The lexicographic use of corpora and computational tools for disambiguation

Dr Petra Storjohann
Institut für Deutsche Sprache Mannheim (IDS)

1. Traditional vs. computerised lexicography

Contemporary German lexicography is still mostly intuition-based and quite far from integrating corpus analysis into the process of dictionary making. Only a few projects are increasingly basing the compilation of their lexicographic structures on electronic corpora. The lexicographic practice remains a traditional "manual" art where compilers investigate numerous texts with the focus on spotting new words or detecting new senses of existing words.[1] They gather material into card file systems which are still the most important instruments for German lexicographers (cf. Scholze-Stubenrecht 2001: 49, Hanks 1990: 31). Nowadays, the citation card files have been replaced by digitised searchable databases, but although such conventional systems have developed into elaborate and comprehensive information resources, they cannot compare to the volume and versatility of data in corpora. As Sinclair (1991: 4) comments:

> Especially in lexicography, there is a marked contrast between the data collected by computer and that collected by human readers exercising their judgment on what should or should not be selected for inclusion in a dictionary.

The limits of such citation card file systems can be perceived when scrutinising any monolingual dictionary that works on such a basis. First, in some cases suitable examples cannot be found due to the paucity of evidence. Secondly, not every sense of a word can be encompassed in a lexicographer's brain. And finally, as corpus-oriented lexicographers would point out, deficits and inconsistencies arise from the lack of operational lexicographic algorithms.

The revolutionary step to move away from intuition and employ a more objective approach through corpus-guided studies is a fairly recent development in German lexicography. The latest editions of monolingual dictionaries such as *Duden* (2000), however, have begun to combine traditional dictionary making with technological innovations, and it can be seen that corpora are being employed more effectively. Computer-based examinations of linguistic data are performed, particularly with regard to the selection of lexemes that are to be included in a new edition. On the one hand, it has been realised that the probability of finding suitable reference material is increased considerably through a large digital data collection and that specific data is accessible by defined search options. On the other hand, the chief asset of corpora is still seen in being a large pool of documentary evidence to support and verify the information gained by introspection (cf. Haß-Zumkehr 2002: 45). As far as the lexicographic process of describing lexemes semantically is concerned, corpora are not being used sufficiently.

Whereas in other fields corpora have brought substantial change in working procedures and have led to the development of new approaches, German lexicography has not yet reached the stage where corpus-derived data delivers new approaches. Finding explanations that fit the evidence and adopting new approaches which enable us to extract new empirical linguistic knowledge still remain the principal objectives of empirical lexicography (cf. Sinclair 1991: 36). The potential of corpora and computer-based text analysing tools for lexicography needs to be fully recognised and to be more effectively employed in order to profitably supplement traditional lexicography.

With the accessibility of comprehensive data and computational tools, lexical disambiguation[2] should in particular be addressed from a new perspective. Currently, there is a demand for an disambiguation technique based on empirical and theoretical grounds. Although the issue of sense distinction has received much attention, most research is predominantly preoccupied in the field of machine translation, information retrieval, and hypertext navigation. Despite a long lexicographic tradition, a sound disambiguation practice for monolingual dictionaries has not been put forward. Today, such a procedure presupposes the existence of an elaborate linguistic disambiguation theory that is compatible with comprehensive empirical research. Its supporting instruments should be corpora, text analysing tools, and it also requires a good understanding of data interpreting. In the following part it is shown how such a need for an operational disambiguation algorithm with a rigorous empirical and theoretical basis can be met.

2. Project *Elexiko*[3]

At the *Institut für Deutsche Sprache Mannheim* (IDS) the project *elexiko* is developing a new corpus-driven lexicographic hypertext dictionary of the German lexicon and its present-day usage. Its aim is to construct a linguistic data warehouse where a diverse spectrum of readers can inexhaustibly explore the German language. By using a hypertext structure we are able to create a comprehensive dictionary with an extensive linking system for illustrating different types of language structures. As Hanks (1990: 35) argues:

> … there is a tendency for human lexicographers to focus on the way words are used to describe the world rather then on the way words interrelate with one another."

Its flexible search system will enable the reader to look for selective information according to individual needs. *Elexiko* shows a radical change from existing lexicographic conventions. The user will face different information with respect to quality, quantity and form of presentation. Our project follows a corpus-based approach, meaning we study linguistic instances on the basis of evidence found in corpus data.

Elexiko has developed a disambiguation technique which is based on empirical and theoretical grounds. The lexicographic prerequisites of this disambiguation procedure are an elaborate theory, corpora, a data-processing software, and the linguistic competency of data interpreting.

2.1 Elaborate disambiguation theory

The principal criticism monolingual German dictionaries face today with respect to the issue of lexical disambiguation is the lack of a linguistic theory and formal grounding. Only a handful of dictionaries

[1] Also known as reading-and-marking method.

[2] In this paper disambiguation is not understood in terms of sense tagging/semantic tagging but refers to the lexicographic procedure of identifying and distinguishing the senses of a word for further semantic/syntactic description in a dictionary entry.

[3] Formerly named *Wissen über Wörter*, see http://www.ids-mannheim.de/wiw/

list meaning discriminating criteria; traditionally paradigmatic and syntagmatic patterns. However, a sound theoretical basis has not been developed so far.

In order to avoid a one-sided theory, *elexiko* has striven to offer a multi-dimensional model which forms the basis for disambiguating content and function words. Crucially, we have constructed a systematic network of criteria consisting of semantic, syntactic, propositional, contextual, conceptual-referential and functional components combined in a complex cross-classification. The main components, however, are the semantic-syntactic and referential functions that correlate with the specific use of a search word within a collocational or sentential context. As Reichmann (1989) emphasises, the interaction of several complementary criteria illustrates differences in linguistic patterns of a word in different contexts and generates adequate results with regard to the correct identification of different functions. As pointed out by Sinclair (1987), the way word patterns relate to the use, function and meaning of individual words requires particular examination. Computer-processed data is hence classified according semantic, syntactic and referential functions with respect to the selected key word and its contexts.

The advantage of describing a lexeme according to functional classes instead of traditional word classes is seen in the illustration of the connection between the semantic form and proposition or illocutionary potential of a communicative unit (cf. Strauß 1989). The classification is the linguistic method underlying the lexicographic exploration of sentential contexts and semantic relations to identify word senses, and it consists of the following levels:

I Level of semantic classes of words

This level is understood as the semantic classification of words into autosemantic or synsemantic classes. Words that contribute semantically to the proposition of a collocational or sentential unit are autosemantic/content words. Words that do not have a characteristic semantic contribution to a propositional and/or illocutionary meaning of a sentence, but are functional or syntactic constituents attached to a noun or verbal phrase are synsemantic/function words.

II Level of sentential-semantic classes/syntactic classes and subclasses

Content words in particular can be classified into propositional types according to their semantic characteristics in sentential contexts. Thus, autosemantic words are grouped into sentential-semantic classes which reflect propositional types and mainly consist of predicators[4] (cf. von Polenz 1988). The class of predicators comprises for example event-denoting predicators, relation-denoting predicators, quality-denoting predicators, state-denoting predicators, classifying predicators etc. Furthermore, sentential-semantic classes also encompass the smaller groups of deictics, quantifiers, and partitives. Function words are disambiguated by their different syntactic functions in a sentence. This group

[4] Linguistically, the term predicators defines a verb in its functional relation to the clause, meaning an expression which takes a subject to form a sentence. We will, however, refer to the term as it is defined in terms of Propositional Logic and Predicate Logic. Here, a predicator designates a property or a relation and they can be ascribed to different objects. Grammatically, different word classes such as nouns, adjective and verbs can function as predicators (cf. Seiffert 1969: 23)

includes conjunctions, prepositions, articles, particles etc. Like autosemantic words, they can be sorted into specific subgroups with corresponding functional properties.

The corresponding semantic or syntactic function for each word sense is mainly identified by the investigation of its paradigmatic and syntagmatic patterns, thematic roles, and modification patterns, as represented in contexts. The analysis of these patterns requires a corpus-driven investigation of surface relations, meaning the analysis of the co-occurrences of the word. Generally, it can be said that a polysemous word has several functions and can hence be grouped into different sentential-semantic classes, which function as lexical disambiguators.

III Level of denotation

The third level of sense differentiation is the level denotation. After each type of propositional function is classified, the reference or denotation of a word is determined. This level illustrates the senses of words by describing their different conceptual values. A system of conceptual classes helps to classify lexemes according to a taxonomic knowledge base which is being developed simultaneously.[5] Words that belong to different entities tend to appear in recognisably different contexts. Therefore, in some cases a knowledge base can function as a context discriminator. The analysis of the denotational and conceptual content also follows a corpus-guided investigation of collocates where distinct word patterns which are associated with functions and the use of a word are examined.

IV Level of specifications

The fourth level comprises semantic specifications of words and provides further distinction of meaning. Specifications are understood as either inherent semantic properties, or semantic features which are identified by complements/adjuncts, e.g. aspectual features (aktionsarten) for process-denoting predicators. Others, like quality-denoting predicators, can be subcategorised according to their specifications into emphasising, classifying and modifying predicators. As far as function words are concerned, they often carry functional specifications. Conjunctions, for example, function as connectors of clauses. The type of connection between the clauses describes an individual specification of conjunctions such as conditional or concessive.

V Level of relational properties

In addition, some lexemes show relational characteristics which enable the lexicographer to further subdivide word senses. Relational properties refer to characteristics such as transitivity, symmetry, and reflexivity. Whereas the classification of sentential-semantic classes, denotation and specifications contain disjoint sets of features, a word can have more than one relational property. The classification of relational properties is a cross-classification which does not necessitate a preceding classification of specification.

e.g.: *This is a table* (*table* = predicator), *This is red* (*red* = predicator), *It moves* (*moves* = predicator)

[5] Disambiguation according to a knowledge base has become a widespread concept in Natural Language Processing (NLP)

Altogether these criteria form a multi-faceted cross-classification functioning as a linguistic sense discriminator. Polysemous words and their different semantic or syntactic functions are classified according to their propositional features. If a word can be categorised into different classes we are able to distinguish word senses. Words that belong to different classes at one level occur in different contexts. Theoretical sense differentiation does not require a classification at every level. Tests have demonstrated that the senses of some words are sufficiently distinguished by classifying their propositional functions at the first or second level only.

While the linguistic disambiguation model provides a theoretical basis of sense distinction, only the actual corpus data can provide an empirical validation of that model. In the next part attention will turn to the question of how theory meets corpus. It is also revealed how the necessary semantic analysis is conducted in a corpus-driven way and which tools are utilised in *elexiko*.

2.2 COSMAS-Korpus-Recherchesystem and concordancing software

The *Institut für Deutsche Sprache Mannheim* has compiled the largest German corpora. Currently, they are composed of about 1,900 million words from contemporary written and spoken texts. These corpora are accessible via a corpus query system called "COSMAS-Korpus-Recherchesystem" (henceforth *Cosmas*[6]). The programme can be adjusted with respect to the settings of specific preference parameters.[7]

Cosmas is an efficient statistical corpus query system with a concordancing software package *Statistische Kollokationsanalyse und Clustering*[8] which has been utilised extensively in our project for lexical disambiguation. Its collocation analysis has been employed to detect statistically significant patterns of co-occurrences of word forms which are evaluated with regard to the use and semantic embedding of a word. Performing a collocation analysis results in the detection of linguistic regularities as well as irregularities within large text samples. Its main advantage is its ability to organise collocational structures by exploring semantic and syntactic neighbourhoods and calculating significance, thereby providing pre-structures which must be analysed systematically by the lexicographer. A further benefit can be seen in its ability to analyse language without introspective expectancy. It also offers empirical access to language in a comprehensive and systematic way, no lexicographer could perform. The lexicographer's task in disambiguating a polysemous word with *Cosmas* follows a procedure at three levels, the collocation-level, the KWIC-level, and the text-level.

[6] *Cosmas* is an abbreviation of Corpus Search, Management and Analysis System. It was developed at the IDS Mannheim and is publicly accessible via the internet (http://corpora.ids-mannheim.de/cosmas/).

[7] In the menu box there are several tabs which can be used to control the calculation of collocations, e.g. define the span of words around the hit or modify the performance of collocation calculation (by defining granularity of clustering, method for resolving cluster ambiguities, de/activating of function words, de/activating lemmatiser etc.). The setting of parameters impinge on the result of collocates and the hierarchy in the collocation list.

[8] The software *Statistische Kollokationsanalyse und Clustering* was developed by Cyril Belica (1995-2002) at the IDS Mannheim and can be used publicly via the internet since 1995.

2.2.1. Collocation-level

As indicated in the theoretical model, we believe that a sense distinction algorithm has to include the analysis of words by exploiting their contexts[9]. The disambiguation of polysemous words starts with the analysis of collocations[10]. The result is a retrieved list of co-occurrences (also called collocates) organised hierarchically and arranged according to the degree of lexical cohesion (lexical density).[11] Although generally the result is only a statistical one, the advantage of a collocation list lies in its organisation and structuring of contexts, and the alignment of sense (cf. Sinclair 1991: 61). The collocation analysis identifies salient words which cluster together in a collective context. These contexts exemplify the semantic and syntactic dependency relationships that the key word participates in. This degree of clustering is expressed by frequency-based statistics where the most significant clustering have the highest score.

Collocational structures are essential for lexicographic disambiguation, as they reveal linguistic patterns of the use and the propositional functions of a word by showing diverse paradigmatic, syntagmatic and syntactic structures, idioms, thematic domains, discourse analysis, and by uncovering co-referential co-occurrences. They reflect the complexity and the network of linguistic structures around the node by demonstrating, for example, the thematic roles of verbs and types of modifications of nouns, or simply any word that is closely associated with the key word. The itemised list of collocates offers the lexicographer different perspectives into the use of a word and its significant, as well as insignificant, semantic and syntactic neighbourhood. Restricted as well as non-restricted collocations give a picture of variant and invariant structures which can be classified into different linguistic categories.

2.2.2 KWIC-level

Although the organisation and structure of collocations offer great insights into the use of a word and its different semantic patterns, the co-occurrences and their significance must be evaluated as indications of senses only. Collocates cannot reveal the complex linguistic characteristics a key word exhibits in a larger empirical study of contexts. A quick insight into the actual contexts of the key word and its corresponding collocates can be gained at the KWIC-level (Key Word in Context[12]). KWICs are generated by concordancing software and are a way of displaying the search term and a selected collocate in a text so that the selected node word is listed in the middle column, with a certain amount of context on either side, usually a single line context.

Lexicographers can draw five main benefits from the KWIC-level. First, the search term and all its corresponding collocates can be investigated individually and systematically by analysing their collective context. The advantage of this level is a selective and systematic analysis of co-occurrences

[9] We will exploit the term "context" when we refer to the notion of surrounding text of a search item, where strictly speaking "co-text" might be more appropriate.

[10] Collocations and the identification of salient words are not restricted to binominal structures. Also see Belica/Steyer (2002) (http://www.ids-mannheim.de/kt/kollok.html)

[11] Collocates are arranged by decreasing lexical density as determined by the corresponding log-likelihood ratio value.

[12] KWIC is a universal format for concordances.

and their relationship to the key word. Secondly, this perspective on a small part of a common context reveals other words within the same semantic neighbourhood. Although they are usually statistically insignificant, they might contribute to the identification of the senses of the search item. Additional semantic partners such as paradigmatic co-occurrences, words functioning as modifiers, or words with semantically recurring patterns, and collocational or lexical restrictions are essential sense differentiating elements. Thirdly, the lexicographer is able to identify irrelevant collocates by establishing whether any collocate occurs, for example, due to faults of the lemmatising software.[13] Fourthly, the KWIC is a small communicative unit which illustrates the proposition and its semantic components. Hence, it is vital to explore the predication in order to classify the involved predicators according to the aforementioned model. Finally, the systematics of this analysing procedure partly reveals unexpected results. As the concordancer works without any introspective expectancy and strict consistent statistical methods, the probability of capturing every sense of a word is higher than that of an intuitive search. Generally, it can be summarised that the KWIC-level is the first stage of a two-step verification process where the indications of the collocation-level are examined.

2.2.3 Text-level

The second step of the verification process is realised at the text-level. A KWIC is a one-line contextual display and in most cases does not display a full sentence. There are cases where the KWIC format is not adequate for the study of some words, as a detailed perspective into the meaning of the full proposition cannot be given. Therefore, the information provided cannot answer all questions in terms of the actual use of a search word. Here, the final level of text display must be consulted for closer study. The extent of the display can be selected according to the number of preceding and following sentences. The broad context must finally be selected to reveal the semantic behaviour of a word in its full environment. The larger the context is chosen, the larger semantic potential it can offer and, thus, the more differentiated an analysis can be performed.[14] Only at this level is the lexicographer able to recognise linguistic patterns around the node word and its collocates as well as indications of referential components and the recurrence of semantic conglomerates which are important aspects of identifying semantic characteristics.

Generally, we can conclude that the first level should be regarded as a means of pre-selection and pre-structuring for the disambiguation procedure. The actual systematic exploration of contexts – smaller and larger ones – is lexicographically simplified and systematised by the alignment of the collocational structures of the word. But only after the examination of the second and third level is the lexicographer able to linguistically categorise senses according to the aforementioned model and finally theoretically verify the disambiguation of the key word.

[13] As Sinclair (1997: 31) claims: "…we do not know which details are essential, which important, which optional, which indicative, which transitory, which random and which distracting."

[14] The number of sentences to be selected is restricted due to copyright contracts.

3 Conclusion

Elexiko has put forward an elaborate disambiguation theory, a linguistic classification model which illustrates the propositional differences words can carry semantically, and hence their various senses can be categorised within a theoretical framework. In this way, lexicographic sense distinction has been elevated from a procedure conducted by introspection to a fully model-based task. *Elexiko* is also the first monolingual German dictionary which works in a corpus-based way. Linguistic information is derived from the complex empirical study of corpora and exemplified through citations. With respect to lexical disambiguation our project illustrates perspicuously the break with conventional German lexicography. Effectively, what has been achieved here is the development of a viable solution to the problems of establishing a disambiguation algorithm which links theoretical and empirical factors.

Initial tests have clearly ascertained that the demands of modern lexicography can be met through a combination of the following: corpora as an empirical base, a corpus-processing software as a method to generate semantic information and as an instrument for structuring senses, and a robust theoretical model which linguistically justifies meaning discrimination. Although we can already derive valuable results from these tests, considerable testing of the proposed disambiguation technique is still required and is currently being conducted. So far, the results have demonstrated more detailed semantic descriptions and more objectively disambiguated senses than have ever been offered by other monolingual dictionaries. The hypertext structure of our dictionary allows a semantic complexity which can only be elicited from the exploration of corpora and the use of computational tools. A number of examples have shown different results in terms of the identified senses of a word. In most cases, compared to *Duden* (2000), the methodological basis, upon which the senses of a key word are identified in our project has led to the identification of additional senses.

However advantageous any performance of computer analysis might be, *Cosmas* and its software components cannot replace the human element. Corpora and *Cosmas* are instruments for lexicographers. A corpus-processing tool can perform statistical analyses which are indicative, but they cannot interpret data linguistically. Although they give strong and necessary measurable evidence, so far we cannot envisage a world of dictionaries without lexicographers, a world where computers analyse a massive amount of linguistic data and where dictionaries would be generated automatically. The decision as to how to interpret corpus data and how to select irrelevant from relevant information remains the task of the lexicographer. At the same time modern lexicography must recognise that computer technology can supply us with the necessary volume of data (and the software for analysing it) and substantially enhance the compilation of a dictionary. It significantly improves, simplifies and systemises the lexicographer's work by performing large empirical explorations of data otherwise unmanageable for the unaided researcher.

Bibliography

Belica C, Steyer K 2002 *Die COSMAS-Kollokationsanalyse – statistisches Modell, Funktionsweise und Interpretationsspielräume*. (IDS-Homepage: http://www.ids-mannheim.de/kt/kollok.html).

Belica C 1995-2002 *Statistische Kollokationsanalyse und Clustering*. Korpusanalysemodul. (IDS-Homepage: http://corpora.ids-mannheim.de/cosmas), Institut für Deutsche Sprache, Mannheim.

Hanks P 1990 Evidence and Intuition in Lexicography. In Tomaszczyk J, Lewandowska-Tomaszczyk B (eds), *Meaning and Lexicography*. Amsterdam/Philadelphia, Benjamins. pp 31-41.

Haß-Zumkehr U 2002 Das Wort in der Korpuslinguistik - Chancen und Probleme empirischer Lexikologie. In Ágel V (ed), *Das Wort: seine strukturelle und kulturelle Dimension.* Festschrift für Oskar Reichmann zum 65. Geburtstag. Tübingen, Niemeyer, pp 45-70.

Polenz v. P 1985 *Deutsche Satzsemantik*. Grundbegriffe des Zwischen-den-Zeilen-Lesens, (=Sammlung Göschen 2226), Berlin/New York, de Gruyter.

Reichmann O 1989 Einführung. In Anderson R R, Goebel U, Reichmann O (eds), *Frühneuhochdeutsches Wörterbuch. Bd.1, a – äpfelkern*. Berlin/ New York, de Gruyter, pp 1-164.

Scholze-Stubenrecht W 2001 Das Internet und die korpusgestützte praktische Lexikographie. In Korhonen J (ed), *Von der mono- zur bilingualen Lexikographie für das Deutsche*. (Finnische Beiträge zur Germanistik Vol 6). Frankfurt, Peter Lang, pp 43-64.

Seiffert H 1969 *Einführung in die Wissenschaftstheorie*. Vol 1: Sprachanalyse - Deduktion - Induktion in Natur und Sozialwissenschaften, München, C. H. Beck.

Sinclair J 1987 Introduction. In Sinclair J et al (eds), *Collins Cobuild English Language Dictionary*. London/Glasgow, Collins, pp XV-XXIII.

Sinclair J 1991 *Corpus, Concordance, Collocation*. Oxford, OUP.

Sinclair J 1997 Corpus Evidence in Language Description. In Wichmann, A et al (eds), *Teaching and Language Corpora*. London/New York, Longman, pp 27-39.

Strauß G 1989 Angabe traditioneller Wortarten oder Beschreibung nach funktionalen Wortklassen im allgemeinsprachigen Wörterbuch? In Hausmann F J, Reichmann O, Wiegand H E, Zgusta L (eds), *Wörterbücher: ein internationales Handbuch zur Lexikographie*. (Handbücher zur Sprach- und Kommunikationswissenschaft Vol 5.1). Berlin/New York, de Gruyter, pp 788-796.

Dictionary

Duden 2000 *Das große Wörterbuch der deutschen Sprache*. 10 Volumes on CD-Rom. Mannheim.