Manually Annotated Corpus of Polish Texts Published between 1830 and 1918

Witold Kieraś, Marcin Woliński
Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warszawa, Poland
wkieras@ipipan.waw.pl, wolinski@ipipan.waw.pl

Abstract

The paper presents a manually annotated 625,000 tokens large historical corpus of Polish. The corpus consists of samples from texts published between 1830 and 1918 — fiction, drama, popular science, essays and newspapers of the period. The corpus provides three layers: transliteration, transcription and morphosyntactic annotation. The annotation process as well as the corpus itself are described in detail in the paper.

Keywords: historical corpora, manually annotated corpora, corpus linguistics, Polish

1. About the project

The paper presents a historical manually annotated corpus of Polish. The corpus consists of samples excerpted from texts published between 1830 and 1918 and is morphosyntactically annotated for the purpose of a larger project aimed at creating a diachronic model of Polish inflection. Together with two other manually annotated corpora (one historical and one contemporary) it will serve as a point of reference for a corpus-driven research in diachronic computational morphology of Polish.

2. Related work

The annotation of the presented corpus took place in parallel with the Baroque corpus of Polish project in which a manual annotation of a gold-standard dataset was also conducted (Bronikowska et al., 2016). The two tasks shared the same web application developed specially for both projects and kept close cooperation in many details. The Baroque corpus covers a 1601-1772 time span leaving over half a century gap between the two projects, which will be hopefully filled in future projects. We are not aware of existence of any other manually annotated historical corpora of Polish.

Among resources for other Slavonic languages a relatively similar project was accomplished for Slovene (Erjavec, 2012; Erjavec, 2015) where a 300,000 tokens large corpus of historical texts was manually annotated and used as gold-standard dataset for automatic annotation of a larger collection of texts.

3. Source Data

We are using an existing collection of samples excerpted from Polish texts published between 1830 and 1918 for the purpose of researching historical inflection and spelling (Bilińska et al., 2016). In literary studies and historical linguistics the period represents the second half of the so called New Polish development stage of the language. The time span of the corpus marks important dates in Polish history having significant impact on social, cultural and political changes which subsequently influenced literary and linguistic developments. Especially year 1918 is considered a turning point in the history of the whole Central-European region and assumed to be the actual end of the 19th century.

The collection consists of 1000 samples of ca. 1000 words each, thus the whole collection is ca. 1 million words large. We will refer to it as F19-1M for short. The samples of the corpus are divided into five separate subcorpora of equal size representing the following functional styles: fiction, essays, science and popular science, short newspaper articles, drama. The samples were excerpted mostly from scans of original first editions of texts stored in digital libraries. They were carefully transliterated with regard to historical spelling rules, including misspelled words in the original editions.

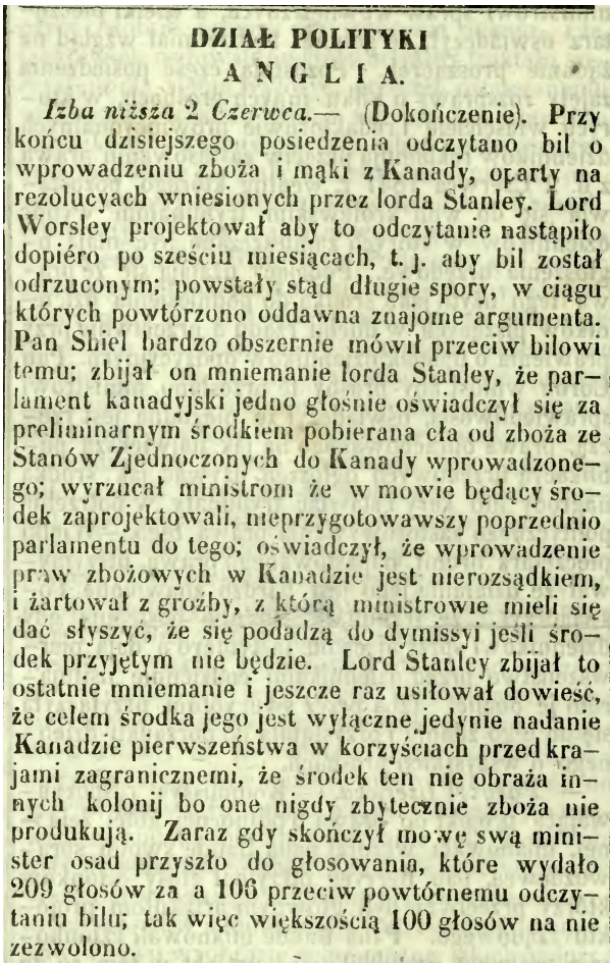
Samples in F19-1M are also evenly distributed between years: for every year and every stylistic subcorpus there is at least one and up to four samples, with an average of 11.24 (standard deviation 1.4) samples per year in the whole corpus. Every sample is accompanied by metadata and source files (scans). An example of source scan is shown in Figure 1.

Although the corpus represents all major Polish literary centres in all five stylistic subcorpora, a bias towards the capital city is significant as nearly 40% of sampled texts were published in Warsaw. Other major publishing centres are Lviv (16%), Cracow (12%), Poznań (7%) and Vilnius (5%), all inhabited by a dominant Polish speaking community at that time. However, the corpus also represents an important part of Polish literary activity performed in exile as about 6% of sampled texts were first published in Paris. In total, publications from over 70 different towns where included in the corpus.

4. Preprocessing

F19-1M is available as a collection of plain text files. Since our goal is to manually annotate only about half of it, we have decided to excerpt 3125 shorter samples of ca. 160 words each. This means that from each F19-1M sample we have extracted three smaller samples for our manually annotated corpus.

Before annotators can start their work, the samples need to be transcribed to modernised spelling and morphologically analysed to provide possible inflectional interpretations of each token. The processing described in this section takes place in Anotatornia web application (Woliński et al., 2017) which then serves the processed samples to annotators.



DZIAŁ POLITYKI
ANGLIA.

Izba niższa 2 Czerwca.— (Dokończenie). Przy końcu dzisiejszego posiedzenia odczytano bil o wprowadzeniu zboża i mąki z Kanady, oparty na rezolucyach wniesionych przez lorda Stanley. Lord Worsley projektował aby to odczytanie nastąpiło dopiéro po sześciu miesiącach, t. j. aby bil został odrzuconym; powstały stąd długie spory, w ciągu których powtórzono oddawna znajome argumenta. Pan Shiel bardzo obszernie mówił przeciw bilowi temu; zbijał on mniemanie lorda Stanley, że parlament kanadyjski jedno głośnie oświadczył się za preliminarnym środkiem pobierana cła od zboża ze Stanów Zjednoczonych do Kanady wprowadzonego; wyrzucał ministrom że w mowie będący środek zaprojektowali, nieprzygotowawszy poprzednio parlamentu do tego; oświadczył, że wprowadzenie praw zbożowych w Kanadzie jest nierozsądkiem, i żartował z groźby, z którą ministrowie mieli się dać słyszyć, że się podadzą do dymissyi jeśli środek przyjętym nie będzie. Lord Stanley zbijał to ostatnie mniemanie i jeszcze raz usiłował dowieść, że celem środka jego jest wyłączne jedynie nadanie Kanadzie pierwszeństwa w korzyściach przed krajami zagranicznemi, że środek ten nie obraża innych kolonij bo one nigdy zbytecznie zboża nie produkują. Zaraz gdy skończył mowę swą minister osad przyszło do głosowania, które wydało 209 głosów za a 106 przeciw powtórnemu odczytaniu bilu; tak więc większością 100 głosów na nie zezwolono.

Figure 1: A fragment of an original 1843 daily newspaper. The exact same fragment is annotated in Figure 2.

4.1. Transcription

Historical texts exhibit significant orthographic variation. For example, the word *komisja* ('commission') appears in F19-1M in the following spellings: *komisja, kommisja, komissja, kommissja, komisya, kommisya, komissya, kommissya, komisyja*. The variation can be coped with by means of transcription, which needs to be done automatically to avoid laborious and time-consuming manual editing. For the purpose of transcription, we use the converter created in the IMPACT project – a rule-based tool[1] (Kresa and Szafran, 2013) for substituting letters or sequences of letters based on the context in which they appear. The procedure itself is simple, but it requires building a relatively large set of rules which are created manually. The number of rules exceeds 3000 and carries not only matching patterns but also a list of exceptions for every rule. Fortunately, a large part of the rules created for the Baroque Corpus of Polish (Bronikowska et al., 2016) could be reused for transcription of 19th c. texts.

4.2. Morphological Analysis

Subsequently, the transcribed samples need to be processed by morphological analyser which provides possible morphological interpretations for text tokens. The annotators may choose one of them but are also allowed to provide their own interpretations in case none of the analyser's answers is correct or the token is unknown to the analyser. The obvious choice for morphological analysis is Morfeusz 2 (Woliński, 2014), the most widely used in Polish NLP and highly configurable analyser. It allows to customise all linguistically sensitive parts of the analysis: inflectional dictionary, tokenisation and tagset.

The linguistic basis of Morfeusz is *Grammatical Dictionary of Polish* (Saloni et al., 2015; Woliński and Kieraś, 2016), consisting of over 330,000 lexemes and nearly seven million word forms, which makes it the largest and most widely used inflectional data source for Polish. SGJP covers the whole list of entries taken from the largest general purpose dictionary of Polish (Doroszewski, 1958–1969) printed on paper in 11 volumes. Doroszewski's dictionary consists of 125,000 lexical entries including a substantial range of archaic, obsolete and dialectal words. Its extensive lexical basis goes back to even last decades of 18th century vocabulary which makes it a perfect lexical source for morphological analysis of 19th century texts. In fact, the extensive coverage of archaic vocabulary in SGJP is usually a curse when processing contemporary data, but in the case of 19th c. it is actually a blessing. Thus the morphological dictionary needed very few lexical additions to cover 19th century vocabulary of the corpus.

SGJP's linguistic data needed to be additionally modified ("aged") in order to cover regular inflectional phenomena non-existing in contemporary Polish but widespread in 19th c. texts. For example, plural instrumental forms of adjectives ending in *-emi* rather then *-ymi* (e.g. *wielkiemi* vs. the only possible contemporary form *wielkimi* 'large'); singular instrumental and locative forms of neuter gender ending in *-em* as opposed to masculine ending *-ym*; plural nominal and accusative forms of some nouns could take an *-a* ending (of Latin origin) instead of contemporary *-y* (*traktata* vs. *traktaty* 'treaties') etc.

Moreover, the analyser's segmentation rules needed to be adjusted to historical joint and disjoint spellings which were significantly different than contemporary ones.

Only 1.72% of tokens in our data did not receive any interpretation from the analyser. This rate is only marginally higher then in the case of contemporary analyser applied to contemporary texts, which proves that the modified analyser performs well.

4.3. Tagging

As described in Section 5, an automatic tagger is being used in the process of manual annotation to simulate one of the annotators simultaneously annotating each sample. Since no training data for tagging 19th c. or any other historical data set for Polish is available, we use a standard contemporary manually annotated corpus of Polish (Przepiórkowski et al., 2012) to train a stochastic tagger (Waszczuk, 2012). The data was only converted to comply to the tagset designed for 19th c. project (to the possible extent as not all grammatical phenomena annotated in the 19th c. project were also annotated in the contemporary corpus). The stochastic model therefore represents only a rough approx-

[1] https://bitbucket.org/jsbien/pol

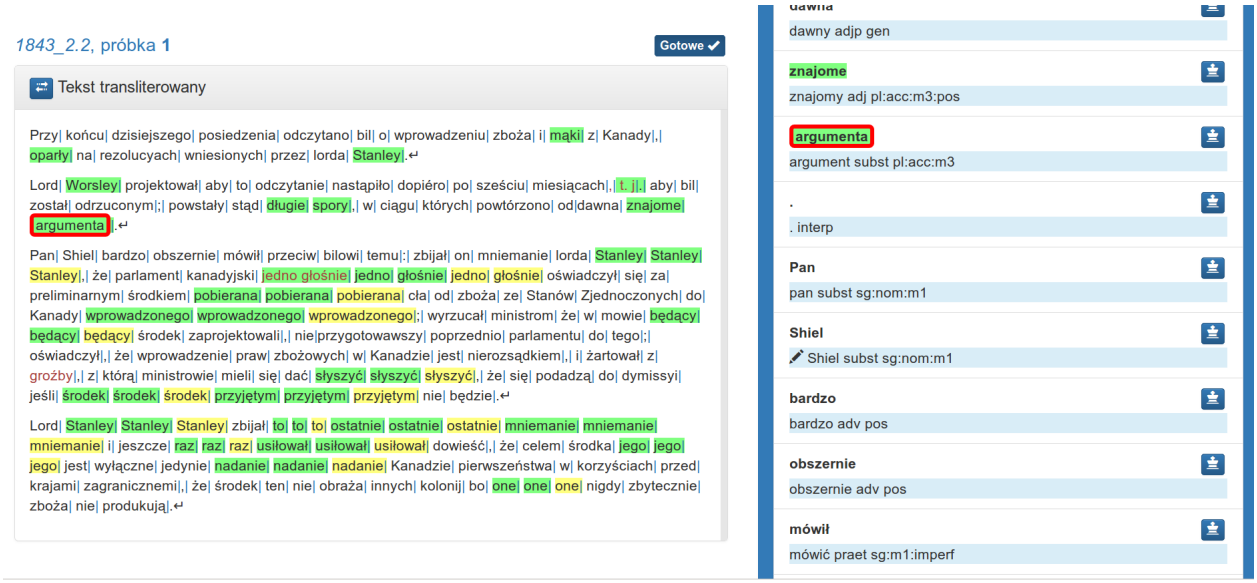


Figure 2: Anotatornia as seen by adjudicator reviewing conflicts between human annotator and tagger. Left hand part of the window shows a running text with annotation discrepancies highlighted. The right hand side shows a list of interpretations provided by an annotator and the tagger. Conflicts between the two are marked in green. The adjudicator can choose one of the two provided interpretations, choose her own from interpretations provided by the morphological analyser, or introduce her interpretation manually.

imation of 19th c. morphosyntax and is not expected to perform flawlessly, but it is expected to be able to handle standard grammatical phenomena such as case, gender and number agreements within phrases. This should be sufficient to pick up simple errors made by human annotators.

During the annotation process the tagger's model is periodically, incrementally retrained together with newly annotated data to improve its performance in the further course of annotation.

5. Annotation

The process of manual annotation of the corpus was conducted in a multi-access web application called Anotatornia (Woliński et al., 2017), which was developed to suite simultaneously two projects devoted to manual annotation of historical Polish text. The other project is the so called Baroque corpus (Kieraś et al., 2017). Thus Anotatornia is focused on satisfying the needs of historical data annotation. It operates on text represented in two layers: transliterated and transcribed.

The annotation is conducted in a hybrid mode conjoining manual work of a qualified linguist and automatic tagging followed by additional verification by human adjudicator ("super-annotator"). Each sample is automatically tagged, but the results of tagging are not disclosed to the annotator on any stage of the process. The annotator can only see possible interpretations provided by the morphological analyser and needs to choose one of them or create her own in case of misinterpreted or unknown tokens. The annotator's choices are then confronted with those made by the tagger and conflicting tokens are highlighted to the annotator, but only her own decisions are shown. This way the annotator is encouraged to check her work for simple mistakes but not tempted to switch to the tagger's version.

After this additional verification, an adjudicator steps in to revise and resolve any remaining conflicting decisions between human annotator and tagger (cf. fig. 2). Adjudicator's work consists mainly of choosing between two possibilities, but she is also allowed to introduce her own interpretation different from those selected by annotator and tagger. Although it is possible that no conflicting decisions between annotator and tagger occur and adjudicator's intervention would not be necessary, in practice every sample in the corpus was additionally reviewed by adjudicators, since the accuracy of the automatic tagger trained on the contemporary data does not exceed 90%.

The annotation process was conducted by a team of nine people, each of them specializing in Polish linguistics with either master's or doctoral degree in the field. Most of them have an extensive experience in various annotation tasks in previous projects. The four most experienced and most active annotators worked also as adjudicators. The possibility of adjudicating one's own conflicts was excluded. The annotators followed a detailed annotation manual. Specific problems were resolved using a mailing list.

Based on the final version of the annotated corpus, the annotation process as described above generates a 14.27% conflict rate between human annotator and tagger. As expected, a large majority of the conflicts are resolved in favour of human annotators (87.22%) but a significant number of human errors are also found and corrected as the remaining 12.78% have been either resolved in favour of the tagger (6.69%) or changed to an alternative interpretation provided by the adjudicator (6.09%).

As a result of the annotation process, 2944 samples were annotated by one human annotator and tagger. Each sample required additional verification by the adjudicator as the situation of full agreement between human annotator and tagger hasn't occurred even once. The number of conflicts ranged between 8 and 152 (large number of conflicts usually involved serious segmentation problems) with an average of 32.34 per sample (median 30). Thus, in the hybrid annotation mode presented above adjudicator's workload is significantly higher comparing to the standard annotating situation with two human annotators followed by adjudicator. On the other hand, time and financial cost of the whole annotation process performed in the hybrid mode drops radically nearly to the level of single annotator mode without the necessity of complete abandoning of any additional quality control. We believe that the hybrid annotation proved to be useful and would apply the same strategy in future projects.
The total number of 625,000 tokens were annotated in the project.

6. Tagset

The tagset of the presented corpus generally reflects ideas behind the tagset of the National Corpus of Polish (NCP) as well as the one used by Morfeusz morphological analyser. The two are similar, yet not identical. The crucial difference between them concerns the grammatical gender. The 19th century tagset basically follows Morfeusz's tagset, however some minor differences were introduced motivated mainly by the Baroque tagset and the desire for basic coherence of the two historical tagsets.
In Polish historical linguistics it is assumed that the main morphosyntactic processes are over by 1830 and Polish morphosyntax of the presented period is basically the same as the contemporary one. The linguistic differences are reflected in phonetics, vocabulary, surface morphology and word order but they do not affect the morphosyntactic tagset. However, some useful extensions of the contemporary tagset were introduced to ease some corpus linguistic research. The 19th century tagset as well as the Baroque one marks the auxiliary verbs of pluperfect tense and future imperfective tense assigning different tags to past and future forms of BYĆ ('to be') verb in constructions of those tenses than in the case of other syntactic constructions (such as passive voice). This will allow to track the decline of pluperfect tense in Polish from early 17th to early 20th century as well as the variability of future imperfective tense construction according to word order (*będę robić* vs. *robić będę* 'I will do') and the use of either infinitive or past verb forms (*będę robić* vs. *będę robił* 'I will do').
Another minor difference between both historical tagsets and the contemporary one is the introduction of numeral adjectives and adverbs as separate parts of speech. Grammatically, numeral adjectives and adverbs share the same features as regular adjectives and adverbs and thus they are not distinguished in the contemporary data, but were introduced to comply to the traditional diachronic description of Polish.
Regardless of those similarities between historical tagsets and their differences comparing to the contemporary one, the 19th century tagset far more resembles the latter. The Baroque tagset needs to cover a much longer time span and the range of morphosyntactic phenomena involved is much more extensive, therefore it is significantly more complex.

	19c 625k	**NKJP 1.2M**
CRF	90.48%	91.44%
bi-LSTM	93.38%	95.28%

Table 1: 10-fold cross-validation accuracy results for two taggers based on 19th and contemporary training data.

7. Usage

Manually annotated resources typically serve as training data for machine learning applications. In the case of the presented corpus, so far two stochastic taggers representing different methodologies, namely conditional random fields (Waszczuk, 2012) and bi-LSTM neural networks (Krasnowska-Kieraś, 2017), were trained and evaluated based on the historical data. As shown in Table 1, the taggers obtained slightly lower results comparing to their evaluation on the manually annotated 1.2 mln subcorpus of the National Corpus of Polish which is commonly used as a gold-standard dataset for contemporary Polish. The results however are not fully comparable since both corpora were annotated with slightly different tagsets. The size of the training data is also relevant to the results as the 19th c. corpus is twice smaller than the manual subcorpus of NCP.
Another NLP tool scheduled to be built in near future using a manually annotated corpus is an automatic transcriber not relying on manually crafted rules. Our rule-based transcriber has yielded good results, but the set of rules became relatively large, which makes them hard to manage and causes some efficiency difficulties. Since corpus annotators were also required to correct transcription errors, the resulting corpus can serve as training data for a machine learning application which can possibly obtain better or at least as good results as the rule-based transcriber, while being more computationally efficient.

8. Conclusions and Future Work

The current project does not cover a task of building a large, automatically annotated corpus of 1830-1918 period or any other large corpora, however the corpus described in this article as well as tools used for its annotation provide a sufficient technical environment for building an extensive corpus of 19th c. Polish. Although strictly philological problems such as text acquisition, balancedness etc. are still open to the future creators of the corpus, the technical leg of the potential project is ready to use. A proof of concept of such corpus will be built based on the tools mentioned above and publicly available collections of historical texts such as Wikisource.[2]
The final version of the manually annotated corpus is publicly available both as a collection of TEI XML files and as a searchable web-based resource.[3] The later is based on MTAS search engine (Brouwer et al., 2017), which allows

[2] https://pl.wikisource.org/
[3] http://korpus19.nlp.ipipan.waw.pl

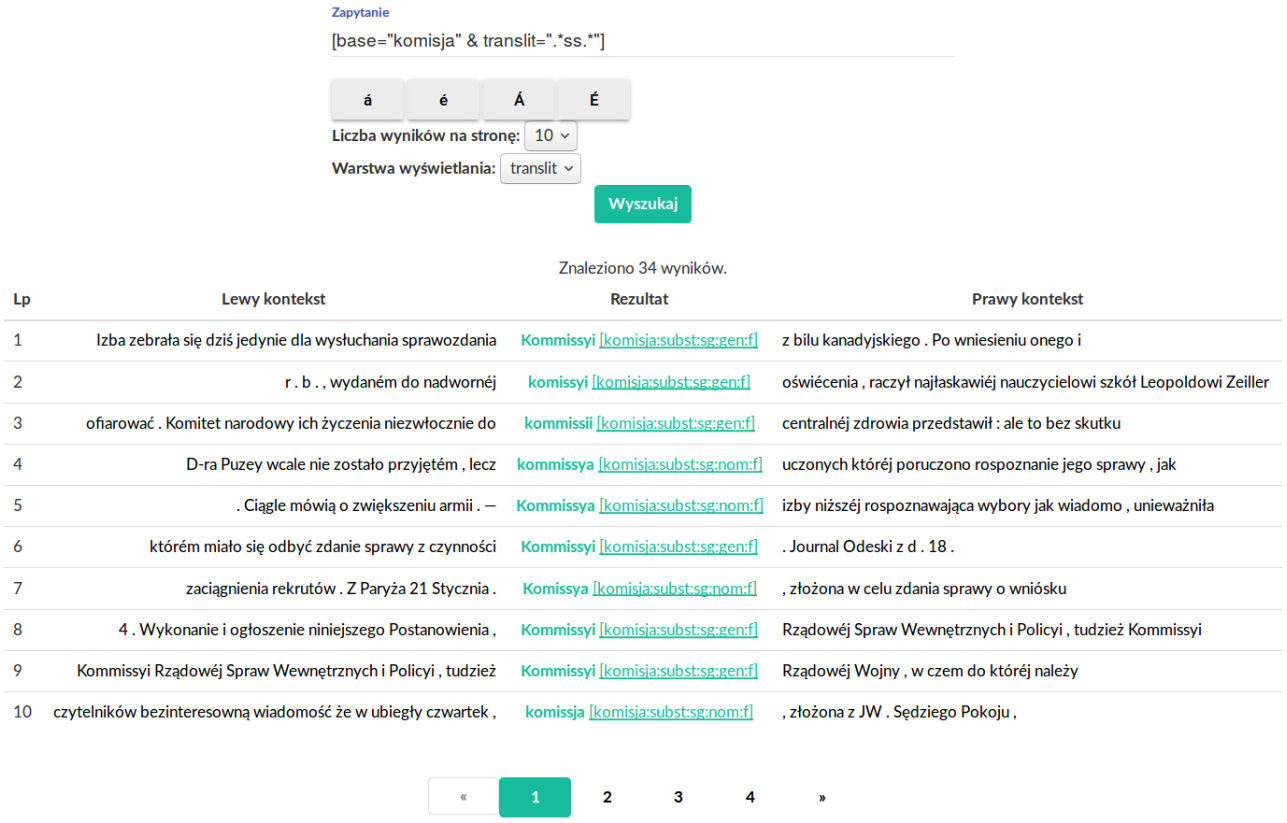


Figure 3: A corpus search query in which all occurrences of a noun *komisja* 'commission' are found with restriction only to those spelled with double *s* in original transliterated document.

to refer to both transcription and transliteration text layers of the corpus, as well as to morphosyntactic annotation layer. Conditions referring to various layers may be combined in one corpus query, so for example all occurrences of a word lemmatized as *komisja* 'commision' restricted to only those spelled originally with double *s* can be found. The query together with several first hits can be seen in Figure 3.

9. Acknowledgements

The work being reported was financed by a National Science Centre, Poland grant DEC-2014/15/B/HS2/03119.

10. Bibliographical References

Bilińska, J., Derwojedowa, M., Kieraś, W., and Kwiecień, M. (2016). Mikrokorpus polszczyzny 1830-1918. *Komunikacja specjalistyczna*, 11:149–161.

Bronikowska, R., Gruszczyński, W., Ogrodniczuk, M., and Woliński, M. (2016). The Use of Electronic Historical Dictionary Data in Corpus Design. *Studies in Polish Linguistics*, 11(2):47–56.

Brouwer, M., Brugman, H., and Kemps-Snijders, M. (2017). MTAS: A Solr/Lucene based Multi Tier Annotation Search solution. In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure*, number 136, pages 19–37. Linköping University Electronic Press, Linköpings universitet.

Doroszewski, W. editor. (1958–1969). *Słownik języka polskiego PAN*. Wiedza Powszechna – PWN.

Erjavec, T. (2012). The goo300k corpus of historical slovene. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Erjavec, T. (2015). The imp historical slovene language resources. *Lang. Resour. Eval.*, 49(3):753–775, September.

Kieraś, W., Komosińska, D., Modrzejewski, E., and Woliński, M. (2017). Morphosyntactic Annotation of Historical Texts. The Making of the Baroque Corpus of Polish. In *International Conference on Text, Speech, and Dialogue*, pages 308–316. Springer, Cham.

Krasnowska-Kieraś, K. (2017). Morphosyntactic disambiguation for Polish with bi-LSTM neural networks. In *Proceedings of 8th Language & Technology Conference*, pages 367–371. Poznań, Poland.

Kresa, M. and Szafran, K. (2013). Przykład nowego zastosowania słownika polszczyzny historycznej. *Prace Filologiczne*, LXIV:159–171.

Przepiórkowski, A., Bańko, M., Górski, R.L., and Lewandowska-Tomaszczyk, B., editors. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Saloni, Z., Gruszczyński, W., Woliński, M., Wołosz, R.,

and Skowrońska, D. (2015). *Słownik gramatyczny języka polskiego*. 3. edition.

Waszczuk, J. (2012). Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India.

Woliński, M. and Kieraś, W. (2016). The on-line version of Grammatical Dictionary of Polish. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, pages 2589–2594, Portorož, Slovenia. ELRA, European Language Resources Association (ELRA).

Woliński, M., Kieraś, W., and Komosińska, D. (2017). Anotatornia 2 – an annotation tool geared towards historical corpora. In *Proceedings of 8th Language & Technology Conference*. Poznań, Poland. under review.

Woliński, M. (2014). Morfeusz reloaded. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111, Reykjavík, Iceland. ELRA.

11. Language Resource References

Magdalena Derwojedowa. (2017). *Corpus of Polish Texts of 1830-1918 period*. University of Warsaw, `http://www.f19.uw.edu.pl`.

ZIL IPI PAN. (2014). *Morfeusz 2*. `http://sgjp.pl/morfeusz`, version 2.0.

Zygmunt Saloni and Marcin Woliński and Robert Wołosz and Włodzimierz Gruszczyński and Danuta Skowrońska. (2015). *Grammatical Dictionary of Polish*. `http://sgjp.pl`, version 3.