A corpus-based morphological analyzer of 19th century Polish

Magdalena Derwojedowa (University of Warsaw), Joanna Bilińska (University of Warsaw), Monika Kwiecień (University of Warsaw) and Robert Wołosz (University of Pécs)

Our aim is to present a morphological analyzer capable of processing Polish texts written in the past 200 years with respect to their original spelling and inflection. Despite of the analyzer, our project involved making a corpus of historical texts as a resource of linguistic data.

In the analyzer part, we simply adapted an analyzer made for contemporary texts. This task implied setting up new patterns and modifications to the existing ones. New paradigms are in majority historical alternations of modern patterns. The adjustments concerned primarily syncretisms and obsolete endings (cf. 1a-4). Some of the stems evolved in a way that affected endings, e.g. a pre-ending syllable got contracted (cf. 5b and 5a). Otherwise we added historical variants of the case endings (cf. 2). Newly added paradigms and forms (i.e. combination of a stem and an ending) are marked "19c", so in the output of the analyzer "obsolete"/"outdated" and "currently in use" forms can be easily spotted. Increasing or decreasing frequency of a form or a pattern over the period of time is valuable information of the trend and may be used as a predictor whether a variant may be assigned a class in general. However, in the current project we stick to the corpus evidence and do not reconstruct or extrapolate.

A difference in spelling that does not affect alternations of the stem was regarded a spelling variant (cf. 6). Such entries were added to the analyzer's lexicon, together with words that went out of use in the course of time and lexemes that changed in morphological characteristics (e.g. from feminine to masculine, cf. 7). New entries in the lexicon are time stamped with the year of the earliest corpus evidence. In addition, the spelling variants and words that shifted from one subclass to another are linked to their contemporary equivalents. A very similar idea of hyperlemma was proposed by Kučera (2007). In the lexical part of the project, our principle was not to miss a word, i.e. we ruled out only ill-formed words, puns, obvious dialect forms etc.

Any change in the analyzer was based on the data from the corpus compiled from texts published for the first time in years 1830-1918. The corpus is divided into five equal subcorpora to provide stylistic variety. A subcorpus consists of 200 samples, for every year there is at least one (but no more than 3) sample, in average 12 samples per year in the whole corpus. 1000 token text file is accompanied by metadata and source files. In basic statistical test held after the corpus was completed the subcorpora are clearly sorted out into fiction, drama, essays, scientific text for general public and news (cf. fig. 1). Because of relatively small samples, the diversity of the corpus in many respects (places, authors, printed sources etc.) is quite satisfactory. Several tests passed on the corpus proved that it can be used as a versatile resource to identify linguistic phenomena, trace their dynamics (cf. fig. 2) and turning points or to confront emerging rules of orthography and good usage from the grammar handbooks with everyday practice.

Examples

(1) a. dobrém (adj. inst., loc. sg. masc./neut, pres. dobrym 'good')
b. dobrym (adj. inst., loc. sg., dat. pl. masc./neut., pres. dobrym)

(2) ładnej, ładnéj (adj. gen., dat., loc. sg., pres. ładnej, 'pretty')

(3) pięcią (num. inst., pres. pięcioma 'five')

(4) każdę (adj. acc. sg. fem., pres. każdą 'every'), księżnę (subst., fem. acc. sg., pres. księżną 'princess, duchess')

(5) a. lekcyi (subst. fem. gen.,dat.,loc. sg., pres. lekcji 'lesson')
b. lekcyj (subst. fem. gen. pl., pres. lekcji/lekcyj 'lesson')

(6) triumphator (pres. triumfator 'triumpher'), patłażan (pres. bakłażan 'aubergine, eggplant'), papiér (pres. papier 'paper')

(7) dreszcz (fem. pres. masc. 'shiver'), planeta (masc. pres. fem. 'planet')

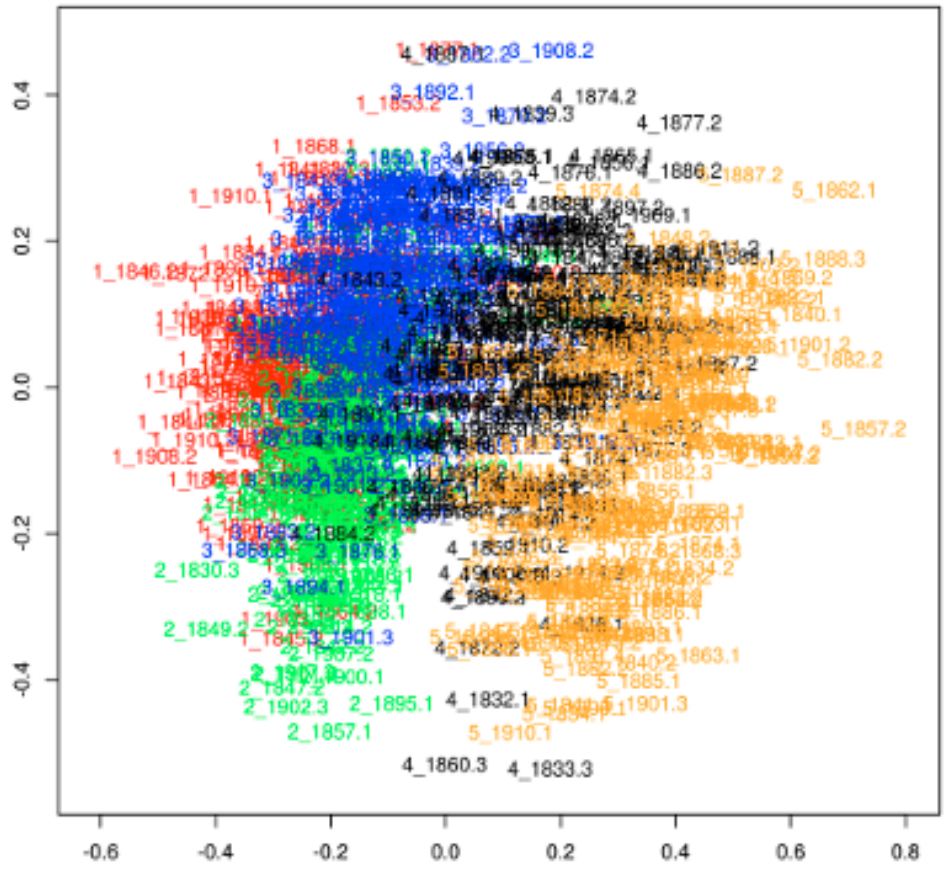

Figure 1. The five styles of the corpus grouped with stylo's MDS

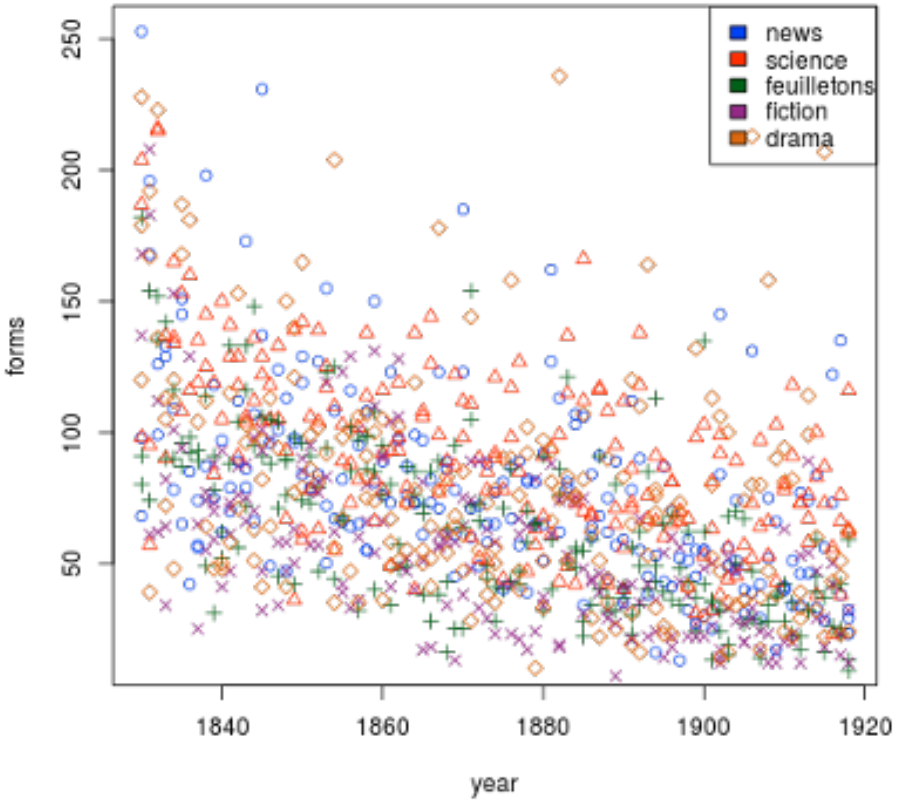

Figure 2. Number of 19th c. words and forms per year

References

Derwojedowa, M., Kieraś, W., Skowrońska, D., and Wołosz, R. (2014). Współczesne narzędzia leksykograficzne a analiza tekstów dawniejszych. Polonica, XXXIV:21--27.

Klemensiewicz, Z. (2002). Historia języka polskiego. Warszawa.

Kučera, K. (2007). Hyperlemma: A concept emerging from lemmatizing diachronic corpora. In Levická, J. and Garabík, R., editors, Computer Treatment of Slavic and East European Languages, pages 121--125. Bratislava.

Čavar, D., Jazbec, I.-P., and Stojanov, T. (2009). Cromo – morphological analysis for standard croatian and its synchronic and diachronic dialects and variants. In Proceedings of the 2009 Conference on Finite-State Methods and Natural Language Processing: Post-proceedings of the 7th International Workshop FSMNLP 2008, pages 183--190, Amsterdam, The Netherlands, The Netherlands. IOS Press.