Ulrich Schnörch

Co-occurrence and comparative analyses: methodological considerations relating to lexicographic description of paronyms

This paper is divided into two major subsections. In the first, *Technologie* (*technology*) is used as an example to outline how co-occurrence analysis (cf. Belica 1995) was used for a lexicographic analysis of individual headwords as part of the general monolingual online dictionary *elexiko*. In the second part, *Technik* (*technique*/*technology*) and *Technologie* (or *technisch*/*technologisch*, *technical*/*technological*) are used to illustrate the advantages and disadvantages of co-occurrence analysis when comparing at least two paronymous headwords.

Within the framework of corpus-based lexicography, co-occurrence analysis is a proven adaptive and adaptable method which can be used in many different ways. Its fundamental purpose is to pre-structure large amounts of linguistic data. Using the monolingual general online dictionary *elexiko*, this paper exemplifies which dictionary data can be identified or derived based on a corpus. The example of *Technologie* was chosen for this purpose. Research using COSMAS II revealed 178,748 hits in the *elexiko* corpus (as of 14. 11. 2017). The co-occurrence analysis then produced a list which includes, amongst others, the co-occurrence partners and any (relevant) syntagmatic patterns. The first stage in the analysis, which was primarily editorial (and interpretative), leads to a whole series of data which are copied directly into the word's entry. This **first group of data** includes, amongst others,

(groups of) Collocates, e. g.:

- *alternativ*, *ausgereift*, *effizient*, *energiesparend*; *anwenden*, *beherrschen*, *einsetzen*, *entwickeln* (*alternative*, *well-established*, *efficient*, *energy-saving*; *to deploy*, *to control*, *to use*, *to develop*). These and numerous other adjectives and verbs match the meaning of the word in the sense of 'production technology'.
- *modern*, *veraltet*; *bauen*, *besitzen*, *exportieren*, *kaufen* (*modern*, *outdated*; *to build*, *to own*, *to export*, *to buy*). These and numerous other adjectives and adverbs rather match a meaning in the sense of 'technical equipment'.

Syntagmatic patterns, e. g.:

- *auf saubere Technologien setzen* (*to focus on clean technologies*), *mit westlicher Technologie* (*with western technology*), which again suggest the meaning of 'production technology'.
- *die neueste Technologie auf den Markt bringen* (*to launch the latest technology on the market*), *die Technologie an China verkaufen* (*to sell the technology to China*), *mit der Technologie vertraut sein* (*to be familiar with the technology*), which fit in with the meaning of 'technical equipment'.

Semantic relations for the headword, i. e. in terms of similarity, e. g.: *Produktionsmethode*, *Verfahren*, *Technik* (*production method*, *process*, *technology*), which are synonymous with *Technologie* in the sense of 'production technology'.

The second type of information which can be extracted from the co-occurrence list only really differs from the first in that the editorial-analytic processes are brought more strongly to the fore. The most important focus is definitely on a semantic analysis. For a lexicographic description, it is essential to analyse the 'meanings' of words, regardless whether they are called **readings** ('Lesarten', as in *elexiko*) or **aspects of use** ('Verwendungsaspekte', as in the dictionary of paronyms). As such, the grouping of collocates is a first step in the right direction: semantically 'coherent' groups of collocates can be used to derive individual meanings of the word in that one group corresponds to one meaning. Using this procedure in *elexiko* resulted in five different interpretations for *Technologie*. The interplay between selection, grouping and categorisation has been simplified somewhat here, but with the help of further, largely editorial steps, the process of semantic differentiation makes (made) it possible to subcategorise the collocates within the five interpretations in relation to sentence-semantic factors and to assign the syntagmatic patterns and sense-related expressions obtained from the co-occurrence lists to the different interpretations.

The second, longer part of the paper shows how the co-occurrence analysis was applied and, where necessary, modified for a comparative (lexicographic) investigation of paronyms, including a discussion on the advantages and disadvantages.

The paronym dictionary "Paronyme – Dynamisch im Kontrast" (Paronyms – Dynamic in Contrast) also generates (generated) its data/information from a corpus, from compiling a list of headwords to the information in the entries on each word. The different types of information presented in the first part of this paper also appear in the dictionary of paronyms, albeit partially in strongly modified form: in quantitative terms, the choice is much more selective and in qualitative terms, the (types of) information going by the same name is (are) based on different concepts.

The structural linchpin of the dictionary of paronyms lies in the concept of comparison: a lexicographic description of paronyms involves comparing and contrasting two or more headwords. For a lexicographic description of semantically related headwords, i. e. of at least two paronyms, a comparison is decisive, in terms of both content and form. The aim of the dictionary is to ascertain and present this comparability in a comprehensible manner. To this end, the data in the entry must be selected, edited and then presented. This will now be illustrated with the help of the paronyms *Technik*/*Technologie* and *technisch*/*technologisch:*

Searching the corpus of paronyms and co-occurrences analyses for the headwords at hand produced the following results:

Technik:	199,422 hits in the corpus, 3,574 co-occurrences
Technologie:	48,369 hits in the corpus, 2,839 co-occurrences
technisch:	282,122 hits in the corpus, 3,327 co-occurrences
technologisch:	12,518 hits in the corpus, 1,112 co-occurrences

These figures should always be borne in mind as they can influence the interpretation of the data in a later stage of the analysis.

Selection and grouping of the **collocates**: there is clearly a limit to the number on display. Up to ten exemplary collocates are selected for the presentation, starting with the most significant. It is possible to create several groups whose members belong together in the

broadest sense due to their sentence-semantic function, such as properties and actions, etc., for example:

Technik 'production process, method'

- *neu*, *modern*, *digital*, *ausgereift*, *innovativ*, *zukunftsweisend* (*new*, *modern*, *digital*, *well-established*, *innovative*, *forward-looking*)
- *erlernen*, *beherrschen*, *entwickeln*, *funktionieren* (*to learn*, *master*, *develop*, *work*)

Technologie 'production process, method'

- *umweltfreundlich*, *innovativ*, *zukunftsweisend*, *nachhaltig*, *veraltet*, *ausgereift*, *aufwendig*, *biometrisch* (*environmentally friendly*, *innovative*, *forward-looking*, *sustainable*, *obsolete*, *well-established*, *complex*, *biometric*)
- *einsetzen*, *ersetzen*, *fördern* (*to use*, *replace*, *encourage*)

Selection and grouping of the **syntagmatic patterns**: again the number is limited, this time to a maximum of five patterns, although it is also possible to include variations. The most significant pattern is given first. The contextual patterns for *Technologie* 'production process, method' are listed here as an illustration:

- *in* [*umweltfreundliche*/*zukunftsträchtige*] *Technologien investieren* (*to invest in* [*environmentally friendly*/*future-oriented*] *technologies*)
- *Technologien für den intelligenten Verkehr* (*technologies for intelligent transport*)
- *Technologien der Zukunft* (*technologies for the future*)
- *die Technologie ist noch nicht ausgereift* (*the technology is not yet well established*)
- *Technologien und* [*Produkte*/*Dienstleistungen*/*Märkte*] (*technologies and* [*products*/*services*/*markets*])

Selection (and grouping) of **semantic relations**: this is limited to synonyms, and sometimes antonyms. Only the top five most significant examples are documented for each aspect of use, along with appropriate evidence from the corpus.

Selection of **evidence from the corpus**: a maximum of three examples serves to clarify a particular aspect of use. Each example serves as an illustration of the relation between the paronymous headword and a synonym or antonym (see above).

The restriction on data included in the entry is the prerequisite for being able to compare headwords lexicographically and to ensure that the comparison is easy to grasp, when possible at one glance. The goal is to highlight the data-driven commonalities and differences, in other words to be able to present them in the new dictionary of paronyms on a sound footing.

In conclusion, the paronymous status of a pair of words is not a 'given' but has to be investigated in detail based on their use(s). In other words, the lexicographic work which is necessary to produce a comparative description of potential pairs of paronyms has to start 'from scratch' every time, i. e. by consulting the corpus. The co-occurrence analysis is a tried-and-tested method for a corpus-driven analysis of significant phenomena in language use. One advantage is that it was based on a corpus which was specially compiled to comply with the requirements of the dictionary of paronyms. The greatest shortcoming for co-occurrence analyses is that the automated process can only ever be used for a single headword. The actual comparison has to be carried out editorially by lexicographers, usu-

ally with much greater effort. This means, though, that there is a good chance that the entries in the dictionary are well founded, balanced, descriptive and tailor made for comparisons. This ensures a high degree of reliability which cannot be guaranteed for all reference works.

References

Belica, Cyril (1995): Statistische Kollokationsanalyse und -clustering. Korpuslinguistische Analysemethode. Mannheim: Institut für Deutsche Sprache. www.corpora.ids-mannheim.de (accessed on: 20. 9. 2018).

COSMAS II: Corpus Search, Management and Analysis System. www.ids-mannheim.de/cosmas2 (accessed on: 20. 9. 2018).

elexiko: www.elexiko.de or www.owid.de (accessed on: 20. 9. 2018).

Paronyme – Dynamisch im Kontrast: www.owid.de/parowb (accessed on: 20. 9. 2018).