Corpus similarity measure in grammatical analysis

Similarity measures are useful tools for comparing corpora or different parts of a single corpus, providing a large ('bird-eye') perspective. At the same time, keyword analysis methods allow for much closer introspection of such collections (Kilgariff 2001). This paper presents a newly developed similarity measure, based on cumulative word frequencies, which allows both for the measuring of distance between two frequency lists and for identifying their keywords, and discusses its usefulness in its application to grammatical studies.

Given two frequency lists, containing words ordered from the most to the least frequent, Cumulative Frequencies Measure (henceforth CFM) assigns a certain position to each word, based on the word's frequency as well as its rank. Then, the difference between the word's position on both lists is calculated. The average of such differences for all the words yields the list dissimilarity score. At the same time, an individual difference score for a given word indicates its keyness (the higher the score, the more characteristic the word is).

Though universal in nature (it can work with any pair of ordered lists), this method is well suited for performing various grammatical tasks. To illustrate this, the proposal describes three such tasks.

In the first one, CFM is employed to measure a recent diachronic change of part of speech patterns in the spirit of Mair and Leech (2006). In a diachronic corpus of Modern Polish (since we still lack of it we use the Polish part of the Polish-German parallel corpus, which is divided into six chronologically ordered subcorpora from 1750 to 1989) for each subperiod the frequency list of POS tags was obtained. Differences between the lists from adjacent time periods show the changes in the grammatical structure of texts in time; for example, while the differences between the eighteenth and the first half of the nineteenth century are rather small, there is a clear change in the late nineteenth century and an even more pronounced one after the World War II. The change in importance of individual POS tags can also be traced: for example, the usage of adverbial participles constantly declines with time, but their adjectival counterparts become increasingly more common.

The second experiment focuses on the study of nominal inflection in contemporary Polish (cf. Janda and Lyashevskaya 2011 for similar approach to Russian verbs). A frequency list for each grammatical case was obtained from the balanced subcorpus of the Polish National Corpus. The each-to-each comparison allows for detecting groups of cases which are similar in their lexical preferences. It seems, for instance, that accusative and locative are much closer to each other with this regard than dative and instrumental or nominative and vocative. Keyword analysis of each pair of cases may be used for the creation of 'grammatical profiles' of nouns, which show the cases of a given noun which are more likely to appear.

The last experiment uses the keyword approach to detect lemmata which are most typically connected with one of five verbal forms (namely: past, present, gerund, active adjectival participle and passive adjectival participle). Each list of lemmata is compared against the rest, yielding four keyword lists for each type. Thus, for each category, words with the highest keyness are found (for past tense the first two are *dojść* 'to come' and *wygrać* 'to win', for the present: *pozdrawiać* 'to greet' and *dziękować* 'to thank', and for the active participle: *dotyczyć* 'to concern' and *umożliwać* 'to enable', and so on).

References

Janda, Laura A., and Olga Lyashevskaya. "Grammatical profiles and the interaction of the lexicon with aspect, tense, and mood in Russian." *Cognitive linguistics* 22.4 (2011): 719-763.

Kilgariff, Adam. "Comparing corpora". *International Journal of Corpus Linguistics* 6 (2001):97–133.

Mair, Christian, and Geoffrey Leech. "Current Changes in English Syntax." *The handbook of English linguistics* 36 (2006): 318.