Extracting linguistic knowledge about collocations from corpora

Ulrich Heid

We start from the assumption that (lexical) collocations and verbal idioms are a type of multiword expressions which deserve a detailed linguistic and lexicographic description (Gouws/Heid 2006: treatment units of their own in dictionaries). If this is so, then there is a need for corpus-based tools which allow us to find out about the contextual (= syntagmatic) and paradigmatic properties of collocations. We intend to show how much of these data can be identified with acceptable quality in large enough corpora.

Syntagmatic properties have to do with the distributional behavior of collocations: preferences in number (*have high hopes*, pl.), determination (article use) or modifiability are well-known examples; some collocations and many verbal idioms have their own syntactic valency constructions (cf. *be in a position [+to+INF]*) or they co-occur preferentially with certain lexical items, e.g. as modifiers (cf. DE *Kritik üben* ("criticize") which prefers adjectives that typically collocate with *Kritik: harsche, scharfe Kritik üben* ("criticize severely"), cf. Häcki-Buhofer et al. 2014).

Examples of paradigmatic properties include the exchangeability of lexical elements of the collocation against synonyms, or the availability of nominalizations (*submit a proposal – submission of a proposal*) or of compounds in Germanic languages (DE *Antrag einreichen* ("submit a proposal") – *Einreichung eines Antrags* – *Antragseinreichung*). Another example are pragmatic marks and preferences with respect to domain-specific languages.

We will give examples of such data from German, English, French and Italian and we will assess to what extent such linguistic knowledge may be needed in translation and in (mother tongue or foreign language) text production. Thereafter, we intend to show which types of data of the above kind can be extracted with acceptable quality from corpus texts, and with which language processing techniques; we claim that state of the art dependency parsing provides a fair amount of such data thus facilitating the description work of terminologists and lexicographers.

References

Gouws, Rufus H. And Heid, Ulrich: "A model for a multifunctional dictionary of collocations", in Corino, Elisa et al. (Eds.): *Proceedings of the XIIth EURALEX International Congress*, (Alessandria: Edizioni dell'Orso) 2006, 979 – 988.

Häcki-Buhofer, Annelies et al.: *Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag*, (Tübingen: Francke) 2014.