Between collocation and construction: Lexical preferences in non-idiomatic word combinations

While there is a substantial body of work on the identification and lexicographic description of collocational word pairs (e.g. OCDSE for English) as well as idiomatic multiword expressions (e.g. Fellbaum 2007 for German), only few studies have addressed longer non-idiomatic word combinations (MWC) so far (e.g. Zinsmeister & Heid 2003).

One type of such MWC are collocational patterns involving three or more lexical items, which often form a series of semantically related MWC, e.g. DE {*scharfe* | *massive* | *heftige* | *harsche*} *Kritik üben* ("to criticize massively"). Such examples can be seen as MWC involving a variable slot with strong lexical and/or semantic restrictions. At the other end of the range are grammatical constructions with marked lexical (or semantic/morphosyntactic) preferences. An example is pattern (2) of the EN verb [*to*] *earn*, whose direct objects are almost always selected from a narrow semantic field different from pattern (1):

(1) sbdy *earns* sth (*money*, *dollars*, *wages*, ...)
(2) sth *earns* sbdy sth (*nickname*, *reputation*, *sobriquet*, *title*)

In this paper, we report ongoing research towards the description of such phenomena and the automatic identification of MWC candidates. We extract co-occurrence data for general syntactic patterns such as (3) below from large dependency-parsed corpora of German (Wikipedia, SDeWac, DECOW) and English (BNC, UKCOW). The data are stored in a tabular database comprising tuples of the form

(3) Adj – Subj – Verb – Adj – DObj – Adj – IObj

In this example, we collect instances of lexical verbs with the head lemmas of subject, direct object and indirect object, as well as pre-nominal adjectives. Any element that is not realized in a given instance is marked with the placeholder "–". Our quantitative approach builds on two premises:

1. Co-occurrence patterns between words cannot be reduced to a one-dimensional association score, but comprise multi-faceted aspects including frequency (measured by conditional probability or ΔP), salience (measured by a suitable confidence interval for MI) and the type-token distribution of each slot. For example, it is useful for EFL teaching that the indirect object in (2) is realized as a pronoun in 80% of all instances and that the form *him*/*himself* is significantly more frequent than for other ditransitive constructions.

2. The complex interrelations between different slots of a MWC can be modelled in terms of nested hypothesis tests, taking into account both significance and association strength. For example, a test for the independence of subject and object given construction (1) shows salient association for *I earn money* and *women earn less* in the British National Corpus. Such nested hypotheses may also involve semantic or morphosyntactic restrictions on the slots, or test whether a larger MWU is composed of overlapping smaller MWU (e.g. DE *scharfe Kritik üben* from *scharfe Kritik* + *Kritik üben*).

Our long-term goals are to design mathematically sound extraction procedures for MWC and to develop corpus-linguistic tools supporting descriptive work at the intersection of collocations, constructions and multiword expressions.

References

OCDSE: *Oxford Collocations Dictionary for Students of English*, 2nd edition, Oxford: OUP, 2009.

Fellbaum, Christiane (ed.): *Collocations and Idioms: Corpus-Based Linguistic and Lexicographic Studies*. Birmingham: Continuum Press, 2007. http://kollokationen.bbaw.de/htm/idb_de.html

Zinsmeister, Heike & Heid, Ulrich: 'Significant Triples: Adjective+Noun+Verb Combinations'. In: Ferenc Kiefer, Gábor Kiss, Júlia Pajzs (Eds.): *Papers in Computational Lexicography – COMPLEX 2003*, Budapest, 2003.