Studies in Polish Linguistics
vol. 11, 2016, issue 2, pp. 47–56
doi:10.4467/23005920SPL.16.003.4818
www.ejournals.eu/SPL

Renata Bronikowska,
Włodzimierz Gruszczyński,
Maciej Ogrodniczuk,
Marcin Woliński

Polish Academy of Sciences

The Use of Electronic Historical Dictionary Data in Corpus Design

Abstract
The History of the 17th and 18th c. Polish Language Laboratory, Institute of Polish Language, Polish Academy of Sciences, is in the process of creating two large databases: *The Electronic Dictionary of the 17th–18th c. Polish* and *The Electronic Corpus of the 17th and 18th c. Polish Texts (up to 1772)*, the latter in cooperation with the Institute of Computer Science, Polish Academy of Sciences. It is expected that combining these two sets of data will help to achieve the objectives established for both database projects. The present article shows the benefits that the Corpus creators can get from the data gathered in the dictionary, with special emphasis put on the use of grammatical information included in the dictionary entries to design tools for automatic text annotation in the Corpus.

Keywords
text corpus, text annotation, historical dictionary, historical corpus, Old Polish, inflectional analysis

Streszczenie
W Pracowni Historii Języka Polskiego XVII i XVIII w. Instytutu Języka Polskiego Polskiej Akademii Nauk powstają obecnie dwie obszerne bazy danych: *Elektroniczny słownik języka polskiego XVII i XVIII w.* oraz *Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do roku 1772)* – ten ostatni we współpracy z Instytutem Podstaw Informatyki PAN. Połączenie tych dwóch zasobów może pomóc zrealizować cele obu projektów. Niniejszy artykuł przedstawia korzyści, jakie mogą odnieść twórcy korpusu, używając danych słownika, m.in. poprzez wykorzystanie informacji gramatycznej z haseł słownika do budowy narzędzi do automatycznej anotacji tekstu.

Słowa kluczowe
korpus tekstowy, anotacja tekstu, słownik historyczny, korpus historyczny, język staropolski, analiza gramatyczna

1. Project factsheets

The Electronic Dictionary of the 17th–18th c. Polish (Gruszczyński 2004) is a continuation of the paper version of the *Dictionary*, whose first volume was published in 1999–2004 (Siekierska 1999–2004). In 2004, the decision to convert the Dictionary to an electronic form was made, and successive entries have since been made available on the dictionary website (http://sxvii.pl, henceforth e-SXVII). At present, the dictionary contains ca 20K entries at differing stages of development; few of the entries are complete, most of them are still being modified and supplemented, and some have the so-called germ status (i.e. they only include grammatical forms of the lexeme). Grammatical forms within the entries amount to ca 52K. It must be noted that the inflectional paradigms cover only those forms which were found in the examined sources, and thus, in the majority of cases, the paradigms are not complete either. The meanings of described lexemes are illustrated by quotations transliterated from ca 850 sources.

The Electronic Corpus of the 17th and 18th c. Polish Texts (up to 1772) is also referred to as *The Baroque Corpus* (Pl. *korpus barokowy* – hence the acronym KORBA). Its data will eventually amount to 12M tokens (a remarkably large number for any historical corpus). The aim of the Corpus is to supplement the *National Corpus of Polish* (Przepiórkowski et al. 2012; henceforth, NKJP) with a diachronic aspect, crucial for investigation of the history of the Polish language. The Corpus will feature rich structural annotation (e.g. locating a searched expression in the appropriate text page) and text annotation (e.g. tagging foreign languages). Part of the corpus data (0.5M tokens) will be annotated with morphosyntactic tags manually, while the remaining data will be annotated automatically.

2. Using corpora in the lexicographic work and dictionaries in corpus design

The benefits that the Corpus gives to lexicographers are self-evident. Corpus quotations allow one to reconstruct the meaning (or several meanings) of a given lexeme, to find its spelling, phonetic or morphological variations, and to establish its syntactic and lexical collocations. Information on the texts contained in the Corpus provides additional data on the use of the lexeme, including its frequency, chronology, geography, etc., genres in which the lexeme is attested, as well as authors who would often use it. Quotations from the Corpus also serve as an illustration of the word meanings, idioms or collocations.

What distinguishes the Corpus from other textual databases is the morphosyntactic tagging of tokens, which is particularly important in inflectional

languages. As a result of such annotation, individual tokens have been identified as forms of a specific lexeme with inflectional interpretation assigned. This enhances the efficiency of the Corpus search, which can now yield all inflected forms of a given lexeme. Morphosyntactic annotation is usually performed with the help of IT tools, i.e. an inflectional analyser and a tagger. The former assigns all possible interpretations to an analysed grammatical form. It often happens, however, that one form can be interpreted in a number of ways, as a result of homonymy. Thus, the next stage involves using the tagger, to disambiguate grammatical information based on the syntactic and semantic context.

The annotation tools, especially the inflectional analyser, derive data from a dictionary, more specifically, a grammatical dictionary which contains lexemes and their inflectional paradigms. This type of annotation of contemporary Polish lexemes is found in the *Grammatical Dictionary of Polish* (Saloni et al. 2015; henceforth, SGJP), which constitutes the basis for the inflectional analyser, Morfeusz (Woliński 2006), employed in the tagging of the *National Corpus of Polish*.

3. Building a Baroque inflectional analyser with the help of the Baroque dictionary

In order to annotate a historical corpus (e.g. KORBA), the Corpus creators need an inflectional analyser based on a grammatical dictionary of a particular historical period (in this case, the 17th and 18th c. Polish language). Although no such dictionary exists, certain grammatical information contained in the e-SXVII can be used. Figure 1 illustrates such information in an entry BAŃKA 'a type of vessel'.

The first portion of grammatical information directly following the lexeme defines the part of speech (*rzecz* [N]) and, since the word is a noun, its grammatical gender (*ż* [fem.]). Below, there follows a paradigm which contains forms found in textual sources, including their phonetic and morphological variants (*BAŃKA, BANKA*). In this particular case, the grammatical category of number includes three values: besides singular (*lp*) and plural (*lm*), it also indicates a dual number (*lplw*), attested in the 17th c. texts. The paradigm is not complete, e.g. a form of Dat. pl and all dual forms except Nom. are missing. Grammatical forms are also indicated in the quotations, which illustrate the word's usage. Because each source in the Corpus is dated, a form can be linked with the date of its use. It is particularly important in the case of forms which disappeared from use during the period examined. Such information will help to determine the probability of each outcome in the process of automatic inflectional analysis. For instance, at the beginning of the 17th century, the form *bańce* can be interpreted as Nom. Du (last quote), while at the end of that century, it is almost certain that the form represents Dat. or Loc. sg.

Figure 1. Grammatical information in e-SXVII entries

Source:?

Figure 2 presents selected data regarding IT tools supporting the contemporary corpus (NKJP) and the Baroque corpus (KORBA), cf.:

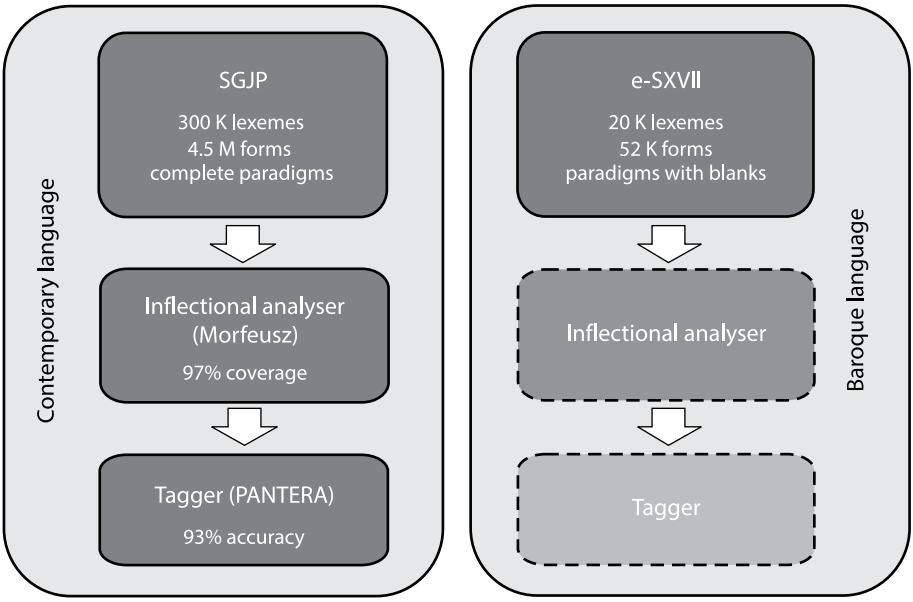

Figure 2. IT tools supporting the contemporary corpus (NKJP) and the Baroque corpus (KORBA)

Source:?

SGJP, which the contemporary inflectional analyser is based on, contains ca 300K lexemes and 4.5M inflected forms. The grammatical paradigms are complete, since lexicographers, who were native speakers of the contemporary Polish language, were able to determine each form. Morfeusz, the inflectional analyser, provides 97% coverage of contemporary texts, i.e. it recognizes 97% of inflected forms in the contemporary texts it analyses (only 3% of the forms are unrecognized). The tagger PANTERA, which disambiguates forms interpreted by Morfeusz, provides 93% accuracy, i.e. 93% of the forms are correctly interpreted. Thus, the IT tools supporting the contemporary corpus of Polish prove to be efficient.

When it comes to the Baroque corpus, designing the tools to support it has proved more difficult because of a weaker database. The e-SXVII, which will serve as a grammatical dictionary of the 17th and 18th c. Polish language, so far has accommodated 20K lexemes and 52K forms. As mentioned above, the paradigms are not complete; they only include forms found in the examined sources, and, since lexicographers themselves are not the native speakers of 17th and 18th c. Polish, they prefer not to reconstruct the old forms. This data shall be used for the development of the Baroque inflectional analyser (Morfeusz 17th) and a Baroque tagger.

Besides the grammatical data from the e-SXVII, Morfeusz 17th will use "aged" SGJP paradigms. In order to increase the reliability of the Baroque inflectional analyser, it will be designed by successive approximations. Figure 3 shows sequential steps leading to its final design. The aim of the first stage will be to create an imperfect, "dirty", version of Morfeusz 17th based on two data sources, i.e. e-SXVII and SGJP. Incomplete paradigms from the e-SXVII will be reconstructed,[1] while contemporary paradigms from SGJP will be modified to include such historical forms of lexemes as the dual number. Then, the "dirty" Morfeusz 17th will be used to analyse a sample of texts from KORBA (0.5M tokens). Subsequently, a group of annotators will manually disambiguate and validate the data. As a result, we shall get a sample of properly annotated Baroque texts. This sample and the reconstructed paradigms will enrich the grammatical part of the e-SXVII, and in turn, the extended version of the baroque dictionary will serve as the database for an improved version of Morfeusz 17th. Ultimately, the new version of the analyser will be used to annotate the whole baroque corpus.

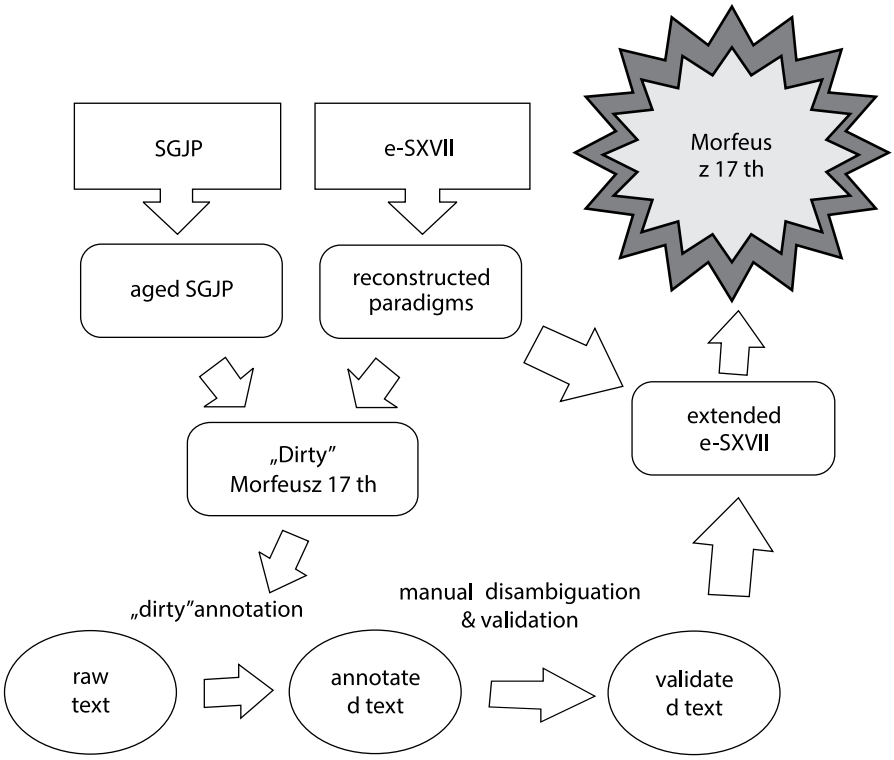

Figure 3. Building the Morfeusz 17th by successive approximations

Source:?

[1] This will be done only in order to build Morfeusz 17th. The decision not to reconstruct old forms is still valid in the design of the Baroque dictionary itself.

4. Morphology goes baroque – "aging" SGJP paradigms

"Aging" the SGJP will mainly consist in adding historical forms to contemporary paradigms. For instance, the paradigms of imperfective verbs will include a perfective participle (e.g. *widziawszy* 'having seen' – in contemporary Polish only perfective verbs take such a form), and adjectival paradigms will see an adjectival predicative form added (e.g. *słuszna* 'it is proper'). Some grammatical categories will be supplemented with additional values. For instance, besides singular and plural, the category of number will also include dual (e.g. *żabie* '(two) frogs'), and the category of tense will gain past perfective (e.g. *widziałem był* 'I had seen'). "Aging" the SGJP may also result in building new entries containing regularly created derivatives with their paradigms, e.g. the comparative forms of active and passive participles (*bolątszy* 'more hurting'), which will then be linked to their appropriate forms in the positive. Yet another means to "age" the SGJP is modification of the category structure, as rendered necessary in the case of gender.

In the contemporary grammatical description accepted in NKJP, there are five grammatical genders, i.e. masculine personal, masculine animate, masculine inanimate, feminine and neuter. In the contemporary corpus, each nominal form of a lexeme will be annotated alike; in other words, all nominal forms will have one gender. In the case of the Baroque corpus, its annotators will often have doubts regarding the gender of nouns (especially in the case of lexemes which did not survive in contemporary Polish). That is why the category of gender will be treated hierarchically (see Figure 4) and the forms will be annotated with a varying degree of accuracy, depending on the lexical material.

For instance, the form *abrysom* is attested in a particular text, but the gender cannot be defined because the ending *-om* in Dat. pl is common to all genders.[2] It cannot be stated with any certainty how it should be lemmatized, i.e. as ABRYS, ABRYSA or ABRYSO. Thus, such a form will be tagged as neutral. The form *abrysu* will be tagged as masculine (and lemmatized as ABRYS), but it cannot be defined more precisely as animate or inanimate. Only *abrys*, i.e. the form in Acc. Sg, is marked for masculine inanimate. Had other forms not existed, all the above-mentioned forms could have been interpreted as the lexeme ABRYS and understood as masculine inanimate. However, Baroque texts also include feminine forms, e.g. *abrysę* (lemmatized as ABRYSA). Consequently, the form *abrysom* can be interpreted either as a masculine inanimate or as a feminine noun.

[2] In the first half of the 17th century, the ending –am, originally characteristic of feminine declination, occurred residually; eventually, all feminine nouns adopted the masculine ending –om.

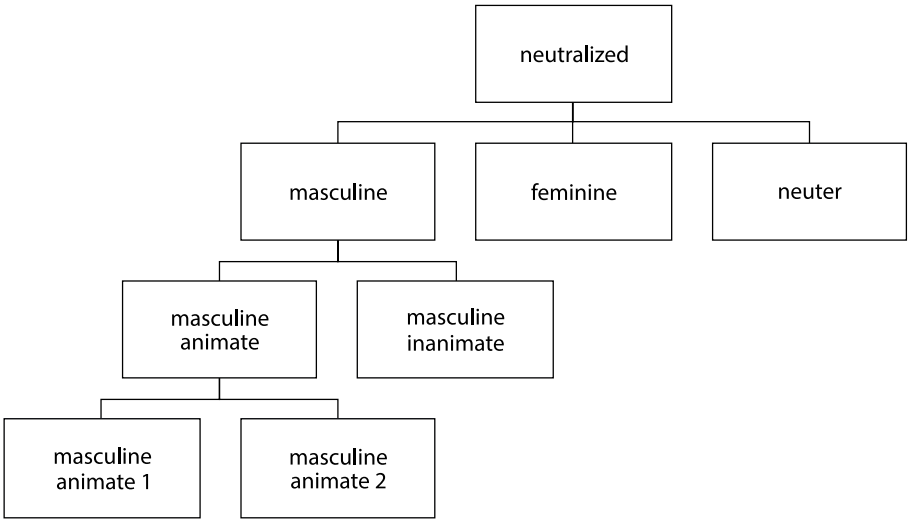

Figure 4. Hierarchy of gender in KORBA

Source:?

5. The experiment – analysing baroque texts with the contemporary Morfeusz

Before creating the first version of Morfeusz 17th, the authors wanted to test how the contemporary Morfeusz managed with Baroque texts, and thus they designed a simple experiment. Of all texts presently gathered in KORBA, a random sample (about 16.5K tokens from 19 texts) was selected and analysed.

Prior to the analysis, all texts were converted from a transliterated to a transcribed version. In order to do that, we used the converter created by Janusz Bień and his team for the IMPACT project[3] (Figure 5 illustrates two versions of the text sample). As a result, the transcribed version became similar to modern texts, e.g. each *a* with a dash was converted to the contemporary *a* without a diacritic mark, or each *y* between spaces was converted to *i*. Additionally, words in sparse print were converted erroneously, because some rules were formulated too rigorously, e.g. a rule "each *s* after *be* and before a voiceless vowel is converted to *z*" was the reason for the wrong conversion from *bestie* to **beztie*.

[3] The converter is available at: https://bitbucket.org/jsbien/pol (accessed December 15, 2015).

Transliterated version	Transcribed version
B. Nie záda mi żaden tey rzeczy, ktoraby bárziey owemu niżeli mnie należeć nie miáłá. A ták śmierć y Kupidyn pozárszy się iáko bestie, posnęli w Kośćiele Bacchusowym, gdźie niewyszumiawszy z przepićia śmierć Cupidinow, á Cupido śmierći należący przypasawszy sáydak do boku szły, swych należących odpráwować powinnośći.	B. Nie zada mi żaden tej rzeczy, któraby barzyej owemu niżeli mnie należeć nie miała. A tak śmierć i Kupidyn pozarszy się jako beztie, posnęli w Kościele Bacchusowym, gdzie niewyszumiawszy z przepicia śmierć Cupidynów, a Cupido śmierci należący przypasawszy sajdak do boku szły, swych należących odprawować powinności.

Figure 5. A text sample before and after conversion to a transcribed version

Source:?

The experiment yielded 1552 unrecognized forms, comprising 9% of the entire sample. Surprisingly, many forms proved to be recognizable (although some were probably misinterpreted), because Morfeusz contains a number of old lexemes and forms, e.g. *słowy* – an archaic form (Instr. sg) of the lexeme SŁOWO 'word'. Some forms were unrecognized due to the converter's mistakes. They will be recognized after the implementation of new rules, such as the segmentation rule, which will separate prepositions from nouns and adjectives (*wtym*), or a conversion rule which will change some words starting with *puł-* to *pół-* 'half' in compliance with contemporary spelling norms. Finally, and as expected, some forms were unrecognized because they disappeared from use and are not included in Morfeusz paradigms, e.g. *moję* – old form (Acc. sg, fem.) of the lexeme MÓJ 'my'. The hope is they will be recognized by the new Morfeusz, which will include modified paradigms and forms exported from the e-SXVII.

6. Conclusion

Morphosyntactic tagging encounters more difficulties when annotating a historical corpus than a contemporary corpus. In the case of morphosyntactic tagging of a Polish Baroque corpus, the lack of a Baroque grammatical dictionary of Polish which could constitute the basis for the Baroque inflectional analyser is the most problematic issue. That is why, in order to build IT tools to support the corpus, it is necessary to combine incomplete data from different sources, such as the grammatical dictionary of the contemporary Polish, and grammatical forms gathered in the Baroque dictionary. Some grammatical information will also be gained in additional manual tagging of part of the corpus.

References

Gruszczyński Włodzimierz (ed.) (2004–). *Elektroniczny słownik języka polskiego XVII i XVIII wieku*. [URL: http://sxvii.pl/; accessed December 15, 2015].

Przepiórkowski Adam, Bańko Mirosław, Górski Rafał L., Lewandowska-Tomaszczyk Barbara (eds.) (2012). *Narodowy Korpus Języka Polskiego*. Warsaw: Wydawnictwo Naukowe PWN. [URL: http://nkjp.pl; accessed December 15, 2015]

Siekierska Krystyna (ed.) (1999–2004). *Słownik języka polskiego XVII i 1. połowy XVIII wieku*. Vol. 1. Kraków: Wydawnictwo Instytutu Języka Polskiego PAN.

Saloni Zygmunt, Woliński Marcin, Wołosz Robert, Gruszczyński Włodzimierz, Skowrońska Danuta (2015). *Słownik gramatyczny języka polskiego*. 3rd ed. Warsaw. [URL: http://sgjp.pl; accessed December 15, 2015]

Woliński Marcin (2006). Morfeusz – a practical tool for the morphological analysis of Polish. In *Intelligent Information Processing and Web Mining, Advances in Soft Computing*. Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, Krzysztof Trojanowski (eds.), 503–512. Berlin: Springer-Verlag.

Renata Bronikowska
Instytut Języka Polskiego
Polska Akademia Nauk
Al. Mickiewicza 31
31-120 Kraków
[r.bronikowska (at) wp.pl]

Włodzimierz Gruszczyński
Instytut Języka Polskiego
Polska Akademia Nauk
Al. Mickiewicza 31
31-120 Kraków
[wlodekiewa (at) poczta.onet.pl]

Maciej Ogrodniczuk
Instytut Podstaw Informatyki
Polska Akademia Nauk
ul. Jana Kazimierza 5
01-248 Warszawa
[maciej.ogrodniczuk (at) gmail.com]

Marcin Woliński
Instytut Podstaw Informatyki
Polska Akademia Nauk
ul. Jana Kazimierza 5
01-248 Warszawa
[wolinski (at) ipipan.waw.pl]