Rainer Perkuhn

The comparability of co-occurence patterns

Semantic relations like paronymy are usually discussed by referring to the linguistic unit of the "word" (or, to be more precise, a "word sense"). The meaning of words can change over time, depending on the context of use. In general, the influencing factors can be uncovered from their use in a particular thematic field, register or genre, or discourse.

In order to carry out an empirical investigation of semantic relations, there are several very powerful corpus-linguistic instruments. The *Institut für Deutsche Sprache* (Institute for the German Language) runs the *Deutsche Referenzkorpus* DeReKo (German Reference Corpus) as an empirical basis (Institut für Deutsche Sprache 2017) while Cosmas II (Institut für Deutsche Sprache 2018) is an online system that facilitates the analysis of co-occurrences, alongside standard corpus queries. Co-occurrence analyses reveal the range of typical patterns of use, presenting them clearly as a co-occurrence profile. When applying the methods embedded in Cosmas II, users can stipulate the data (subset) they want to use as well as the search query and other options.

Co-occurrence profiles provide more or less conclusive indications as to the meaning potential of linguistic units which can be interpreted in different ways. In the embedded, interactive version, they can inspire the researcher to refine a hermeneutic analysis in relation to the above-mentioned determining factors in order to control for the effects of specific parameters. In other words, specific periods of time, thematic fields or the distribution of sources can be viewed separately in the summary of results. For a follow-up analysis, peculiarities can then taken special account of when refining the considered dataset.

To the extent that co-occurrence profiles indicate aspects of meaning, it is possible to compare the profiles to shed light on the semantic relation between the chosen units. A large number of similarities or parallels imply that there is a close relationship while differences suggest specific fields of application and, therefore, specific senses. The relationship between and extent of similarities or differences can certainly be influenced by the composition of the data.

When a linguistic reference unit is compared to each of two other linguistic units, there are cases where a high degree of similarity is proposed for both comparisons. This can also happen even though the joint commonalities are virtually negligible, i. e. when the manifestations of similarities presumably differ in nature. The one type of similarity can be the result of a specific topic while the other type can be due to another topic or even to a specific characteristic of one of the other dimensions named above. In case this characteristic – e. g. a thematic or discursive field – is not idiosyncratic for the comparison of a pair of words but, instead, has a value which overlaps with other linguistic units, this can be exploited methodologically in order to get an idea of this characteristic. When this characteristic is reflected in the typical patterns of use of several linguistic units, the profiles of all of these units will be very similar to the reference unit's profile. Other characteristics could lead to other groups which also have a remarkable similarity. When the elements of the groups are compared with each other internally, the groups are revealed to have a certain inner coherence: on account of their belonging to the same topic, for example, all

elements which are similar to the reference word are also similar to each other as they can be assigned to the same topic. The fuzziness of the boundaries between groups depends on the extent to which the characteristics of various dimensions can touch or overlap.

These methodological approaches have been implemented in the co-occurrence database CCDB (Belica 2007), which includes co-occurrence profiles of more than 220,000 entries. Profiles can be called up for the reference units and partner words can be searched for. Using the above-indicated degree of similarity between profiles, lists of entries can be compiled which have a similar profile to the reference word. This lists of similar profiles can then be further processed with the help of self-organising maps, which attempt to identify and visualise the groupings hinted at in the previous paragraph. The resulting maps can ideally be used to identify areas which can be associated with – and explained by – the above-mentioned characteristics.

In almost all cases, the methods employed by the CCDB present impressive and helpful insights. For certain constellations, however, it can happen that phenomena can be over- or underestimated due to the assumptions and constraints of the set-up. The limitations include the fact that the list of keywords was compiled in 2007 and has not been updated since as well as the corpus it is based on (a virtual corpus taken from the DeReKo 2007). In addition, co-occurrence profiles are only available for a small number of settings for the search and co-occurrence queries with their respective options and parameters. The consequences are that it is not possible to capture language phenomena arising since 2007. Fine-tuned explorative follow-ups or cross-sectional studies must fall back on Cosmas II, although the results from the two corpora are very difficult to compare and cannot be used in the CCDB set-up either.

Alongside these well-known and obvious effects of the different approaches, one aspect should be amplified which has received little attention so far. At many points in this text, the somewhat awkward formulation “linguistic unit” was deliberately used. Studies on meaning or semantic relations often start out from the intuitive concept of a “word” and may then relativise it as a “word sense”; they normally mean, on a morphological level, a collection of inflected forms, which can also be called a “lemma” or a “lexeme”, depending on the domain, and which is named after a representative of the inflection paradigm which is as unmarked as possible. The co-occurrence database was also provided with an operationalised counterpart as its reference unit to make it technically feasible and manageable from a user perspective. A more differentiated consideration of all forms would be far too complex, requiring the user to manually collate the individual results of the inflected forms of a paradigm. From a theoretical linguistic perspective, however, there has been little discussion as to whether all forms of a lemma (or a word sense) contribute homogenously to its meaning to the same extent or whether they would have to be considered separately by form or groups (e. g. participles with verbs or comparatives with adjectives). Should this be the case, the quantitative distribution of the forms might be critical and plays a major role in processes which evaluate the forms within one paradigm and aggregate the assessments to a single lemma. But even independently of the quantitative benchmark data, the characteristic as such can already provide indications of special meanings for specific forms. When it turns out, for example, that the word *kindisch* (*childish*) is used relatively often in its uninflected form, this already reveals that the essence of a sentence is covered by its predicative use, in contrast to *kindlich* (*childlike*), which attributively modifies only one of several constituents in a sentence. While the word *effektiv* (*effective*)

in the proximity of *Prozess* (*process*) is either uninflected or used as an inflected pre-modifier as in "effektiver Prozess" (*effective process*), it is necessary to ascertain whether the similar pattern of *effizient* (*efficient*) plus *Prozess* as in "Prozesse … effizienter …" is more likely to hint at a comparative form (*processes ... more efficient ...*) rather than an inflected pre-modifier. In both examples, it can certainly be argued that these special meanings should be captured and described. According to the quantitative constellations in the data and depending on the settings of the search and analysis parameters, it could be the case that these connections are over- or underestimated. Searching for a lemma with a simultaneous lemmatisation of the partner words presents the biggest threat of distortion. The search for word forms – no matter whether with or without the partner words being lemmatised – increases the effort required of possibly conflating the individual results back to a lemma. The most practical way is the one which was also chosen for the CCDB: searching for a lemma while considering the partner words as word forms. Syntagmatic patterns are a source of additional information in order to solve the above-mentioned coincidences.

In addition to this, it is always possible to carry out increasingly fine-tuned analyses in Cosmas II with any settings, whereby it would be advisable to pay attention to the distribution of forms within paradigms. Alongside the methodological deliberations and reflections described above, the full report also presents shorter studies which illustrate how outliers in the paradigm (taking account of upper/lower case in potential adjectival paronyms, the distribution of comparative/inflected forms) can be detected and visualised.

References

Belica, Cyril (2007): Kookkurrenzdatenbank CCDB – V3. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. http://corpora.ids-mannheim.de/ccdb (accessed on: 18.12.2018).

Institut für Deutsche Sprache (2017): Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache 2017-I (Release date 8.3.2017). Mannheim: Institut für Deutsche Sprache. www.ids-mannheim.de/DeReKo (accessed on: 18.12.2018).

Institut für Deutsche Sprache (2018): Cosmas II-Recherchesystem. https://cosmas2.ids-mannheim.de/cosmas2-web (accessed on: 18.12.2018).