Milena Hebal-Jezierska (Warsaw, Poland), Neil Bermel (Sheffield, United Kingdom)

Frequency and oppositions in corpus-based research into morphological variation

Abstract

Research into variation using corpora is changing our methods of describing variation itself. This change is connected first and foremost with the quantity of data available to us. For the majority of researchers, it dictates a description based on the frequency of forms as they appear in the corpus. Some such descriptions make use of individual examples alone and avoid frequency data, while others use frequency data as their main categorising tool, either with or without recourse to statistical measures.

In both frequency-based and non-frequency-based descriptions of variation, the concept of opposition has proved useful as a starting point, and has informed much of our work on morphological variation, which has been carried out in conjunction with projects hosted at the Czech Language Institute at the Czech Academy of Sciences (Exploring the Core and Limits of Czech Grammar through Corpora, 2003-5 and Chapters in a Grammar of Czech, 2006-8). Through comparing the relationship between two or more variants, we can reach a fuller picture of the linguistic situation. Comparison can be investigated in several ways, a few of which will be mentioned here. Of special interest are the mutual frequency relationship and the distribution of elements investigated.

In our contribution, we will attempt to classify variants from the point of view of their role in an opposition, and will exemplify this on Czech morphological variation. Oppositions can be constructed on the basis of the interplay of several different factors, for example: the relationship of frequency and acceptability of data; the relationship between frequency and usage in a certain text type; the distribution of forms in phrases and co-texts, etc.

Part of our talk will focus on categories formed by the frequency of the items in question. Some researchers have used frequency labels such as dominant, majority, equifrequential, minority, sporadic (see Šimandl, Hebal-Jezierska), or central and peripheral (see Tušková). We will attempt to answer the following questions: What these labels tell us and how they relate to the judgments of native speakers and to usage patterns (see Oliva/Doležalová, Bermel). We will also mention possible reasons for mismatches between corpus data and native-speaker judgments, and between corpus data and that from other large-scale text databases, in all cases using the concept of the opposition as our starting point. Here we believe morphology offers a particularly useful tool for analysis, as morphological variation (as opposed to syntactic variation) most often occurs in a closed set of two or perhaps three variants, and as such, data from one variant can provide useful information about the status of the other(s).

Our conclusions are specifically directed at the situation in Czech, but we propose that these suggestions may be more widely applicable.

References

Bermel N. (2008), 'Pilotní studie o vztahu mezi korpusovými daty a hodnoceními přijatelnosti konkurujících tvarů', A lecture given in Brno (in press).

Hebal-Jezierska, M. (2008), *Wariantywność końcówek fleksyjnych rzeczowników męskich żywotnych w języku czeskim*, Warszawa.

Oliva, K. / Doležalová, D. (2004): 'O korpusu jako o zdroji jazykových dat'. In: Karlík, P. (ed.): *Korpus jako zdroj dat o češtině.* Brno: Masarykova univerzita, s. 7-10.

Šimandl, J. (2008), *Dnešní stav skloňování substantiv takzvaných typů kámen a břímě* (PhD disertation), http://ktf.cuni.cz/~simandl/trf/dis/1.pdf (in press).

Tušková, J.M. (2006), *Variantní a dubletní tvary v současné deklinaci apelativních feminin,* Brno.