GermaNet – a Lexical–Semantic Net for German

Birgit Hamp and Helmut Feldweg
Seminar für Sprachwissenschaft
University of Tübingen
Germany
email: {hamp,feldweg}@sfs.nphil.uni-tuebingen.de

Abstract

We present the lexical–semantic net for German "GermaNet" which integrates conceptual ontological information with lexical semantics, within and across word classes. It is compatible with the Princeton WordNet but integrates principle–based modifications on the constructional and organizational level as well as on the level of lexical and conceptual relations. GermaNet includes a new treatment of regular polysemy, artificial concepts and of particle verbs. It furthermore encodes cross–classification and basic syntactic information, constituting an interesting tool in exploring the interaction of syntax and semantics. The development of such a large scale resource is particularly important as German up to now lacks basic online tools for the semantic exploration of very large corpora.

1 Introduction

GermaNet is a broad–coverage lexical–semantic net for German which currently contains some 16.000 words and aims at modeling at least the base vocabulary of German. It can be thought of as an online ontology in which meanings associated with words (so–called *synsets*) are grouped according to their semantic relatedness. The basic framework of GermaNet is similar to the Princeton WordNet (Miller et al., 1993), guaranteeing maximal compatibility. Nevertheless some principle–based modifications have been applied. GermaNet is built from scratch, which means that it is neither a translation of the English WordNet nor is it based on a single dictionary or thesaurus. The development of a German wordnet has the advantage that the applications developed for English using WordNet as a resource can be used for German with only minor modifications. This affects for example information extraction, automatic sense disambiguation and intelligent document retrieval. Furthermore, GermaNet can serve as a training source for statistical methods in natural language processing (NLP) and it makes future integration of German in multilingual resources such as EuroWordNet (Bloksma et al., 1996) possible.

This paper gives an overview of the resource situation, followed by sections on the coverage of the net and the basic relations used for linkage of lexical and conceptual items. The main part of the paper is concerned with the construction principles of GermaNet and particular features of each of the word classes.

2 Resources and Modeling Methods

In English a variety of large–scale online linguistic resources are available. The application of these resources is essential for various NLP tasks in reducing time effort and error rate, as well as guaranteeing a broader and more domain–independent coverage. The resources are typically put to use for the creation of consistent and large lexical databases for parsing and machine translation as well as for the treatment of lexical, syntactic and semantic ambiguity. Furthermore, linguistic resources are becoming increasingly important as training and evaluation material for statistical methods.

In German, however, not many large–scale monolingual resources are publically available which can aid the building of a semantic net. The particular resource situation for German makes it necessary to rely to a large extent on manual labour for the creation process of a wordnet, based on monolingual general and specialist dictionaries and literature, as well as comparisons with the English WordNet. However, we take a strongly corpus–based approach by determining the base vocabulary modeled in GermaNet by lemmatized frequency lists from text corpora[1]. This list is further tuned by using other available sources such as the CELEX German database. Clustering methods, which in principle can apply to large corpora without requiring any further information in order to give similar words as output, proved to be interesting but not helpful for the construction of the core net. Selectional restrictions of verbs for nouns will, however, be automatically extracted by clustering methods. We use the Princeton WordNet technology for the database format, database compilation, as well as the Princeton WordNet interface, applying extensions only where necessary. This results in maximal compatibility.

3 Implementation

3.1 Coverage

GermaNet shares the basic database division into the four word classes noun, adjective, verb, and adverb with WordNet, although adverbs are not implemented in the current working phase.

For each of the word classes the semantic space is divided into some 15 semantic fields. The purpose of this division is mainly of an organizational nature: it allows to split the work into packages. Naturally, the semantic fields are closely related to major nodes in the semantic network. However, they do not have to agree completely with the net's top–level ontology, since a lexicographer can always include relations across these fields and the division into fields is normally not shown to the user by the interface software.

GermaNet only implements lemmas. We assume that inflected forms are mapped to base forms by an external morphological analyzer (which might be integrated into an interface to GermaNet). In general, proper names and abbreviations are not integrated, even though the lexicographer may do so for important and frequent cases. Frequency counts from text corpora serve as a guideline for the inclusion of lemmas. In the current version of the database multi–word expressions are only covered occasionaly for proper names (*Olympische Spiele*) and terminological expressions (*weißes Blutkörperchen*). Derivates and a large number of high frequent German compounds are coded manually, making frequent use of cross–classification. An implementation of a more suitable rule–based classification of derivates and the unlimited number of semantically transparent compounds fails due to the lack of algorithms for their sound semantic classification. The amount of polysemy is kept to a minimum in GermaNet, an additional sense of a word is only introduced if it conflicts with the coordinates of other senses of the word in the network. When in doubt, GermaNet refers to the degree of polysemy given in standard monolingual print dictionaries. Additionally, GermaNet makes use of systematic cross–classification.

3.2 Relations

Two basic types of relations can be distinguished: **lexical relations** which hold between different lexical realizations of concepts, and **conceptual relations** which hold between different concepts in all their particular realizations.

Synonymy and **antonymy** are bidirectional lexical relations holding for all word classes. All other relations (except for the 'pertains to' relation) are conceptual relations. An example for synonymy are *torkeln* and *taumeln*, which both express the concept of the same particular lurching motion. An example for antonymy are the adjectives *kalt* (cold) and *warm* (warm). These two relations are implemented and interpreted in GermaNet as in WordNet.

The relation **pertains to** relates denominal adjectives with their nominal base (*finanziell* 'financial' with *Finanzen* 'finances'), deverbal nominalizations with their verbal base (*Entdeckung* 'discovery' with *entdecken* 'discover') and deadjectival nominalizations with their respective adjectival base (*Müdigkeit* 'tiredness' with *müde* 'tired'). This pointer is semantic and not morphological in nature because different morphological realizations can be used to denote derivations from different meanings of the same lemma (e.g. *konventionell* is related to *Konvention (Regeln des Umgangs)* (social rule), while *konventional* is related to *Konvention (juristischer Text)* (agreement).

The relation of **hyponymy** ('is–a') holds for all word classes and is implemented in GermaNet as in WordNet, so for example *Rotkehlchen* (robin) is a hyponym of *Vogel* (bird).

Meronymy ('has–a'), the part–whole relation, holds only for nouns and is subdivided into three relations in WordNet (component–relation, member–relation, stuff–relation). GermaNet, however, currently assumes only one basic meronymy relation. An example for meronymy is *Arm* (arm) standing in the meronymy relation to *Körper* (body).

[1]We have access to a large tagged and lemmatized online corpus of 60.000.000 words, comprising the ECI–corpus (1994) (*Frankfurter Rundschau, Donau–Kurier, VDI Nachrichten*) and the *Tübinger NewsKorpus*, consisting of texts collected in Tübingen from electronic newsgroups.

For verbs, WordNet makes the assumption that the relation of **entailment** holds in two different situations. (i) In cases of 'temporal inclusion' of two events as in *schnarchen* (snoring) entailing *schlafen* (sleeping). (ii) In cases without temporal inclusion as in what Fellbaum (1993, 19) calls 'backward presupposition', holding between *gelingen* (succeed) and *versuchen* (try). However, these two cases are quite distinct from each other, justifying their separation into two different relations in GermaNet. The relation of entailment is kept for the case of backward presupposition. Following a suggestion made in EuroWordNet (Alonge, 1996, 43), we distinguish temporal inclusion by its characteristics that the first event is always a subevent of the second, and thus the relation is called **subevent** relation.

The **cause** relation in WordNet is restricted to hold between verbs. We extend its coverage to account for resultative verbs by connecting the verb to its adjectival resultative state. For example *öffnen* (to open) causes *offen* (open).

Selectional restrictions, giving information about typical nominal arguments for verbs and adjectives, are additionally implemented. They do not exist in WordNet even though their existence is claimed to be important to fully characterize a verbs lexical behavior (Fellbaum, 1993, 28). These selectional properties will be generated automatically by clustering methods once a sense–tagged corpus with GermaNet classes is available.

Another additional pointer is created to account for **regular polysemy** in an elegant and efficient way, marking potential regular polysemy at a very high level and thus avoiding duplication of entries and time–consuming work (c.f. section 5.1).

As opposed to WordNet, connectivity between word classes is a strong point of GermaNet. This is achieved in different ways: The cross–class relations ('pertains to') of WordNet are used more frequently. Certain WordNet relations are modified to cross word classes (verbs are allowed to 'cause' adjectives) and new cross–class relations are introduced (e.g. 'selectional restrictions'). Cross–class relations are particularly important as the expression of one concept is often not restricted to a single word class.

Additionally, the final version will contain examples for each concept which are to be automatically extracted from the corpus.

4 Guiding Principles

Some of the guiding principles of the GermaNet ontology creation are different from WordNet and therefore now explained.

4.1 Artificial Concepts

WordNet does contain artificial concepts, that is non–lexicalized concepts. However, they are neither marked nor put to systematic use nor even exactly defined. In contrast, GermaNet enforces the systematic usage of artificial concepts and especially marks them by a "*?*". Thus they can be cut out on the interface level if the user wishes so. We encode two different sorts of artificial concepts: (i) lexical gaps which are of a conceptual nature, meaning that they can be expected to be expressed in other languages (see figure 2) and (ii) proper artificial concepts (see figure 3).[2] Advantages of artificial concepts are the avoidance of unmotivated co–hyponyms and a systematic structuring of the data. See the following examples:

In figure 1 *noble man* is a co–hyponym to the other three hyponyms of *human*, even though the first three are related to a certain education and *noble man* refers to a state a person is in from birth on. This intuition is modeled in figure 2 with the additional artificial concept *?educated human*.

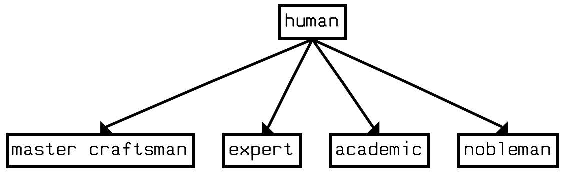

Figure 1: Without Artifical Concepts

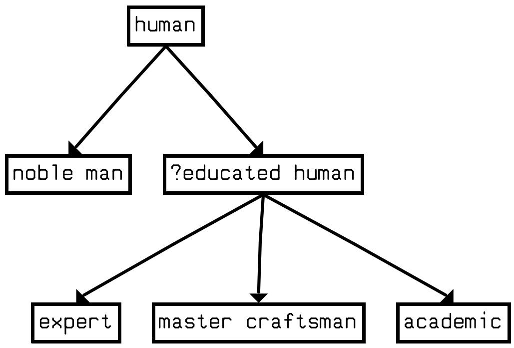

Figure 2: Lexical Gaps

In figure 3, all concepts except for the leaves are proper artificial concepts. That is, one would not expect any language to explicitly verbalize the concept of for example manner of motion verbs which specify the specific instrument used. Nevertheless such a structuring is important because

[2]Note that these are not notationally distinguished up to now; this still needs to be added.

it captures semantic intuitions every speaker of German has and it groups verbs according to their semantic relatedness.

4.2 Cross–Classification

Contrary to WordNet, GermaNet enforces the use of cross–classification whenever two conflicting hierarchies apply. This becomes important for example in the classification of animals, where folk and specialized biological hierarchy compete on a large scale. By cross–classifying between these two hierarchies the taxonomy becomes more accessible and integrates different semantic components which are essential to the meaning of the concepts. For example, in figure 4 the concept of a cat is shown to biologically be a vertebrate, and a pet in the folk hierarchy, whereas a whale is only a vertebrate and not a pet.

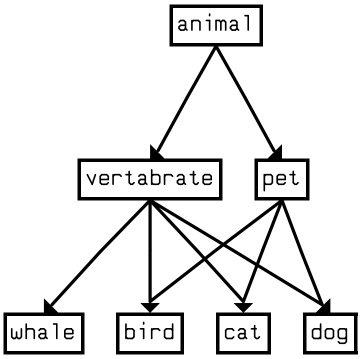

Figure 4: Cross–Classification

The concept of cross–classification is of great importance in the verbal domain as well, where most concepts have several meaning components according to which they could be classified. However, relevant information would be lost if only one particular aspect was chosen with respect to hyponymy. Verbs of sound for example form a distinct semantic class (Levin et al., in press), the members of which differ with respect to additional verb classes with which they cross–classify, in English as in German. According to Levin (in press, 7), some can be used as verbs of motion accompanied by sound (*A train rumbled across the loop–line bridge.*), others as verbs of introducing direct speech (*Annabel squeaked, "Why can't you stay with us?"*) or verbs expressing the causation of the emission of a sound (*He crackled the newspaper, folding it carelessly*). Systematic cross–classification allows to capture this fine–grained distinction easily and in a principle–based way.

5 Individual Word Classes

5.1 Nouns

With respect to nouns the treatment of **regular polysemy** in GermaNet deserves special attention.

A number of proposals have been made for the representation of regular polysemy in the lexicon. It is generally agreed that a pure sense enumeration approach is not sufficient. Instead, the different senses of a regularly polysemous word need to be treated in a more principle–based manner (see for example Pustejovsky (1996)).

GermaNet is facing the problem that lexical entries are integrated in an ontology with strict inheritance rules. This implies that any notion of regular polysemy must obey the rules of inheritance. It furthermore prohibits joint polysemous entries with dependencies from applying for only one aspect of a polysemous entry.

A familiar type of regular polysemy is the "organization – building it occupies" polysemy. GermaNet lists synonyms along with each concept. Therefore it is not possible to merge such a type of polysemy into one concept and use cross–classification to point to both, *institution* and *building* as in figure 5. This is only possible if all synonyms of both senses and all their dependent nodes in the hierarchy share the same regular polysemy, which is hardly ever the case.

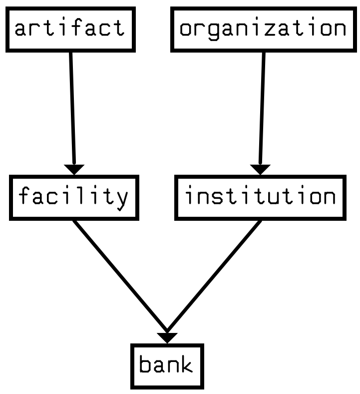

Figure 5: Regular Polysemy as Cross–Classification

To allow for regular polysemy, GermaNet introduces a special bidirectional relator which is placed to the top concepts for which the regular polysemy holds (c.f. figure 6).

In figure 6 the entry $bank_1$ *(a financial institution that accepts deposits and channels the money into lending activities)* may have the synonyms *depository financial institution, banking concern,*

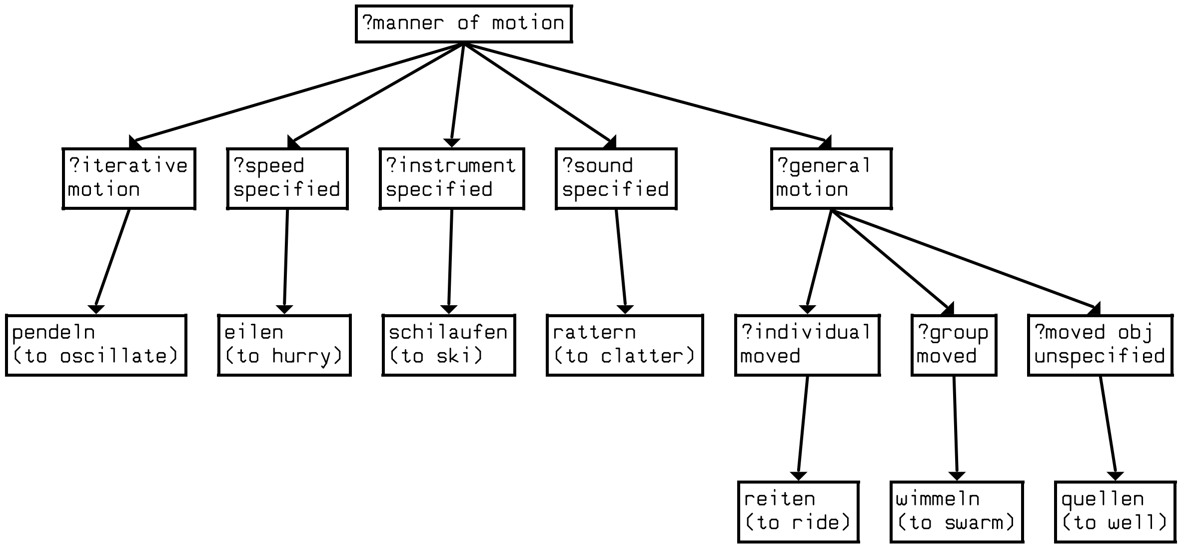

Figure 3: Proper Artificial Concepts

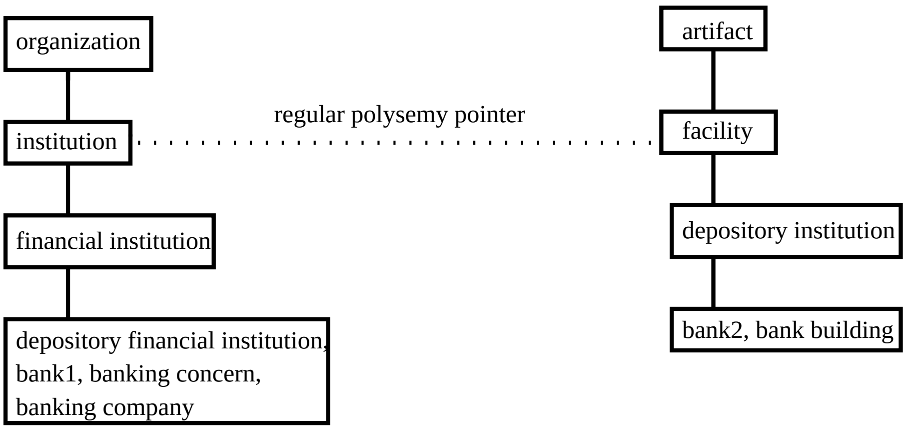

Figure 6: Regular Polysemy Pointer

banking company, which are not synonyms of *bank*$_2$ *(a building in which commercial banking is transacted)*. In addition, *bank*$_1$ may have hyponyms such as *credit union, agent bank, commercial bank, full service bank*, which do not share the regular polysemy of *bank*$_1$ and *bank*$_2$.

Statistically frequent cases of regular polysemy are manually and explicitly encoded in the net. This is necessary because they often really are two separate concepts (as in *pork*, *pig*) and each sense may have different synonyms (*pork meat* is only synonym to *pork*). However, the polysemy pointer additionally allows the recognition of statistically infrequent uses of a word sense created by regular polysemy. So for example the sentence *I had crocodile for lunch* is very infrequent in that crocodile is not commonly perceived as meat but only as animal. Nevertheless we know that a regular polysemy exists between meat and animal. Therefore we can reconstruct via the regular polysemy pointer that the meat sense is referred to in this particular sentence even though it is not explicitly encoded. Thus the pointer can be conceived of as an implementation of a simple default via which the net can account for language productivity and regularity in an effective manner.

5.2 Adjectives

Adjectives in GermaNet are modeled in a taxonomical manner making heavy use of the hyponymy relation, which is very different from the satellite approach taken in WordNet. Our approach avoids the rather fuzzy concept of indirect antonyms introduced by WordNet. Additionally we do not introduce artificial antonyms as WordNet does (*pregnant, unpregnant*). The taxonomical classes follow (Hundsnurscher and Splett, 1982) with an additional class for pertainyms[3].

5.3 Verbs

Syntactic frames and particle verbs deserve special attention in the verbal domain. The frames used in GermaNet differ from those in WordNet, and particle verbs as such are treated in WordNet at all.

Each verb sense is linked to one or more **syntactic frames** which are encoded on a lexical rather than on a conceptual level. The frames used in GermaNet are based on the complementation codes provided by CELEX (Burnage, 1995). The notation in GermaNet differs from the CELEX database in providing a notation for the subject and a complementation code for obligatory reflexive phrases. GermaNet provides frames for verb senses, rather than for lemmas, implying a full disambiguation of the CELEX complementation codes for GermaNet.

Syntactic information in GermaNet differs from that given in WordNet in several ways. It marks expletive subjects and reflexives explicitly, encodes case information, which is especially important in German, distinguishes between different realizations of prepositional and adverbial phrases and marks *to*–infinitival as well as pure infinitival complements explicitly.

Particles pose a particular problem in German. They are very productive, which would lead to an explosion of entries if each particle verb was explicitly encoded. Some particles establish a regular semantic pattern which can not be accounted for by a simple enumeration approach, whereas others are very irregular and ambiguous. We therefore propose a mixed approach, treating irregular particle verbs by enumeration and regular particle verbs in a compositional manner. Composition can be thought of as a default which can be overwritten by explicit entries in the database.

We assume a morphological component such as GERTWOL (1996) to apply before the compositional process starts. Composition itself is implemented as follows, relying on a separate lexicon for particles. The particle lexicon is hierarchically structured and lists selectional restrictions with respect to the base verb selected. An example for the hierarchical structure is given in figure 7 (without selectional restrictions for matters of simplicity), where *heraus-* is a hyponym of *her-* and *aus-*. Selectional restrictions for particles include Aktionsart, a particular semantic verb field, deictic orientation and directional orientation of the base verb.

The evaluation of a particle verb takes the following steps. First, GermaNet is searched for an explicit entry of the particle verb. If no such entry exists the verb is morphologically analyzed and its semantics is compositionally determined. For example the particle verb *herauslaufen* in figure7 is a hyponym to *laufen* (walk) as well as to *heraus-*.

Criteria for a compositional treatment are separability, productivity and a regular semantics of the particle (see Fleischer and Barz (1992), Stiebels (1994), Stegmann (1996)).

6 Conclusion

A wordnet for German has been described which, compared with the Princeton WordNet, integrates principle–based modifications and extensions on the constructional and organizational level as well as on the level of lexical and conceptual relations. Innovative features of GermaNet are a new treatment of regular polysemy and of particle verbs, as well as a principle–based encoding of cross–classification and artificial concepts. As compatibility with the Princeton WordNet and EuroWordNet is a major construction criteria of GermaNet, German can now, finally, be integrated into multilingual large–scale projects based on ontological and conceptual information. This constitutes an important step towards the design of truly multilingual tools applicable in key areas such as information retrieval and intelligent Internet search engines.

[3]Adjectives pertaining to a noun from which they derive their meaning (financial, finances).

References

A. Alonge. 1996. Definition of the links and subsets for verbs. Deliverable D006, WP4.1, EuroWordNet, LE2-4003.

L. Bloksma, P. Diez-Orzas, and P. Vossen. 1996. User Requirements and Functional Specification of the EuroWordNet Project. Deliverable D001, WP1, EuroWordNet, LE2-4003.

G. Burnage, 1995. *The CELEX Lexical Database, Release 2.* CELEX – Centre for Lexical Information; Max Planck Institute for Psycholinguistics, The Netherlands.

C. Fellbaum. 1993. A Semantic Network of English Verbs. In G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, editors, *Five Papers on WordNet.* August. Revised version.

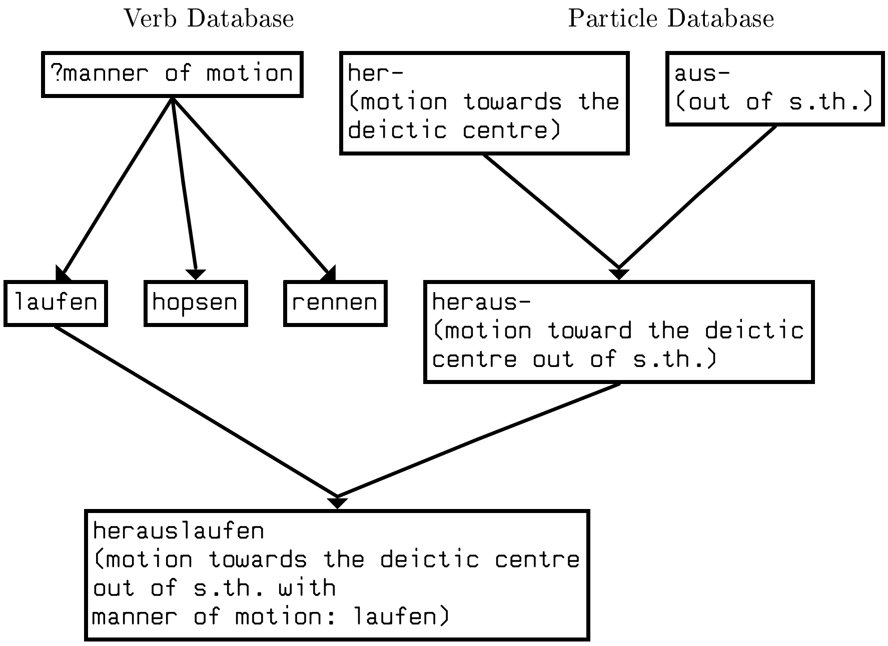

Figure 7: Particle Verbs

Wolfgang Fleischer and Irmhild Barz. 1992. *Wortbildung der deutschen Gegenwartssprache.* Max Niemeyer Verlag, Tübingen.

GERTWOL. 1996. German morphological analyser. http://www.lingsoft.fi/doc/gertwol/.

Franz Hundsnurscher and Jochen Splett. 1982. *Semantik der Adjektive des Deutschen: Analyse der semantischen Relationen.* Westdeutscher Verlag, Opladen.

European Corpus Initiative. 1994. European Corpus Initiative Multilingual Corpus.

B. Levin, G. Song, and B.T.S. Atkins. in press. Making sense of corpus data: A case study of verbs of sound. *International Journal of Corpus Linguistics*, page 41 pages.

G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1993. Five Papers on WordNet. Technical report, Cognitive Science Laboratory, Princeton University, August. Revised version.

James Pustejovsky. 1996. *The Generative Lexicon.* MIT Press.

Rosmary Stegmann. 1996. Semantic Analysis and Classification of Verbs of Direction. MA-Thesis. Seminar für Sprachwissenschaft, Universität Tübingen.

Barbara Stiebels. 1994. Lexikalische Argumente und Adjunkte. Zum semantischen Beitrag von verbalen Präfixen und Partikeln. Dissertation. Philosophische Fakultät, Universität Düsseldorf.