Abs-145

Václav Cvrček (Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague)

How large is the core of language?

Corpus research is based on the hypothesis that large and representative collections of texts reflect the language reality truly and precisely. In lexicography we assume that general corpus is a sufficient source for representation of core lexical units. However, the comparison with traditionally elaborated dictionaries shows us that there are still lexemes missing either in dictionaries or less frequently in corpora (peripheral lexemes, scientific terms etc.). Can we determine the range of lexical core exactly by corpus methods?

Let us begin with the assumption that the core vocabulary is the part of lexicon which is common to majority of texts and speakers. We can use the proportion of hapax legomena (i.e. words that occur only once) to all word-types in relation to the growing corpus size to identify the frequency range in which core elements occur. In hypothetically small corpus (a few sentences) the hapax-type ratio will be equal to one (each word-type is also a hapax). As we add texts to corpus (up to a few million words) the hapax-type ratio decreases (the number of new words including hapaxes is continuously increasing but the majority of added tokens are new instances of words already present in the corpus) from its maximal value (=1) to the local minimum (between 0.35 and 0.45). This is the turning point (in graph it is represented by a plateau) and from now on with extending the corpus the ratio increases because the amount of hapaxes grows at a faster pace than the number of types with frequency higher than one. The graph of the hapax-type ratio (which has similar shape in different languages regardless of the types of texts or their order) resembles pipe or chibouque (hence "pipe-graph").

This empirical finding tested on corpora of Czech, English and Italian brings us closer to exact determination of the range of core lexicon (this range differs, of course, in languages with typologically distinct structure). Subsequently, we can deduce the approximate size of a corpus sufficient for compiling a dictionary covering the core lexicon.

Shape of the hapax-type ratio function also suggests that there are still some unknown differences between text and language. Some of the quantitative laws discovered by exploring individual texts might therefore be biased by this phenomenon of qualitative change in data structure when the corpus exceeds certain size. We might thus interpret the pipe-graph (with slight exaggeration) as a special case of the parole-langue distinction: the first part of the graph (decreasing function) reflects the properties of texts (parole), then there is the transient part (plateau) and the third part (increasing function) reflects the whole language or domain (with corpus large enough to eliminate idiosyncrasies of individual texts or speakers).