modern data. Because of that, comparisons across different n-grams starting with the same tag are not reliable.

Regardless of that, certain interesting trends can be noted. The relevant bigrams are displayed in Table 14 Although the sentences in the memoir tend to be longer than in PDB, the <BOS> CCONJ bigram is still relatively more numerous in historical data, which appears to be a peculiarity of the author's language, perhaps indicating a more informal style of writing. There is also a large discrepancy when it comes to <BOS> PUNCT, this time in favor of PDB, also indicating that there are more examples in that set that are marked as utterances in a dialogue or quotes. The historical data seems to feature more ADJ ADJ bigrams, which could indicate the omission of a comma that normally separates those. Relatively, the modern text features more ADJ NOUN bigrams than the memoir; it also has a higher relative count of NOUN ADJ bigrams, which indicates that a comparison against the whole selection of bigrams and not just within a subcategory can be less informative, as one needs to factor in the frequency of the constituent tags themselves.

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

Table 14: Relative frequencies for the modern and historical data for selected UPOS bigrams, rounded to two decimals.

The historical text seems to feature more ADP PROPN bigrams, which is reasonable given the prevalence of the PROPN tag in the memoir. There appears to be some variation in terms of the bigrams starting with AUX, with a surprising disparity in favor of PDB when it comes to the AUX VERB bigram; this is especially noticeable given the fact that proportionally the counts for AUX and VERB separately are not very different in both test sets. One possible explanation is that PDB may contain more conditional and future clauses, where auxiliaries are used, whereas

the memoir is mostly recounting the past, where auxiliaries are not nearly as common. AUX appears to be followed more often by ADP or NOUN in the historical data. The DET NOUN and DET PROPN combinations are noticeably more frequent in the historical data than in the modern data, but there is no large difference between the frequencies for DET ADJ for the two sets. Similarly, the historical data shows proportionally many more NOUN DET combinations, hinting at a possible word order variation. The same can be noticed for PROPN DET and DET PROPN. While verbs seem to be more often directly followed by nouns in the historical data, it is impossible to determine whether that is due to some word order inversion or simply because of the object immediately following the verb. There is a small number of bigrams that only appear in one of the sets of data, partly due to INTJ and SYM only being present in the PDB data.

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

Table 15: Relative frequencies for the modern and historical data for selected UPOS trigrams, rounded to two decimals.

As far as trigrams are concerned, many more combinations are available, and, unfortunately, not all of them can be discussed within this subsection, so a special focus will be put on the trends that were already previously noticed, such as the word order in adjectival phrases, and the relevant trigrams and their relative frequencies are can be found in Table 15. The ADJ DET NOUN and ADJ NOUN DET trigrams appear more common in the historical data. DET ADJ NOUN appears to be equally frequent, while DET NOUN ADJ and NOUN DET ADJ are again more common in the historical data. While these results may be hinting at a tendency for unusual word order in certain phrases, this method does not allow for the distinction of whether the ADJ and DET elements were actually modifiers for the NOUN token. While this could be more noticeable in XPOS bigrams and trigrams, the number of possible combinations makes it impossible for any more in-depth analysis of those results without further computational processing, which is why the discussion of the results obtained for XPOS bigrams and trigrams is omitted (while the XPOS unigrams are discussed below).

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

Table 16: The top 10 XPOS unigram % frequencies with the largest difference between the modern and historical test data. The higher relative frequency is indicated in bold. The full comparison of unigrams can be found in Appendix A.

While due to the sheer number of the XPOS tags appearing in the test data an equally detailed analysis cannot be conducted, certain interesting trends can be identified, especially when the classes with the largest frequency differences are considered (Table 16). Once more, punctuation appears on the top of the frequency table, with the PDB test set having proportionally more such tokens. What is interesting is that the PDB test set features significantly more verbs in the present tense (fin) as opposed to the memoir, in which past forms (praet) prevail. This is not unexpected given the genre of the memoir, but it could also hint at certain forms being underrepresented in the PDB. The prevalence of tokens marked as subst:sg:acc:m1, which refers not just to masculine, but masculine animate human nouns, indicates a certain bias in the memoir, where male characters were given more spotlight by the author. Looking at the full data, contrary to the expectations, vocative forms do not appear to be proportionally severely underrepresented in the PDB, but they are not frequent in general, which may lead to issues with detecting those forms by tools trained on the PDB. There is a number of tags that only appear in one of the test sets; their total number is still smaller than that of the classes used for training the BERT model (based on the test sets and the PDB training set). However, it is possible that even this collection of tags is incomplete given the number of possible feature combinations.

While this method does reveal distributional differences, its usefulness is mostly constrained to unigrams and UPOS bigrams, as more processing of the results would be needed for data where there are thousands of n-gram combinations. Additionally, while presenting the bigram or trigram frequency as a fraction of the whole set of n-grams does allow for some comparisons, representing the frequency within the category of bigrams that start or end with a given element could yield additional insights. The lack of explicit indication as to what relation the elements in an n-gram have to each other makes this method inferior to the one employed by Johannsen et al. (2015), although the annotation required for it may be less time-consuming given that the data in question is historical and the performance of pre-trained dependency parsers on it may be low. Nevertheless, since an annotation effort already has to be made, including syntactic relations in the annotation

and then conducting an analysis of the syntactic treelets should yield clearer results.

4.5 The National Corpus of Polish vocabulary comparison

As discussed in subsection 3.9, a comparison of the unique tokens and lemmas from both of the test sets against a subsection of the data from the National Corpus of Polish was conducted. As visualized in Table 17, noticeably more tokens and lemmas from the historical data were not found in the National Corpus of Polish when compared to the results from the PDB test set. What is worth noting is that while PDB is based on the texts from the National Corpus of Polish, it does include some sentences that do not come from that corpus and can therefore include tokens and lemmas that are not present in the corpus.

An inspection of the items that were returned as having no matches in the National Corpus of Polish for the PDB test set (which can be found in Appendix D) reveals that many of them overlapped across the lemma and token category (i.e. if a token was not found in the corpus, neither was its lemma and vice versa). A number of the missing items include punctuation, which appears to not be included in the corpus searches, as well as different numbers and numerals, which, due to their practically infinite number, is understandable. Additionally, a number of place names, names, surnames, and brand names were not found in the National Corpus of Polish. Finally new borrowings with nonstandard spelling reflecting the original pronunciation (the diminutive *lajwik* 'live (stream)') as well as highly specialized vocabulary (*trichlorobenzen* 'trichlorobenzene') were not found in the National Corpus of Polish either, which is warranted as these are either new or very rarely used words.

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

Table 17: Raw and % numbers of tokens and lemmas unique to the modern or historical test sets when compared with a subset of the National Corpus of Polish.

A similar trend in terms of both the token and the lemma missing from the National Corpus of Polish can be noticed for the historical data, although here the comparison is more difficult, as the unique tokens were extracted from a larger text sample than the unique lemmas (which had to be manually annotated). Once again, some punctuation is listed as not found due to the search engine's limitations. However, almost no standard numbers or numerals are listed as not found, with the exception of *cwansiger* '20 (coin/bill),' which appears to be a borrowing from German. Similarly as with modern data, a large number of proper names and surnames was not found in the corpus. A large part of the vocabulary with no hits in the National Corpus of Polish consists of either words spelled in a nonstandard fashion (with the spelling of *nie* together with the verb it modifies and the use of the grapheme *y* for /j/ being the most prominent; there is a number interesting instances of words featuring the sound /ʐ/ in standard Polish being spelled as *rż*, while the two accepted modern spellings are *rz* and *ż*. Only one of those words was problematic for the taggers,

and therefore this tendency was not singled out as a separate error category. Nevertheless, similarly to the tendency to replace *a* with *e*[11], this appears to be specific to Borderlands Polish, where, as discussed in subsection 2.2, at least in pronunciation, the phoneme /r̪/ (historically spelled as *rz*) did not merge with /ʐ/ (spelled as *ż*), but instead evolved into /rʐ/ or /r/[12]; while not detected in the previous experiments, this supports the claim that the text displays regional variation as well. Other words that were not detected in the National Corpus of Polish were words that appear to be out of use or highly specific to the sociohistorical context of the text (such as *mandatariat* 'the position of being a potentiary' or *mortyfikować* 'to self-flagellate'), with a potential overlap between the two categories. What is also noticeable is that some of the words reappear with multiple variations of spelling (e.g. *jurysdykcya*, *juryzdyksya* 'jurisdiction' or *mandataryusz*, *mandatyrusz* 'potentiary') indicating a certain degree of inconsistency when it comes to spelling; simultaneously, when it comes to tokens, and not lemmas, some words reappear with the same nonstandard spelling but various inflectional endings.

Having previously noted that the historical text in question appears to bear some features of Borderlands Polish (as outlined in subsection 4.1), the items with zero hits returned for the memoir, with the exception of proper names and foreign words, were compared with a lexicon of Borderlands Polish, as provided by Kurzowa (1983). Most of the items were not present in the lexicon, with the exception of *cwansiger* '20 (coin/bill)' could be found, albeit with a different spelling (*cwancygier* in the lexicon). Alongside a number of words featuring the aforementioned pronunciation differences (*rż*, *e* errors), this strongly supports the claim that the text bears features of the Borderlands dialect of Polish. Additionally, one may expect to find more overlap between the lexicon of the memoirs and the vocabulary typical for Borderlands Polish, but those words may have been found in the corpus as well, excluding them from the output in this experiment.

A simple comparison of the vocabulary of a given text or dataset to that of a large corpus of a given language appears to yield informative results as far as those texts' divergence from the standard is concerned. With the exclusion of terms that are naturally unlikely to appear in the corpus (surnames, proper names), there still remain tokens and lemmas that were not identified in the corpus due to their spelling or rarity, and some of the same conclusions as to the nature of these differences can be drawn here as from the experiments subsection 4.1, subsection 4.2, and subsection 4.3. While this method does not return as many items that were not found (or, in the case of the taggers, errors), it only requires the text to be lemmatized, not annotated for the part of speech — and even simply searching for the tokens, and not their lemmas, yields interesting results that are not extremely different from those for lemmas. Overall, this kind of comparison appears to be quite rewarding for the amount of preprocessing or annotation required.

4.6 Discussion of Results

With the aim of the thesis, as specified in section 1, being simultaneously trying to identify the variation present in the memoir and to assess the usefulness of selected methods in identifying variability, a summary and discussion of the results from both of these perspectives is in place. The tool-based methods (measures and error review for lemmatization and two kinds of part-of-speech

[11] Which could, to some extent, be noticed in this comparison, but which was much more prominent in the other experiments, indicating that this type of spelling or pronunciation is also present in the National Corpus of Polish.

[12] In all of these cases the fricative could also be realized as voiceless, depending on the surrounding sounds.

tagging) have all revealed similar kinds of variation, albeit at varying levels, since clearly it did not hinder their tagging and lemmatizing performance in all the cases; much of what was detected in these experiments was also observed in the vocabulary comparison with the National Corpus of Polish. In relation to the rest, the n-gram experiments did not provide as much information about the variation, but they did give insights into the aspects of the language in the memoir that were not accessible in other experiments, such as syntax. Certain features could have been considered negligible, had it not been for their importance in the dialectal context, as described in subsection 2.2, e.g. the cases of *rż*, the vocative, or attested dialectical vocabulary that was not found in the National Corpus of Polish.

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	-

Table 18: A comparison of the kinds of variation identified in various experiments.

While, unfortunately, there is no benchmark to compare the performance of these methods to, it should be noted that not much other variation has been noticed by the author whilst reading the memoir itself than what was detected in the experiments. The only peculiarity which was found while reading through the text but not in the results presented in this thesis is the way in which the word order and sentence construction seemed to differ from written modern Polish. Overall, the tool-based methods seem to be quite proficient at picking out spelling and pronunciation differences, with lemmatizers being likely the easiest tools for that. Part-of-speech taggers additionally reveal the unification of vocative and nominative, but with only one example. The major disadvantage of these three methods is that it not only requires the manual annotation of the text in question, but also subsequent error annotation. Given the differences in prominence of different errors on the same data, they cannot be used to assess how widespread the phenomenon is either. The corpus vocabulary comparison requires, at best, the text to be lemmatized, but that is not that necessary, as the results are similar for tokens and lemmas. This method does show a lot of the

spelling variation, with the exception of the irregular capitalization, as the API and search engine appear to ignore capitalization. While this method requires significantly less preparation, it clearly can lead to some variation being overlooked. Finally, the n-gram analysis does show differences, but it is difficult or impossible to draw many conclusions from them; the multiplicity of classes and the n-grams built from them leads to very sparse results that are difficult to read and interpret; additionally, there is no indication of the kind of relation between the elements in question, meaning that e.g. not all NOUN ADJ bigrams need to be a combination of a noun and an adjective that describes that noun, they can just be adjacent. All in all, while the methods should not be considered perfect, they do paint a picture of the variation present in the memoir, which can inform further inquiries.

4.7 Results and prior research

While this topic has been discussed, to an extent, in the previous subsections, it is important to summarize the relation of these experiments and their findings to existing research. As far as subsection 2.1 and subsection 2.2 are concerned, the variation identified in the memoir does not diverge significantly from what has previously been described; the text shows a number of features typical for Borderlands Polish, alongside more than one spelling convention for the /j/ sound. With subsubsection 2.3.1 in mind, the annotated excerpt from the memoir can serve as a basis for future analyses, while the comparison of tool performance can inform the selection of a tool for some task related to processing historical or dialectical data. As previously discussed, the suggested n-gram method for syntactic variation detection is somewhat lacking in comparison with Johannsen et al. (2015), although it evades the need for syntactic relation annotation, which could be problematic for a historical text.

Results from the POS tagger experiments using non-preprocessed data show a relatively high tagging performance compared to what was presented in subsubsection 2.3.3; while in this thesis the results are used to assess whether variation is responsible for the errors made by the tools, it is not the focus of the aforementioned papers. Simultaneously, the evaluation of multiple tools on that historical data that has been conducted as a part of this experiment constitutes a relevant contribution to the discussion of the use of modern POS-tagging tools on historical data. While no data normalization methods were employed in these experiments, the conclusions made by Dipper & Waldenberger (2017) that mappings and rules used for normalization can be used to assess variation are relevant to this thesis; if what needs normalization is what displays some kind of variation, then the errors made by tools on a non-normalized text should also be informative, as they have proven throughout this thesis.

5 Ethical Considerations

Ethical concerns are ever-present within the field of Natural Language Processing. As Hovy & Spruit (2016) point out, these concerns can revolve both around the data itself and the impact that NLP can have on society. Bender et al. (2021) point out the environmental impact of computationally-heavy processes (such as training very large language models) and draw the readers' attention to how biases existing in the training data can impact the aforementioned models. A number of different tools and resources were utilized in this thesis, and many of them deserve to be discussed from an ethical point of view.

To begin with, the experiments conducted as a part of this thesis did not involve training any large models from scratch — and the most computationally expensive part was the fine-tuning of two BERT-based part-of-speech taggers. While training a large transformer model like BERT is definitely impactful, its ability to be fine-tuned for different applications eliminates the need to train another costly model from scratch. Utilizing pre-existing, optimized code for token tagging suited for this model likely streamlined the process as well. With the exception of Marmot, the training of which is not computationally expensive, the other taggers were already pre-trained, which made this investigation much more justifiable than training many models from the start would be with the environmental impact in mind.

While a lot has been written about different biases in NLP, Blodgett et al. (2020) find that many such discussions are "vague, inconsistent, and lacking in normative reasoning." They adopt a division of biases into allocational, meaning ones where the system bias distributes some resources unfairly to some social groups, and representational, where some groups are misrepresented or omitted by the system. The authors present recommendations for future work with bias in NLP, and while some of them are not relevant to this thesis, their suggestion to explicitly state what behaviors and what kinds of biases exist in the system or the data, how they could be harmful, and to whom, is of high importance. One instance of a representational bias displayed by the tools tested in this thesis has been described in subsection 4.3. According to the UD XPOS annotation, the first- and second-person singular pronouns are annotated for gender despite not overtly displaying it. During the tagging process, the tools did annotate pronouns used by a female speaker as masculine. This kind of a system behavior could be harmful if displayed at a larger scale, as it disregards the presence of women as speakers — depending on what this tagging is used for in a downstream task, this could lead to e.g. a dialogue system addressing its interlocutor using incorrect pronouns and forms.

Biases in large language models and other NLP tools do not necessarily stem from the code itself. As Garimella et al. (2021) note, "unstructured data often contain several biases, and natural language processing (NLP) models trained on them learn and sometimes amplify them." It is therefore important to discuss the kinds of data used in this thesis and whether or not they can contain such biases (potentially leading to the aforementioned gender representation bias). The author of the Polish version of BERT, Kłeczek (2021), points out that the data used to pre-train his models may include biases and stereotypes, and, consequently, these could be visible in downstream tasks. While the National Corpus of Polish is claimed to be balanced and representative of the language at large, its creators do not address the issue of biases, and one can assume that certain prejudices can be reflected in the texts that constitute the corpus (Przepiórkowski et al., 2012). While, as Wróblewska (2018) explains, the Polish Dependency Bank is largely based on the National Corpus of Polish, it does include sentences from other sources and a similar conclu-

sion can be drawn about this dependency bank as about the National Corpus of Polish. As far as the historical data discussed in this thesis is concerned, it is likely to contain biases, as it only comes from one author. As pointed out in subsection 4.4, in comparison with the PDB, the author of the memoir uses proportionally more masculine animate human nouns, potentially leading to an over-representation of men at the cost of other genders, a bias that could be amplified had this data been used to train a tool for later use.

Another tangentially related issue worth considering in the light of this thesis is the representation of small languages or dialects and the regional and diachronic variation in NLP. McEnery et al. (2000), Soria et al. (2016), as well as Hovy (2018) point out that the lack of corpus data for such languages severely impedes the development of appropriate NLP tools, which can, in turn, lead to some social groups being excluded from utilizing such tools or result in their language becoming more endangered. While the data analyzed in this thesis is not a sample of a currently spoken modern minority language or dialect, some of the methods tested in the experiments could be used for exploring contemporary language variation as well, potentially contributing to solving this issue.

Finally, while both the out-of-context sentences provided in PDB and the limited access that users have to the texts that constitute the National Corpus of Polish are methods for dealing with copyright and privacy issues, working with an independently transcribed and annotated text may pose its own ethical problems. However, in the case of this thesis, the data in question is historical, and its author passed away around a century ago, which largely voids the issue of the author's consent for the use of his text.

6 Critiques and Limitations

While certain limitations of this thesis project have already been mentioned in previous sections, it is worthwhile to summarize them and discuss other potential issues. The first few of those pertain to the data itself, as outlined in subsection 3.1. While the original manuscript is available physically at a library in Poland, the version of the data utilized in this project has been manually copied more than once on its way into a `.docx` file, by people who were neither involved in the writing of this thesis, nor trained in historical data transcription. This introduces possibilities for transcription errors, making the data less reliable than if it were transcribed directly from the manuscript.

Another kind of limitation related to the data could be that it only comes from one author and is not very large. However, the aim of this thesis was to explore this particular memoir, not the entirety of late 19th-century Polish. One issue connected with this is that, consequently, the genre of the memoir and the texts featured in the PDB are not necessarily the same, and some of the differences may stem not from language variation at large, but from this genre mismatch. While it is relevant to keep in mind that this data is not very representative, this limitation does not necessarily invalidate the thesis project.

Finally, the project would have greatly benefitted from all the experiments being run on an additional sample of older historical data. Unfortunately, due to differences in the tagsets, the Korba Corpus could not be used. While results for historical data do exist for one of the taggers, this, unfortunately, meant that the other tools' performance on the memoir (both in terms of their intended use and identifying variation) in question could not be compared to older data, and thus the effectiveness of the methods could not be judged based on such a comparison.

As far as the annotation is concerned (subsection 3.2), the limited time and womanpower available meant that only a part of the data could be annotated. Ideally, this annotation should be proofread by another trained native speaker, but, unfortunately, that was not possible within the given timeframe. This is likely to have negatively impacted the quality of the annotation. Additionally, as discussed in subsection 4.4, including the syntactic dependency annotation could have yielded more interesting results; once again though, the choice to omit this annotation was made due to the aforementioned reasons. On the topic of annotation, one of the limitations of many of the discussed methods is that they require the historical data to already be annotated. However, those methods could also point in the direction of what kind of pre-processing would be needed for a more reliable automated annotation of historical data.

While the tagging performance of BERT-based tools was outstanding, it is possible that it could have been better. As described in subsection 3.3, the hyperparameters used to fine-tune the taggers were the ones suggested by default by the authors of the tagging framework. It is possible that these were not optimal for the task.

The error annotation (subsection 3.10) conducted in this project is largely subjective, and some of the categories are partly overlapping. Some tokens could likely be classified as more than one category. It is important to keep in mind that this annotation was not intended to yield a strict measure but to offer more general insights into what kinds of tokens are the most problematic for the taggers and lemmatizers.

Finally, as discussed in more detail in subsection 4.4, the time limitations made it impossible to

analyze and discuss all of the obtained n-gram results. The number of possible XPOS tags is staggering, and the decision to construct n-grams out of them without clustering them into more general categories resulted in an output that was too large to analyze in the allotted time; additionally, the multiplicity of categories made the results very sparse, and, therefore, hard to analyze regardless of the time constraints.

7 Future Work

A number of the issues mentioned in section 6 could be better addressed had the scope of the project been wider and had it been possible to allot more time to it — and it is some of these discarded ideas that form the basis of the potential future work.

As far as the memoir itself is concerned, completing the annotation thereof with lemmas, UPOS, and XPOS tags would have constituted a small but valuable contribution to the body of annotated historical Polish. It would also have been interesting to see this data with full UD-style annotation, including dependency relations. Furthermore, the library where the manuscript is held appears to be in possession of some of the correspondence by the same author, which could be similarly transcribed and annotated; the memoir's digital version could also benefit from being compared to the contents of the manuscript to eliminate potential transcription errors.Following what has been said in subsubsection 2.3.5, it might be worth it to try to adapt existing tools for better automatic annotation when it comes to extending the scope of the annotation of this data.

Including more of the data and a fuller annotation could potentially reveal more kinds of variation that may not be evenly distributed within the text. The presence of dependency relation annotation would enable the use of the methods implemented by Johannsen et al. (2015) for a higher-quality analysis of the syntax of the memoir. Refining the methods for utilizing n-gram counts, especially when it comes to the XPOS tags, could yield new insights as well.

Simultaneously, such inquiries pertaining to language variation could be conducted on more data. Both older and more contemporary non-standard data, as well as, potentially, data contemporaneous to the memoir could be explored, and, perhaps, some trends could be identified. Hearkening back to the idea of including more of the author's writing, and referring back to the Korba corpus, which features 17th- and 18th-century texts, the construction and annotation of a diachronic corpus of Polish for a different time period or spanning a larger time period, with the use of a tagset compatible with the UD XPOS tags (if such annotation were to be included) could be extremely beneficial for quantitative investigations into the history of Polish. Alternatively, the focus could be put on regional variation (or on both historical and regional one), as the text discussed in this thesis does display features characteristic of a group of regional dialects. What could be particularly interesting is utilizing the various methods for modelling language variation and change discussed in subsubsection 2.3.6 when it comes to e.g. various Borderlands Polish texts synchronically or diachronically.

The experiments reveal certain issues that the tools that were tested struggled with when faced with nonstandard data. While it was not the goal of this project, it could be useful to analyze these issues and explore pre-processing or normalization methods that could be implemented if such tools were to be used for the automatic annotation of larger amounts of historical data, which could be incredibly helpful given the scarcity thereof.

One direction in which the tagger testing could be developed could be to only review certain tokens where the tagger confidence was below some threshold. Such a method could also be utilized with no golden standard available (on unannotated data). Unfortunately, this idea could not be applied to all of the taggers utilized in this project, as not all of them return confidence scores, at least not in an obvious way. Another alternative that could eliminate the need for manual annotation would be to process the input text using multiple tools (e.g. multiple lemmatizers or part-of-speech taggers)

and focus on the error annotation of tokens where the tools do not return the same kind of tag or lemma; the drawback here could be tokens that are confusing all of the tools in the same way, so there is cross-tool agreement on the tag, but it is not the correct tag nonetheless.

8 Conclusions

The aim of this thesis was to explore some of the potential methods for identifying language variation in Polish on the example of a late 19th-century memoir by Juliusz Czermiński and the ways the language of the memoir differs from modern Polish. The text was expected to differ from modern standard Polish in some ways due to its age and geographical origin. A part of the text was manually annotated with lemmas, UPOS, and XPOS tags according to the UD standards for such annotation. Subsequently, a number of experiments were conducted, where the memoir was compared to the test set of PDB-UD, the largest existing UD treebank for modern Polish. The experiments included comparing the performance and output of various tagging and lemmatization tools on the two sets of data, reviewing the features of the most problematic tokens, part-of-speech tag statistics analyses, and a review of which tokens and lemmas from the data are not present in a specific subsection of the National Corpus of Polish.

The results, presented and discussed in section 4, show that the memoir does differ from resources that are available for modern Polish, and some major trends in terms of spelling variation (the use of *y* for the /j/ phoneme, spelling the negation of a verb together with the verb itself) and spelling that reflects potential phonological differences (the use of *e*, likely signifying /ɛ/, where modern Polish features the phoneme /a/, the use of *rż* in place of *rz*, reflecting the pronunciation of /rzˌ/ in place of the standard /zˌ/) are identified. A noticeable drop in tagging or lemmatizing performance can be observed for all three categories of annotation tools, regardless of their architecture (although some perform better than others). N-gram counts of the part-of-speech tags suggest possible word order or syntactic differences but are inconclusive. A comparison of the memoir's vocabulary reveals a number of tokens that are not present in the National Corpus of Polish in the selected timespan; while some of those are proper names, other examples show spelling and vocabulary variation. Inevitably, the methods explored in this project have their drawbacks, the major one being that most of them require the data to be annotated in some way; they are also not equally reliable at detecting all kinds of variation. Simultaneously, this thesis offers a small contribution to the body of annotated historical data for Polish and advocates for the usefulness of constructing larger collections of diachronic data with annotation compatible with other large annotated corpora. One more side effect of the tagger and lemmatizer experiments is that they provide a comparison of the performance of various tools on modern data.

The experiments and results presented in this thesis explore the ways in which existing tools for modern languages can help identify language variation in historical texts. While the presented solutions may not be perfect, they encourage further discussion and research into utilizing them for diachronic linguistics, not only for simply identifying the language variation but also as an intermediate step in the process of automatizing the annotation of historical data for the creation of larger corpora.

References

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' 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* (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.

A Resources

Within this Appendix, links to a number of resources, both created and used in the thesis, will be provided, alongside explanations; these resources have been cited in the text according to the resource-specific guidelines (if provided):

- Thesis repository, including the text of the memoir with and without various annotation (the repository structure is explained in the README): https://github.com/Turtilla/swe-ma-thesis
- Instructions and contact information for the PELCRA search engine for the National Corpus of Polish[13]: http://www.nkjp.uni.lodz.pl/help.jsp
- The University of Sheffield UD-based POS tagger[14]: https://cloud.gate.ac.uk/shopfront/displayItem/tagger-pos-pl-maxent1
- Morfeusz2: http://morfeusz.sgjp.pl/
- Concraft-pl: https://github.com/kawu/concraft-pl
- Transformers' Token Classification: https://github.com/huggingface/transformers/tree/main/examples/legacy/token-classification
- BERT for Polish (cased): https://huggingface.co/dkleczek/bert-base-polish-cased-v1
- Marmot: http://cistern.cis.lmu.de/marmot/
- Stanza: https://stanfordnlp.github.io/stanza/

[13] If one desires the so-called programming access, one must contact the person listed on this page, as this access is limited.

[14] As of 26.04.2023 this tagger appears to not work due to an "Internal Server Error".

B Error Type Definitions

Error Type	Definition	Included Subtypes
spelling	Any spelling-related differences, both intentional and not	*y*, *nie*, spelling, capitalization, *e*
name	Any type of proper names	proper name, surname, name
abbreviation	Abbreviated tokens	abbreviation
ambiguous	The whole token or its part is ambiguous in some way	ambiguous, problematic
unidentified	The reason for the error cannot be identified	unidentified
vocabulary	The token is likely OOV due to being specialized, dialectical, archaic, or foreign	foreign
grammar	The token displays an unusual grammatical feature	grammar

Table 19: General types of errors made by lemmatizers.

Error Type	Definition	Example	Predictions	Standard
spelling: *y*	The grapheme *y* is used instead of *j* to signify the /j/ sound	*suchey* 'dry'	suchey	suchy
name: other	Potentially unfamiliar proper name token	*Bludniki* 'Bludniki'	Bludnik bludnik	Bludniki
spelling: *nie*	Spelling of the negation with the negated word in word classes that normally do not allow it	*niemają* '(they) don't have'	niemaja nie	niemieć
spelling: other	Other spelling differences	*ładąn* 'pretty'	ładąn	ładna
name: surname	Potentially unfamiliar surname token	*Polanowski* 'Polanowski'	polanowski	Polanowski
spelling: capitalization	Nonstandard capitalization	*Dziedzica* 'of the heir'	Dziedzic dziedzica	dziedzic
abbreviation	The token is abbreviated	*Stan* 'Stan'	Stan stan	Stanisław
spelling: *e*	The grapheme *e* is used instead of another vowel (commonly *y*)	*tem* 'this'	tema tem	to
ambiguous: other	The token could have more than one interpretation	*dobra* 'goods'	dobry	dobra
name: given name	A potentially unfamiliar first name token	*Kleosię* 'Kleosia'	Kleosię kleosia	Kleosia
unidentified	No apparent reason	*łania* 'doe'	łani łanie	łania
ambiguous: problematic	The choice of the lemma is up to the annotator because of two acceptable spelling variants or the word being on the verge of being independent or being a derivational form of another word	*bombardowaniu* 'of the bombing'	bombar- dować	bombardo- wanie
vocabulary: foreign	The token is foreign	*Toje* '-'	Toje tój	toje
grammar: other	The token displays an unusual grammatical ending	*człowiecze* 'human'	człowieczy człowiec	człowiek

Table 20: Types and examples of errors made by lemmatizers. The translations into English are not ideal since they do not capture all of the encoded information, such as case, gender, number.

Error Type	Definition	Included Subtypes
spelling	Any spelling-related differences, both intentional and not	capitalization, *y*, *e*, *nie*, spelling
name	Any type of proper names	surname, proper name, name
abbreviation	Abbreviated tokens	abbreviation
ambiguous	The whole token or its part is ambiguous in some way	ambiguous, UD, ending, problematic, digits
unidentified	The reason for the error cannot be identified	unidentified
vocabulary	The token is likely OOV due to being specialized, dialectical, archaic, or foreign	archaic, foreign, uncommon, special
grammar	The token displays an unusual grammatical feature	impersonal, grammar

Table 21: General types of errors made by UPOS taggers.

Error Type	Definition	Example	Predictions	Standard
ambiguous: other	The token is ambiguous	*jego* ‘his’	PRON	DET
spelling: capitalization	Nonstandard capitalization	*Patrona* ‘patron’	PROPN NOUN	NOUN
spelling: *y*	The grapheme *y* is used instead of *j* to signify the /j/ sound	*móy* ‘my’	PROPN ADJ VERB	DET
unidentified	No apparent reason	*wyłącznie* ‘exclusively’	PART ADV NOUN	ADV
vocabulary: archaic	The token is somewhat archaic or regional	*pono* ‘supposedly’	PUNCT VERB PROPN ADJ	PART
ambiguous: UD	The token has more than one possible tag due to UD guidelines	*był* ‘(there) was’	VERB AUX	VERB
name: surname	Potentially unfamiliar surname token	*Ostaszewskiej* ‘Ostaszewska’	ADJ PROPN	PROPN
spelling: *e*	The grapheme *e* is used instead of another vowel (commonly *y*)	*małem* ‘small’	ADJ NOUN	ADJ
spelling: *nie*	Spelling of the negation with the negated word in word classes that normally do not allow it	*niechciały* ‘(they) didn’t want to’	VERB NOUN ADJ	VERB
spelling: other	Other spelling differences	*wkońcu* ‘in the end’	ADV NOUN	NOUN
ambiguous: ending	The ending of the word can be indicative of more than one class	*chwała* ‘glory’	NOUN VERB	NOUN
name: other	Potentially unfamiliar proper name token	*Dąbrowy* ‘Dąbrowa’	PROPN ADJ	PROPN
ambiguous: problematic	The choice of the tag is up to the annotator because of two spelling variants or the word having been derived	*służąca* ‘servant’	NOUN ADJ	NOUN
ambiguous: digits	The token is in digits	*1* ‘1’	ADJ NUM	NUM

vocabulary: foreign	The token is foreign	*daruju* '-'	NOUN VERB	X
vocabulary: uncommon	The token is uncommon	*czółno* 'canoe'	ADV ADJ	NOUN
abbreviation	The token is abbreviated	*5-cioro* 'five'	NUM NOUN MIS-PARSED	NUM
grammar: impersonal	The token is an impersonal verb form	*wierzono* '(it was) believed'	VERB ADJ ADV	VERB
name: given name	A potentially unfamiliar first name token	*Wiktorów* 'Wiktors'	PROPN NOUN	PROPN
grammar: other	The token features a nonstandard inflectional ending	*egzamina* 'exams'	NOUN ADJ	NOUN
vocabulary: stylized	The token is a intentionally spelled in a nonstandard fashion	*psipiólki* 'quails'	NOUN ADJ	NOUN

Table 22: Types and examples of errors made by the UPOS taggers. The translations into English are not ideal since they do not capture all of the encoded information, such as case, gender, number.

Error Type	Definition	Included Subtypes
spelling	Any spelling-related differences, both intentional and not	*y*, *nie*, spelling, *e*
name	Any type of proper names	proper name, surname, name
abbreviation	Abbreviated tokens	abbreviation
ambiguous	The whole token or its part is ambiguous in some way	ambiguous, digits, problematic, currency,
unidentified	The reason for the error cannot be identified	unidentified
vocabulary	The token is likely OOV due to being specialized, dialectical, archaic, or foreign	archaic, foreign, uncommon
grammar	The token displays an unusual grammatical feature	gender, vocative

Table 23: General types of errors made by XPOS taggers.

Error Type	Definition	Example	Predictions	Standard
ambiguous: other	The token's meaning is ambiguous	*parafii* 'parish'	subst:sg:gen:f	subst:sg:loc:f
unidentified	No apparent reason	*Ciotka* 'aunt'	subst:sg:acc:m1 subst:sg:nom:m1 subst:sg:nom:f	subst:sg:nom:f
name: other	Potentially unfamiliar proper name token	*Brzeżan* 'Brzeżany'	subst:pl:gen:n:pt subst:sg:gen:n:ncol subst:pl:gen:m1 subst:sg:gen:m1	subst:pl:gen:n:pt
spelling: *y*	The grapheme *y* is used instead of *j* to signify the /j/ sound	*arye* 'arias'	subst:pl:acc:m3 adj:pl:nom:m2:pos	subst:pl:acc:f
ambiguous: digits	The token is in digits	*1808* '1808'	adj:sg:gen:m3:pos dig	adj:sg:gen:m3:pos
ambiguous: problematic	The choice of the tag is up to the annotator because of two spelling variants or the word having been derived	*oczytany* 'learned'	adj:sg:nom:m1:pos adj:sg:nom:m3:pos ppas:sg:nom:m1:perf:aff	ppas:sg:nom:m1:perf:aff
spelling: *nie*	Spelling of the negation with the negated word in word classes that normally do not allow it	*niema* '(there) aren't'	fin:sg:ter:imperf subst:sg:nom:f	fin:sg:ter:imperf
spelling: other	Other spelling differences	*kończ* 'end'	subst:pl:gen:n:ncol subst:pl:gen:f impt:sg:sec:imperf	subst:sg:gen:m3
vocabulary: archaic	The token is somewhat archaic or regional	*wieleż* 'many'	num:pl:acc:m3:rec subst:sg:nom:m3 subst:sg:nom:m1	num:pl:acc:m3:rec
vocabulary: foreign	The token is foreign	*Toje* '-'	subst:sg:nom:n:ncol xxx subst:sg:nom:m1	ign
name: surname	Potentially unfamiliar surname token	*Zabilskich* 'Zabilscy'	subst:pl:gen:m1 adj:pl:gen:f:pos subst:pl:acc:m1	subst:pl:gen:m1
vocabulary: uncommon	The token is uncommon	*hoża* 'swift'	adj:sg:nom:f:pos subst:sg:nom:f	adj:sg:nom:f:pos
ambiguous: currency	The token is a name of a currency and belongs to the *m2* gender	*dukatów* ducats	subst:pl:gen:m3 subst:pl:gen:m2	subst:pl:gen:m2

spelling: *e*	The grapheme *e* is used instead of another vowel (commonly *y*)	*któremi* '(with) which'	adj:pl:inst:n:pos subst:pl:inst:m3 subst:pl:inst:m1 adj:pl:inst:f:pos	adj:pl:inst:m1:pos
grammar: gender	The token is assigned the wrong gender	*Ja* 'I'	ppron12:sg:nom: m1:pri	ppron12:sg:nom: f:pri
grammar: vocative	The vocative case is not properly recognized	*Ty* 'you'	ppron12:sg:nom: m1:sec	ppron12:sg:voc: m1:sec
abbreviation	The token is abbreviated	*śp* 'may his soul rest in peace'	brev:pun subst:sg:nom:m3 subst:sg:nom:m1 aglt:sg:sec: imperf:nwok	brev:npun
name: given name	A potentially unfamiliar first name token	*Melchior* 'Melchior'	subst:sg:nom:m1 subst:sg:nom:m3 subst:sg:gen:f	subst:sg:nom:m1

Table 24: Types and examples of errors made by the XPOS taggers. The translations into English are not ideal since they do not capture all of the encoded information, such as case, gender, number.

C Extended UPOS measures

POS-tag	Modern		Historical	
	Precision	**Recall**	**Precision**	**Recall**
ADJ	99.11%	99.55%	94.23%	93.31%
ADP	99.77%	99.91%	99.74%	98.74%
ADV	97.73%	98.44%	87.61%	89.94%
AUX	98.93%	98.93%	91.32%	84.03%
CCONJ	97.64%	97.99%	98.84%	93.92%
DET	98.82%	99.06%	94.52%	83.99%
INTJ	87.50%	70.00%	0.00%	0.00%
NOUN	99.58%	99.26%	95.47%	95.27%
NUM	97.06%	99.62%	98.21%	82.71%
PART	97.14%	95.12%	79.45%	84.47%
PRON	99.62%	99.44%	93.69%	91.09%
PROPN	95.88%	98.12%	84.07%	96.87%
PUNCT	99.96%	99.98%	99.59%	100.00%
SCONJ	98.68%	98.40%	87.91%	94.97%
SYM	50.00%	25.00%	0.00%	0.00%
VERB	99.72%	99.72%	95.42%	96.01%
X	95.62 %	91.91%	82.76%	72.73%

Table 25: BERT precision and recall per POS-tag per test set.

POS-tag	Modern		Historical	
	Precision	Recall	Precision	Recall
ADJ	97.25%	97.71%	81.33%	84.57%
ADP	99.46%	99.74%	99.58%	98.99%
ADV	95.59%	95.33%	86.83%	85.80%
AUX	91.67%	95.60%	85.02%	86.31%
CCONJ	96.17%	95.60%	97.20%	95.76%
DET	98.44%	96.93%	95.85%	74.94%
INTJ	46.15%	60.00%	0.00%	0.00%
NOUN	98.23%	98.04%	89.16%	91.16%
NUM	98.04%	94.34%	97.98%	72.93%
PART	93.49%	92.42%	78.39%	75.73%
PRON	99.05%	98.31%	90.66%	86.53%
PROPN	91.30%	94.09%	79.01%	86.20%
PUNCT	99.95%	99.95%	100.00%	100.00%
SCONJ	96.49%	96.21%	86.73%	91.96%
SYM	100.00%	25.00%	0.00%	0.00%
VERB	97.96%	97.43%	91.67%	92.73%
X	89.33%	86.73%	63.16%	54.55%

Table 26: Marmot precision and recall per POS-tag per test set.

POS-tag	Modern		Historical	
	Precision	Recall	Precision	Recall
ADJ	98.17%	98.99%	88.87%	93.85%
ADP	99.46%	99.91%	99.58%	99.07%
ADV	94.58%	96.06%	91.16%	88.46%
AUX	95.44%	97.14%	84.70%	86.31%
CCONJ	95.47%	96.17%	98.14%	97.24%
DET	98.00%	98.47%	94.49%	79.58%
INTJ	100.00%	50.00%	0.00%	0.00%
NOUN	99.17%	98.70%	95.15%	93.44%
NUM	98.48%	98.11%	97.25%	79.70%
PART	95.01%	90.97%	93.33%	74.76%
PRON	98.63%	98.87%	90.08%	91.68%
PROPN	94.14%	96.51%	79.07%	91.89%
PUNCT	99.95%	99.95%	99.59%	100.00%
SCONJ	95.86%	94.61%	86.30%	94.97%
SYM	100.00%	25.00%	0.00%	0.00%
VERB	99.20%	98.66%	93.73%	94.15%
X	93.53%	93.53%	73.58%	59.09%

Table 27: Stanza precision and recall per POS-tag per test set.

POS-tag	Modern		Historical	
	Precision	Recall	Precision	Recall
ADJ	83.86%	91.58%	66.73%	75.08%
ADP	96.65%	98.89%	96.18%	97.64%
ADV	79.69%	75.89%	75.09%	64.20%
AUX	87.67%	82.98%	84.08%	78.33%
CCONJ	93.72%	87.14%	95.58%	95.58%
DET	94.64%	72.96%	94.58%	44.55%
INTJ	0.00%	0.00%	0.00%	0.00%
NOUN	90.82%	92.91%	80.08%	82.56%
NUM	73.62%	70.57%	77.00%	57.89%
PART	89.62%	80.69%	83.77%	62.62%
PRON	94.35%	94.06%	85.54%	81.98%
PROPN	83.40%	92.29%	69.77%	92.60%
PUNCT	99.93%	99.72%	100.00%	99.92%
SCONJ	88.62%	63.56%	83.81%	44.22%
SYM	0.00%	0.00%	0.00%	0.00%
VERB	89.46%	88.86%	81.64%	86.35%
X	52.24%	41.42%	56.60%	45.45%

Table 28: UD Cloud tagger precision and recall per POS-tag per test set.

D National Corpus of Polish vocabulary comparison output

Aside from counts and proportions of the tested vocabulary that could not be found in the National Corpus of Polish, specific tokens and lemmas were returned. The items listed below were not found in the selected subsection of the National Corpus of Polish. They are separated by whitespace.

- **PDB lemmas:** !) 19:15 25-procentowy 642-65-85 9-miesięczny Arasyb Bielsko-Biała Bushill-Matthews Collridge Eija-Riitta HA-Il Hawełko III-1 IRSC IRSR Instagram Lunzie McMillan-Scott Minecraft PPE-DE Palmiak Stallarholmen Winfryd-Bonifacy Yeosol [ajtemik antysubsydyjny bezfabularny bio-obrazowanie ciemku krio-elektronowy lajwik merozoit niekonscjentywny non-profit nudności odmaterializować podwaliny przeciwbiałaczkowy przeciwretrowirusowy tekstilandia trichlorobenzen Ździara
- **PDB tokens:** ! “ % ’ ” ( ) , 19:15 25-procentowy 5-proc 642-65-85 9-miesięczna ; Ajtemików Bushill-Matthewsowi Collridge Eija-Riitta HA-Il III-1 IRSC IRSR Instagramie Kalkilli Lunzie Maggego McMillan-Scott Minecrafcie PPE-DE Palmiak Pirkera Stallarholmen Winfryd-Bonifacy Yeosol [ ] ajtemika bio-obrazowania celekoksybem efawirenzem krio-elektronową lajwika non-profit nukleozydowymi przeciwciałem przeciwretrowirusowego ry(d)zykować sakwinawiru tektilandia trichlorobenzenu tuńczykowymi zięciowskim – — ” ”
- **memoir lemmas:** !) Asińdźka Bludniki Będowszczyazna Bęklowizna Cobary Czołhany Dochorów Dorchów Dłużanin Głowecki Kmińszczyzna Kurypów Lesniowice Muczynowska Nawarya Notiak Pierściorowski Pokasowce Ronantowizna Ruszkowizna Rypnin Rzotoławski Semiginów Siemginów Siemiginów Siemignów Strużewo Stryiskie Swieżaska Szołayska Temerowice Treterówna Treterówną Zebold abbum adlinencja assekuracya całorolny cwansiger daruju domnikalny dotacya dośmierć excentryczność generacya gymnazyum instantacya ioyciec jurysdykcya juryzdyksya kadectwo mandatariat mandataryusz mandatyrusz mojomu mortyfikować mychayłowu niepomiąć nieprzynieść obeymować obeyście ordynarya oycowski pełnłnomocnik przystoyna półgrunt półrolny rarachować separacya spaśne stayermarka submittować sukcessor sukcessorka sukcessya sukcesya szambelanic szyzmatycki treterianum ukochanomu warżyć wdokument świętej pamięci Żółtowizna
- **memoir tokens:** ! ( ) , Abbum Adelunia Asińdźka Badenianką Bełszowcu Bełzkiem Bienkowskey Blizcey Bludnickey Bludnikach Bludnikami Bludniki Bodzowcu Borkoscy Bołszowcu Bołszowice Bołszowickim Będowszczyazna Bęklowizna Chyrowskiey Cobary Czołhanach Czołhany Dobrrzyńskiej Dochorowie Dominikalnym Domnikalnego Dorchów Dołputowie Dołputów Dziduszyckiego Dłużanie Dźurkowie Floyrana Galecyi Galicyiskiego Golejowskiemi Golejowskimi Gwoździu Głoweckiego Głuską Helnkę Horodzyńskiego Inżyniryi Jabłonoscy Jenerałówną Jędrżejowicz Kazimierzostwie Kleofasę Kmińszczyzna Knihinicz Knihiniczach Komornikostwa Komornikowej Kopestyńskich Koropacza Korytyńską Koziobrodzkiego Kołmyiskim Kołomyiskimc Kruszelnicy Kruszelnicę Krzywczas Kunaszowa Kunaszowie Kurypów Kutyszcza Kutyszczach Leboskich Lesniowic Luboscy Maksymowic Mandatariaty Mandataryusz Mandataryusza Mandatyrusza Michałoskich Mohorocie Muczynowską Mychayłowu Nawaryi Naybliższe Nieograniczały Niezabitoski Notiak Obertyńskim Ocyciec Ostaszeskiey Ostaszewskigo Oyciciec Oycowskiey Perekozach Pierściorowskim Pieścioroski Pieściorowski Pokasowce Poldzię Przebysławscy Przebysławski Przebysławskimi Puszczanki Puzyniance Rafałoska Rafałoskiej Romaszówki Ronantowizna Rudźwianach Rusoccy Ruszczewskich Ruszkowizna Rypnin Rzotoławskim Sadogurze

Semiginów Separacya Siemginowa Siemiginowa Siemiginowie Siemiginów Siemignowa Simiginowskiego Sopohów Stamisaw Stojoskiego Strużewo Stryiskim Sukcesya Swieżaski Swieżaskiey Swojey Szamanowskim Szołayska Szołayskiego Tarmowiecki Tarnoscy Temerowiec Tomaszowiec Treterianum Treterowej Treterówną Wincentowey Woyniłowa Woyniłowie Woyniłowskich Woyniłów Wołczyńcu Zabilską Załęskiey Zebold adlinencjami administracyę ambicya arędujący assekuracyi asystencyą austyackie balożowaniu bronzowemi całorolni ciaśnieysze ciemnoblnd ciepłey cwansigera cywilney daruju delkatnieyszey domurowanego dotacyi doyrzałemi drugey drżew dway dwukoleśnym dystynkcy ekwipazach etykietalney excentryczności exkluzyi familiinemi foryszyc generacya generacyi gołey gymnazyum główney iOycu installacyę instantacyi jadałney jednająca jedwabney jendory jurysdykcyi juryzdyksye kadectwo kląby kochającey kollokacyi kompano kompleksya kompleksyi koniaż krzewowe kunaszowskie kupsze mandataryusza miarkmi mieycus mieyscowych mojomu mortyfikował najzabawney natarczywiey nawykręcawszy naybliższych nayjstarszego nayjświętszey naymował naymłodszą naymłodzy naypięknieyszey nayskładniey naystarsza naystarszych nayszczęśliwsze naywiększey nayznacznieyszą nibieszczeją niechwytało nieciosanemi niemiano nienosiło nieodjechał niepomięła nieprzyniósł niepsuło nierobiły niespokóy niespotkałem nieusuwały niewinem niezapalają niezostawili nieśniło normalney obedwie obeymował obeyrzeć obeyściu obliwamy obliwać obruciły oczem odchuchano okoloney ordynaryi osypami owalney ożewionemi perswazye pełnłnomocnik piurami piętnastulaty podedworem podeyrzenie policyu pomarłemi porozpuszczały poselskiey pospolitey powiena powiązanemi pruchniało prużnował przedewsią przepiurek przerodni przypiórki przystoyna próżnemi psipiólki póyść półgruntów półrolni rantuchem rarachować redukcyi roumianey ruwieśnika skończoney skrzętnem spaśnemi stancyę starszey stayermarka submittować sukcessorka sukcessorów sukcessyi swojey szambelanica szczupleyszey szczupłey szkrofuliczny szutrowany szyzmatycki sądowey teraźnieysze traktryerni troygiem ukochanomu uprzężone uprzężonym urzędowey warżenia wekslarek wojażerowi wsąsiedztwo wuyem wypiętnowaną wywruż wzorowem zachorowałł zaprzęgył zatabaczone zatrętwiał zawerbował zaymie zaymował zuchwałey zżadka Łomnickiey Łosiówną ładąn Świeżyńską śmieszney śniadey świcząc Żółtowizna – ” „