Mareike Teichmann

Self-organising maps and contrasting near synonyms as corpus-linguistic tools to analyse paronyms using the example of *technisch*/*technologisch*

This paper describes the use of self-organising maps (SOMs) and contrasting near synonyms (CNS) to investigate the semantic properties of paronyms using the example of *technisch*/*technologisch* (*technical*/*technological*). These tools allow semantically similar pairs of words to be compared in the contexts in which they are used, a procedure which was first tried out on a large scale in the "dictionary of paronyms" project (cf. Storjohann 2013).

A quick look at question-and-answer platforms (like "wer-weiss-was.de") reveals that many users are insecure about the adjective pair *technisch*/*technologisch* in connection with how and in which contexts they should be used. Storjohann/Schnörch (2016, p. 232 f.) call this phenomenon paronymy, defining it as "the linguistic confusion of words which are similar in form and/or meaning or their replacement/shifting into other domains of use more or less deliberately (yet sometimes erroneously)".

More specialised and even well-established dictionaries tend not to include much information on concrete contexts of use. The dictionary of paronyms uses typographical methods to present them more clearly. Self-organising maps are one such methods, visualising the analysis of semantic proximity; contrasting near synonyms is another, which is used to differentiate between synonyms (see http://corpora.ids-mannheim.de/ccdb). A special co-occurrence database (CCDB) has more than 220,000 co-occurrence profiles (cf. Keibel/Belica 2007) which provide a basis for various explorative methods.

The CCDB can be used to call up expressions which have a similar co-occurrence profile to the search term. The results are then grouped and displayed two-dimensionally, based on the degree of semantic similarity, with the help of self-organising maps. Familiarity with the SOM and CNS maps for the potential paronyms *technisch*/*technologisch* is advisable (see figures 1–3 in the full-length paper) in order to understand the following explanations more easily. The colour-coded map gives a kind of overview of the degree of similarity of various expressions which have overlaps with the search term in their co-occurrence profiles. These overlaps reveal thematic or discourse-bound domains (cf. Belica 2011; cf. Keibel/Belica 2007). These expressions are arranged in square boxes in such a way that profiles which are more similar are closer together and are often grouped in the same square. These squares are then arranged in such a way that they are close to or further away from each other depending on their semantic proximity (cf. Perkuhn/Keibel/Kupietz 2012, p. 133 ff.). The individual boxes should not come across as rigidly defined areas, which might be the impression given by a visualisation involving 25 squares, each with a different colour. Instead, the different shades of the same colour used for clusters located close to each other clearly illustrate the relationship continuum.

It is now a question of interpreting the squares and the way in which they are arranged. First of all, a superordinate (more abstract) concept (supersign, cf. Vachková/Belica 2009) is identified for the contents of a square, which also covers the squares in the immediate vicinity. These superordinate concepts, which can be conceived of as domains, topic areas or discourses, represent different aspects of the use of the word. It has proved to be the case that this procedure allows topic areas to be defined more precisely; in contrast, a purely collocational analysis only identifies direct, syntagmatic word partnerships.

In order to analyse *technisch*/*technologisch*, the self-organising maps for the two terms are interpreted systematically in the manner described above. After the basic concepts have been ascertained for *technisch* and *technologisch*, the thematic conceptual domains are presented in tabular form (see table 1), with the similarities and differences between the areas of use clearly shown. There are overlaps in the areas of GESELLSCHAFTLICHE UND ZIVILISATORISCHE BELANGE (ISSUES RELATING TO SOCIETY AND CIVILISATION) for *technisch* and KULTURELLE, ZIVILISATORISCHE UND ETHISCHE BELANGE (ISSUES RELATING TO CULTURE, CIVILISATION AND ETHICS) for *technologisch*. Lexical expressions like *sozial* (*social*), *geschichtliche* (*historical*) and *gesellschaftlich* (*societal*) can be found on both SOMs. There are also areas, however, where there is hardly any semantic overlap.

In the next step, the CNS map of *technisch*/*technologisch* is evaluated. The CNS analysis proceeds along the same lines as the SOM method, except that here the complete list of similar profiles for both expressions is used as a basis. The big advantage of doing so is that a combined SOM can be produced. The results are presented in colour-coded boxes, facilitating a simple and rapid interpretation of the proximity, or otherwise, of the relationship between the expressions at hand. The colour-coding makes use of the yellow-red spectrum: bright yellow or bright red areas indicate that the words are used in (very) different ways while orange signals that they overlap. The more orange a box is, the more likely it is that the underlying field of meaning is equally important for both words (cf. Perkuhn/Keibel/Kupietz 2012, p. 137). This reveals whether the two lexemes under investigation are found in the same contexts and whether they are synonymous.

A quick look at the CNS map for *technisch*/*technologisch* reveals that none of the areas is bright red or bright yellow and that the large central area is orange, indicating where semantic overlaps are to be found. The orange-coloured boxes are, however, arranged in groups by different shades of orange. In the lower right yellow area of the map, words can be found (e. g. *fußballerisch* (*relating to football*), *spieltechnisch* (*playing technique*)) which can be allocated to the abstract concept of CHARAKTERISIERUNG VON BEGABUNGEN/FERTIGKEITEN/FÄHIGKEITEN (CHARACTERISING TALENTS/SKILLS/ABILITIES) for *technisch*. The central part of the CNS map is lighter orange, suggesting that the similarity of the profiles is more oriented towards *technisch*. Here it is clear that both expressions appear in similar contexts: both words relate, for example, to the characterisation of KULTURELLE, ZIVILISATORISCHE UND ETHISCHE BELANGE (e. g. *sozial* (*social*), *zivilisatorisch* (*relating to civilisation*), *demografisch* (*demographic*)) as well as to the fields of WIRTSCHAFTLICHE PLANUNG (ECONOMIC PLANNING) (e. g. *Betriebsablauf* (*operating procedure*), *kundenorientiert* (*customer-oriented*)) and ARBEITSMETHODEN (WORKING METHODS) (e. g. *fachspezifisch* (*subject-specific*), *kaufmännisch* (*commercial*)). This analysis of the CNS map largely confirms the results of the SOMs as interpreted above and makes it possible to differentiate more finely between the concepts of use in specific fields. Overlaps exist in the following domains: ARBEITSBEREICHE FORSCHUNG (RESEARCH-RELATED FIELDS OF WORK); WISSENSCHAFTLICHE BEREICHE (AREAS RELATING TO SCIENCE); GESELLSCHAFTLICHE UND ZIVILISATORISCHE BELANGE; MANAGEMENT UND VERWALTUNG/ORGANISATION, VERÄNDERUNG (MANAGEMENT AND ADMINISTRATION/ORGANISATION, CHANGE); WIRTSCHAFTSLEIS-

TUNGEN/ARBEITSMETHODE, WIRTSCHAFTLICHE PLANUNG, ENTWICKLUNG (ECONOMIC OUTPUT/WORKING METHODS, ECONOMIC PLANNING, DEVELOPMENT).

The final stage involves analysing co-occurrence patterns and including corpus evidence as a further source of information. The co-occurrence analysis establishes which words found in the immediate vicinity of the search term are statistically significant, based on the German Reference Corpus (DeReKo), from which the evidence quoted in the paper is also taken. The results are presented in table 2 as exemplary lexical realisations of reference expressions.

By analysing the results of the SOM and CNS tools, the contextual uses or superordinate thematic categories can be ascertained for a particular word and compared to the contexts and categories associated with a similar expression. Using maps to visualise this information makes the findings easier to interpret, especially in relation to semantic proximity/distance.

References

Belica, Cyril (2011): Semantische Nähe als Ähnlichkeit von Kookkurrenzprofilen. In: Abel, Andrea/Zanin, Renata (eds.): Korpora in Lehre und Forschung. Bozen-Bolzano: Bozen-Bolzano University Press, pp. 155–178.

Keibel, Holger/Belica, Cyril (2007): CCDB: A corpus-linguistic research and development workbench. Proceedings of the 4th Corpus Linguistics conference, Birmingham. http://corpora.ids-mannheim.de/cl2007-134.pdf (accessed on: 12.12.2016).

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): Korpuslinguistik. Paderborn: Fink.

Storjohann, Petra (2013): Korpuslinguistische und lexikografische Ansätze zur Beschreibung deutscher Paronyme. In: Sava, Doris/Scheuringer, Hermann (eds.): Im Dienste des Wortes. Lexikologische und lexikografische Streifzüge. Festschrift für Ioan Lăzărescu. Passau: Stutz, pp. 401–418.

Storjohann, Petra/Schnörch, Ulrich (2016): Wie kann ein Paronymwörterbuch funktionieren? In: Colliander, Peter et al. (eds.): Tagungsband der Internationalen Deutschlehrertagung IDT. Bd. 5: Linguistische Grundlagen für den Sprachunterricht. Bozen: Bozen University Press, pp. 231–242.

Vachková, Marie/Belica, Cyril (2009): Self-organizing lexical feature maps. Semiotic interpretation and possible application in lexicography. In: Rauch, Irmengard/Seymour, Richard K. (eds.): IJGLSA 13, 2. Berkeley: IJGLSA/University of California Press Pp. 223–260. Available online at: http://corpora.ids-mannheim.de/IJGLSA.pdf (accessed on: 2.12.2016).