**Notum per ignotum.**

**The enrichment of the lexical information and the corpus resources by using the results of the morphological analysis of historical texts**

**Renata Bronikowska1, Emanuel Modrzejewski2**

1Institute of Polish Language, Polish Academy of Sciences,

Address: al. Mickiewicza 31, 31-120 Kraków

2Institute of Polish Language, Polish Academy of Sciences

Address: al. Mickiewicza 31, 31-120 Kraków

E-mail: r.bronikowska@wp.pl, modrzejewski.emanuel@gmail.com

The History of the 17th and 18th Centuries Polish Language Laboratory, Institute of Polish Language, Polish Academy of Sciences is in the process of creating two large databases, i.e.: *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. Combining the two sets of data might help achieve the objectives established for both projects. At present, the main task of the project is to morphosyntactically semi-manually annotate the 0.5M-segment sample. The manual annotation is based on the automatically generated morphological interpretations proposed by *Korbeusz* – the morphological analyser with the access to the data of the two dictionaries. The results of this analysis not only have a great impact on revising the lexical data of both dictionary and corpus, but also point out the directions of further improvements of corpus tools. The issue that we focus on is what kind of information we can obtain from the tokens that have not been recognised by morphologicial analyser.

**The** res**ources**

*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 was made to convert the Dictionary to an electronic form, 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 different stages of development: few of them 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. When it comes to paradigms, it must be noted that they cover only those forms which were found in the examined sources, and thus, in the majority of cases, the paradigms are not complete. 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 (which is remarkably large for any historical corpus). The aim of the Corpus is to supplement the *National Corpus of Polish* (further NKJP[[1]](#footnote-1), Przepiórkowski et al. 2012) with a diachronic aspect, crucial for investigation of the history of the Polish language. Part of the corpus data (0.5M tokens) will be annotated with morphosyntactic tags manually, with the remaining data will be annotated automatically.

Currently, there are more than 10.5M tokens from about 700 baroque texts – transliterated, annotated structurally (e.g. locating a searched expression on the appropriate text page) and textual (e.g. tagging foreign languages).

The morphological analyser – *Korbeusz* – operates based on two dictionaries: *Grammatical Dictionary of Polish* (further SGJP[[2]](#footnote-2), Saloni et al., 2012), which contains the inflectional paradigms of the contemporary Polish, and *The Electronic Dictionary of the 17th-18th c. Polish*, which contains the lemmas and the paradigms of the 17th-18thcentury Polish (some of them are still partly incomplete). The harvested information is further processed by *Korbeusz*. Incomplete paradigms from the e-SXVII are partly reconstructed, while the contemporary paradigms from SGJP are modified to include such historical forms of lexemes as e.g. the dual number. *Korbeusz* also allows to formulate custom segmentation rules, which enables the analysis of the specific baroque ways of delimitation. For example, contrary to modern spelling rules, particle *nie* may be spelled together with verbal form (including participles) as one word. Prepositions may sometimes be spelled together with the following noun. Particles *-by, -ć, -ż* may attach to forms of virtually any grammatical category.

The process

First, from the whole corpus the 200-token samples are automatically chosen. These “texts” are converted from a transliterated to a transcribed version. In order to do that, we use the converter created by Janusz Bień and his team for the purposes of the IMPACT project[[3]](#footnote-3). The previous formulas of converter have been completed by new ones, based on the 17th-18thcentury Polish orthographic rules. After the conversion, the samples are going to be segmented and morhologically analysed by *Korbeusz*. In such form all samples are ready to be manually annotated; mostly by disambiguation. The list of the tokens, that have not been recognised by the analyser (tagged as *ign* – from Lat. *ignotum* ‘unknown’), is prepared to be checked by the project crew.

The so-called „ign list” containes what lacks in the e-SXVII’s entries: both lemmas and the irregularities in paradigms. Mostly, the irregular forms are on the top of the list, such as abbreviations, irregular inflectional forms (numerals, lemmas with in-built inflection), forms derived from proper names and inhabitant names. These forms are added to the dictionary (more frequent forms are more likely to be added first). After the complementation, the new data is harvested by *Korbeusz*, which makes it possible to give a correct analysis of the forms, that could not be recognised before.

The information from „ign list” is also used by *Korbeusz* in a more direct way; as a research base of new paradigms (or new parts of the available paradigms). For example, searching for analogies pointed out that the segments as *czasiech*, *Prusiech* are the specific realisations of locative pluralis with the characteristic ending *-ech*, so the paradigm should be extended. This list also can be a point of exploration of the words consisting several inflectional segments syntactically combined into one piece, such as *niwczym*, *nizacz*, or a material to create the rules of distribution of the synsemantic segments (particles, agglutinative suffixes).

Commonly, the lack of proper analysis of the segment occurs when there is no specific formula to apply or when the formula is not specified enough and provides superflous changes in transliteration – changing a correct input to invalid, unrecognisable output (i.e. *świetny* na *\*świętny*). To sum up briefly, the „ign list” can be treated as a to-do-list of the improvements of converter’s formulas.

Finally, the „ign list” is a full repository of the forms that were transliterated in a careless manner: typos, superflous delimitation or combining two words can be also the reason why the form remains unrecognised. Another source of the plenty ign forms is the lack of tagging of words of foreign origin[[4]](#footnote-4). That information can be used in erasing errors of the textual layer of the Corpus.

Conclusion

The benefits that the Corpus provides lexicographers with are self-evident. Using corpus as a main resource of information in creating a dictionary is obligatory. But some of the crucial information can be taken beforehand – in the process of creating the corpus and by annotation tools. The Corpus, the dictionary and the NLP structure should be seen as parts of the supertool. The obtained data can be a crucial incentive to improve all of these resources.

References

Bronikowska, R., Gruszczyński, W., Ogrodniczuk, M., Woliński, M. (2016). The Use of Electronic Historical Dictionary Data in Corpus Design. *Studies in Polish Linguistics*, 11(2), pp. 47-56. Available at: <http://www.ejournals.eu/SPL/2016/Issue-2/art/7035/>).

Gruszczyński, W. (ed.) (2004–). Elektroniczny słownik języka polskiego XVII i XVIII wieku [The Electronic Dictionary of the 17th-18th c. Polish]. <http://sxvii.pl/>.

Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B. (eds.) (2012). Narodowy Korpus Języka Polskiego [The National Corpus of Polish]. Warsaw: Wydawnictwo Naukowe PWN.

Siekierska, K. (ed.) (1999-2004). Słownik języka polskiego XVII i 1. połowy XVIII wieku [The Dictionary of the Polish Language of the 17th and the first half of the 18th centuries]. Vol. 1. Kraków: Wydawnictwo Instytutu Języka Polskiego PAN.

Saloni, Z., Woliński, M., Wołosz, R., Gruszczyński, W., Skowrońska, D. (2012). Słownik gramatyczny języka polskiego [Grammatical Dictionary of Polish]. 2 ed. Warsaw: G. Murzynowski.

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

1. <http://nkjp.pl>. [↑](#footnote-ref-1)
2. <http://sgjp.pl>. [↑](#footnote-ref-2)
3. The converter is available at: <https://bitbucket.org/jsbien/pol> [↑](#footnote-ref-3)
4. The tokens analysed as foreign language are labelled <foreign>. [↑](#footnote-ref-4)