From dictionary to morphological analyzer. The case of Baroque Polish

Witold Kieraś (University of Warsaw; Polish Academy of Sciences)

The aim of the paper is to present a process of building a morphological analyzer for 17th and 18th century Polish. The analyzer constitutes a crucial part of an NLP toolkit aimed at creating a Baroque Corpus (*Korpus Barokowy*, abbrev. *KorBa*, Gruszczyński et al. 2013). The paper focuses on obtaining and automatically extending inflectional data from an existing dictionary of Baroque Polish for the purpose of automatic morphological analysis aimed at building a large annotated corpus of 17th and 18th c. Polish.

An analyzer consists of two parts: linguistic data and computer program. In our case, the latter is Morfeusz 2 (Woliński, 2014), which allows to customize both inflectional data and segmentation rules. The inflectional data is much more difficult to obtain, since it calls for laborious and time-consuming work of human lexicographers. For the purpose of *KorBa*, two sources of data are exploited: inflectional information from the *Electronic Dictionary of 17th–18th c. Polish* (e-SXVII, http://sxvii.pl, Gruszczyński 2004) and modified ("aged") data of *Grammatical Dictionary of Polish* (SGJP, http://sgjp.pl, Saloni et al. 2012, 2015; Woliński and Kieraś 2016) which provides inflectional data of contemporary language.

Inflectional paradigms in e-SXVII are prepared manually based only on forms found in historical texts, therefore are usually incomplete (currently ca. 62,000 forms of 24,000 lexemes), containing on average only 2.5 forms per paradigm. Thus, for the purpose of morphological analysis, the data calls for some automatic and semi-automatic extensions, i.e. reconstructions of forms that are not present in the dictionary.

Semi-automatic extensions are based on syncretisms and some other regularities of forms, e.g. plural nominative and vocative forms of nouns are always identical, so if one is noted in the dictionary, the other can be safely reconstructed. Numerous other such rules can be formulated, which results in reconstruction of ca. 10,000 forms of around 8,600 different lexemes. Yet, still the average ratio of forms per paradigms is small.

A more extensive method of reconstructing inflectional forms is based on formal similarities of paradigms. Numerous lexemes fall into very productive inflectional patterns, thus from incomplete paradigms of many lexemes inflected in the same way it is possible to induce a full or almost full set of endings for the given pattern and later apply these endings to all the lexemes of the type. Due to the fact that not all the lexemes are inflected in a regular and predictable way, the general idea is not faultless and may result in producing errors or forms that are only potential, but the overall error rate should not exceed ca. 10%. Nevertheless, some method of eliminating or correcting errors is needed.

Based on the concept described above, a formal procedure was formulated and applied to the data of e-SXVII which resulted in ca. 170,000 new forms. It is obvious that not all lexemes had their paradigms extended by the method, but ca. 16,000 of them were partially or fully reconstructed. In the paper, a detailed description of the method will be presented. Some examples of successfully reconstructed paradigms will be shown as well as those indicating limitations of the method.

References

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.

Gruszczyński, W., editor (2004). *Elektroniczny słownik języka polskiego XVII i XVIII wieku*. Kraków.

Gruszczyński, W., Adamiec, D., and Ogrodniczuk, M. (2013). Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.) — prezentacja projektu badawczego. *Polonica*, XXXIII:309–316.

Saloni, Z., Gruszczyński, W., Woliński, M., Wołosz, R., and Skowrońska, D. (2015). *Słownik gramatyczny języka polskiego*. publikacja internetowa, 3. edition.

Saloni, Z., Woliński, M., Wołosz, R., Gruszczyński, W., and Skowrońska, D. (2012). *Słownik gramatyczny języka polskiego*. Warszawa, 2. edition.

Siekierska, K., editor (1999-2004). *Słownik języka polskiego XVII i 1. połowy XVIII wieku*, volume 1. Kraków.

Woliński, M. (2014). Morfeusz reloaded. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111, Reykjavík, Iceland. ELRA.

Woliński, M. and Kieraś, W. (2016). The on-line version of Grammatical Dictionary of Polish. In Calzolari, N., Choukri, K., Declerck, T., Grobelnik, M., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., 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).