Extension and use of GermaNet, a lexical-semantic database

Claudia Kunze

Seminar für Sprachwissenschaft University of Tübingen Wilhelmstr. 113, 72074 Tübingen, Germany kunze@sfs.nphil.uni-tuebingen.de

Abstract

This paper describes GermaNet, a lexical-semantic network and on-line thesaurus for the German language, and outlines its future extension and use. GermaNet is structured along the same lines as the Princeton WordNet (Miller et al., 1990; Fellbaum, 1998), encoding the major semantic relations like synonymy, hyponymy, meronymy, etc. that hold among lexical items. Constructing semantic networks like GermaNet has become very popular in recent approaches to computational lexicography, since wordnets constitute important language resources for word sense disambiguation, which is a prerequisite for various applications in the field of natural language processing, like information retrieval, machine translation and the development of different language-learning tools.

1. Introduction

The lexical-semantic network GermaNet¹, which has been developed at the University of Tübingen, is filling the gap with respect to German semantic on-line resources. Within the second phase of the EuroWordNet project, GermaNet has been integrated into EuroWordNet², a multilingual semantic network for eight European languages (Vossen, 1999) which was subject to project-internal and external evaluations. GermaNet is being further enhanced as monolingual generic resource with regard to the coverage of data, the encoding of new lexical-semantic relations, and its applicability within natural language processing.

2. Basic features of GermaNet

GermaNet currently models almost 30.000 concepts, representing more than 40.000 word meanings, and its coverage is constantly being extended. Nouns, verbs, and adjectives are interconnected by their basic semantic relations. The concepts are derived from various monolingual resources, taking into account corpus frequencies. Like in WordNet, the central unit of representation for all lexical categories being implemented is the so-called synset, containing the set of synonymous word meanings that refer to the same concept. Semantic relations are established between concepts (synsets) or word meanings (single synonyms from synsets). The database contains an average of 1.4 synonymy (synonyms per synset) and 1.2 polysemy (word meanings per literal). Like WordNet, GermaNet distinguishes conceptual from lexical relations. Conceptual relations hold between entire synsets, among them the most fundamental hierarchical ones like hyperonymy/hyponymy, which link concepts with their superordinate terms, and meronymy/holonymy between parts and wholes. Further relations concern the cause relation between events and results, and the implication relation between events, a kind of backward presupposition.

Lexical relations, on the other hand, hold between individual synonyms (variants), like synonymy, which is implicitly encoded by the membership of two or more variants in the same synset, and the antonymy relation, which connects pairs of opposite word meanings. In GermaNet, there is an average of 2.4 relations per synyset.

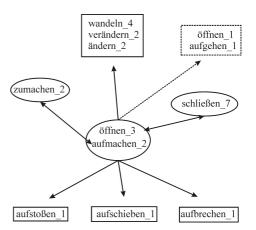


Figure 1. Semantic relations of the causative verb öffnen

Consider the example of the causative verb öffnen 'cause ot get open' in Figure 1 which is realized by the synset {öffnen_3, aufmachen_2}³ and represented with its semantically related concepts in GermaNet, its superordinate concept {wandeln_4, verändern_2, ändern_2} 'cause to change', indicated by the upward arrow, and three hyponyms *aufstoßen* 'push open', *aufschieben* 'slide open' and *aufbrechen* 'break open', indicated by downward arrows. There is a causal relation to the inchoative variant {öffnen_1, aufgehen_1} 'to

¹ GermaNet has been constructed within the SLD-project "Ressourcen und Methoden zur lexikalisch-semantischen Disambiguierung", funded by the Ministery of Research of Baden-Württemberg, Germany, in 1996-1997. The database was built by Helmut Feldweg and his co-workers: Valérie Béchet-Tsarnos, Birgit Hamp, Michael Hipp, Claudia Kunze, Karin Naumann, Susanne Schüle, Rosmary Stegmann, Karen Steinicke, Christine Thielen and Andreas Wagner. URL:http://www.sfs.nphil.uni-tuebingen.de/lsd

² The EuroWorNet project for constructing a multilingual semantic network has been carried out in two stages, EuroWordNet-1 (LE-4003) and EuroWordNet-2 (LE-4 8328), and was funded by the European Community. The whole project was coordinated by Piek Vossen (University of Amsterdam). All project report deliverables can be downloaded from the EWN-webpage: http://www.hum.uva.nl/~ewn/.

³ In this example, the literals are presented with their reading numbers.

get/become open', being illustrated by the dotted arrow pointing to the concept. Both synset variants of causative *öffnen* have different antonyms (*öffnen_3* \leftrightarrow *schließen_7*, and *aufmachen_2* \leftrightarrow *zumachen_2*); thus the variants and not the entire concepts are interrelated by the bidirectional arrows.

Although the design principles and the database technology have been adopted from Princeton WordNet, some principle-based modifications are applied in GermaNet with respect to the employment of artificial nodes in the hierarchy and cross-classification, to the syntax-semantics interface and to the treatment of adjectives.

2.1. Artificial concepts

In contrast to WordNet, GermaNet makes use of artificial concepts⁴ which may refer either to lexical gaps in the language or to non-lexicalized terms, which are introduced to balance the hierarchies and to avoid unmotivated co-hyponymy. Consider the example given in Figure 2 which contains two artificial concepts (?Schullehrer 'teacher of a certain type of school' and ?hierarchischer Lehrer 'teacher w.r.t. to an hierarchical position') that help to structure the partial network within the semantic field Lehrer 'teacher' more adequately. Following Cruse (1986), co-hyponyms should be, on the basis of an underlying similarity, incompatible to each other. For example, the hyponyms of Kind 'child' like Baby 'baby', Kleinkind 'toddler', Schulkind 'pupil' are mutually exclusive. Since a teacher of a certain subject (Fachlehrer) is also a teacher of a certain type of school and as well a teacher in a certain hierarchical position, the lexicalized hyponyms of Lehrer 'teacher' are not mutually exclusive. Incompatible hyponyms are therefore collected under the nodes of the corresponding artificial concepts.

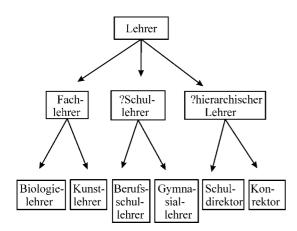


Figure 2. Artificial concepts in GermaNet

Artificial nodes are also introduced for verbs and adjectives. For example, GermaNet distinguishes for descendent nodes of the verb concept *essen* 'eat' between *?Art_essen* 'manner of eating' and *?Zeit_essen* 'time of eating', so that concepts like *schlingen* 'gobble' and

frühstücken 'to have breakfast' do not surface as co-hyponyms.

In GermaNet, adjectives are modeled following the same taxonomic approach (as opposed to the satellite approach to adjectives within WordNet) like for nouns and verbs. Since the hierarchical depth of adjectival taxonomies is quite flat compared to those of nouns and verbs, artificial concepts are necessary to head the subnetworks, i.e. in the field of adjectives of perception (*?farbspezifisch* 'w.r.t. colour' is the superodinate of the colour terms).

2.2. Cross-classification

In GermaNet, concepts, which belong to different hierarchies, are often cross-classified. So the concepts can be accessed according to different meaning aspects, i.e. the cross-classification of animals like *Wellensittich* 'budgerigar' as *Vogel* 'bird' and *Haustier* 'pet'.

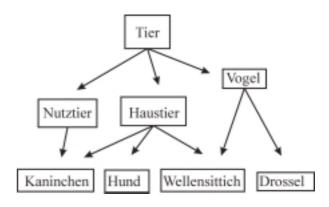


Figure 3. Cross-classification of animals

From a theoretical point of view, systematic crossclassification may help to detect productive patterns of regular polysemy, i.e. *Birke* 'birch' as tree and kind of wood or *Tennis* 'tennis' as sport discipline and event.

2.3. Subcategorization frames

GermaNet contains some 7.000 verbs for which subcategorization frames⁵ are provided, implying full disambiguation of the verb readings and accounting for verb alternations (Kunze, 1999) which are centered around the causation relation like

- 1) the causative-inchoative alternation
 - a) Er kocht die Suppe. 'He cooks the soup.'
 - b) *Die Suppe kocht*. 'The soup boils.'
- 2) the induced action alternationa) *Peter rollt einen Ball.* 'Peter rolls a ball.'
 - b) Der Ball rollt. 'The ball rolls.'
- 3) The change-of-location vs. change-of-possession alternation
 - a) Er gibt sein Mobiliar zu Peter.
 - 'He brings his furniture to Peter.'
 - b) Er gibt Peter sein Mobiliar.

⁴ Artificial concepts are identified by an initial question mark so that they can be retrieved automatically.

⁵ The representation templates of the syntactic frames are based on the complementation codes provided by CELEX (Burnage, 1995) with some minor modifications for reflexive and nominative arguments.

'He gives Peter his furniture.'

- 4) causative-change-of-integrity alternation
 - a) Er zerbricht die Tasse. 'He breaks the cup.'
 - b) *Die Tasse zerbricht*. 'The cup breaks.'

Other diathesis types like eg. the resultative alternation, which is feasible for activity verbs in the Vendlerian sense, are predictable and need not be accounted for (*Er i* β *t den Teller leer*. 'He eats the plate clean.').

The syntactic information should be enriched by providing for the respective selectional restrictions on verb complementation and the semantic roles being assigned to the arguments involved (see section 3.2.2).

3. Future Perspectives

Within a national project⁶ and several co-operations, we are following two main areas of future actions: the extension of our German wordnet, considering the guidelines and standards of the EuroWordNet project and the Global WordNet Association, and the use of GermaNet within NLP tasks.

3.1. Extending GermaNet

Extending GermaNet will imply the enhancement of the database with respect to the quantitative and qualitative coverage of concepts.

3.1.1. Quantitative enhancements

The number of synsets in the database shall be extended to 40.000 concepts, covering 60.000 word meanings. The corpus-based completion of implementing the German base vocabulary is currently being complemented by statistical measures of the overlap with other generic German resources like the PAROLE lexicon.

Frequent and common nominalizations of verbs and adjectives, the lack of which has already become obvious within the EuroWordNet framework, are being accounted for systematically.

We plan the treatment of further lexical categories like adverbs and functional categories, which have so far not been implemented in the database.

Furthermore, a terminological sample⁷ will be encoded for the field of economy, which serves to investigate the relationship between general language resource and terminological extensions, and enables work on the development of ontologies with regard to domain and world knowledge (Mädche & Staab, 2000).

3.1.2. Qualitative enhancements

Along with these tasks, besides the improvement of coverage, we are planning to establish new types of semantic relations in GermaNet. This concerns particularly the adoption of role pointers, which capture semantic roles like AGENT, PATIENT, INSTRUMENT, that are assigned by verbal predicates and adjectives. Role information supports verb sense disambiguation in cases where the syntactic frame is not sufficiently distinctive, i.e. for occurrences of verb alternations of which the patient role can surface either in the subject position or direct object position of the predicate, i.e. *Ball* 'ball' in the following examples:

(i) Er rollte den Ball. 'He rolled the ball.'(ii) Der Ball rollte. 'The ball rolled.'

We are intending to implement more fine-grained relational pointers for subtypes of meronymy⁸ and antonymy⁹. At least three types of meronymy should obtain different pointers: the physical part-whole-relationship (between *wheel* and *vehicle*), the member-group-relationship (between *leader* and *gang*) and the substance-composition-relationship (between *oxygen* and *atmosphere*).

After some further research, a semantic pointer for types of regular polysemy shall be realized.

3.2. Using GermaNet

Semantic wordnets constitute important resources for word sense disambiguation which may feed various applications like information retrieval, text categorization and automatic summarization, since the concept nodes and relational links among nodes can be used for making semantic inferences, for finding alternative expressions, and for expanding words to sets of semantically close concepts.

GermaNet will be studied within the framework of two different applications, exploiting the GermaNet synsets and relations for the acquisition of selectional preferences of verbal and adjectival predicates (Wagner & Kunze, 1999) and for the semantic annotation of large corpora (Buitelaar, 1999).

3.2.1. Acquiring selectional preferences

A verbal or adjectival predicate imposes semantic constraints, the so-called selectional restrictions, on the realizations of its arguments, i.e. *kochen* 'cook' requires a human agent and a patient which denotes some food.

The acquisition of selectional preferences of a predicate for its syntactic complements¹⁰ in a wordnet can aid syntactic and lexical disambiguation both for language processing and human use. Syntactic (i) and lexical (ii) ambiguity can be resolved by referring to selectional restrictions like in

(i) Die Suppe kocht die Frau. 'Soup cooks the woman.' (literal)

(ii) Die Frau kocht einen Auflauf.¹¹ 'The woman prepares a baked pudding.'

Only a few predicates have rigid selectional restrictions like *kalben* 'give birth to a calve' which selects *Kuh* 'cow'

⁶ A national funding for the extension of GermaNet is being provided by the Ministery of Research of Baden-Württemberg, Germany, until September 2001 ("Ausbau des GermaNet als nationale Ressource").

⁷ We already have included some 500 terms from the field of computing as part of a EWN-2 subtask.

⁸ The EuroWordNet specification allows for five specialised meronymy pointers with regard to being part of a location, being substance of a composition, being member of a group, part of a countable thing or part of a mass.

⁹ see Cruse (1986) for a detailed overview on different types of contrast

¹⁰ This investigation will be elaborated in detail in A. Wagner's Ph.D. thesis.

¹¹ Auflauf has also the meaning 'crowd'

as the sole argument, and can be recognized and marked by the lexicographer. The well-known semantic features like abstrakt 'abstract' and konkret 'concrete' are too general for covering the base vocabulary; more specific semantic properties for defining the constraints would be subject to controversial judgements. Statistical methods on analyzing the co-occurrences of predicates and complements in large text corpora (Resnik, 1993; Abe & Li, 1996) help to determine the adequate level of generalization, i.e. $Nahrungsmittel^{12}$ 'food' as preferred kochen 'cook'. Since candidate of the verb Nahrungsmittel can be preferred for both the subject and object complement of kochen (depending on the alternation variant), the selectional preferences should be mapped to the underlying semantic role (see McCarthy & Korhonen, 1998), i.e. the PATIENT role.

The statistical determination of selectional preferences of predicates ideally yields the semantic role preferences for being encoded or verified in GermaNet.

3.2.2. Semantic tagging

GermaNet is being applied in a test phase on the semiautomatic tagging of syntactically disambiguated sentences, which will provide a first step towards the development of reliable tag-sets for the semantic annotation of corpora¹³. Each literal of the syntactically annotated output is assigned the best match from among the GermaNet synsets. Anyway, several concepts are missing in the resource, and some meanings do not match exactly, since either GermaNet senses are too general or too fine-grained for capturing the exact meaning which the (human) sense-tagger has in mind.

For example, the literal *Geschichte* has 7 readings in GermaNet that can be semantically clustered and reduced to 3 senses (*history*: the accumulation of happenings in the past; *history* as school subject; *story*).

Different aspects of the same concept *book* like its physical representation as countable object and its content will not be treated as instances of regular polysemy, but of underspecification.

The degree of polysemy within GermaNet is quite low, but WordNet meanings require sense clustering techniques for reducing the semantic search space in information retrieval (Peters et al., 1998).

Consider the term *Mittelstandsanekdoten* 'anecdotes typical for members of middle classes' which was object to tagging and could not be detected in GermaNet. This ad-hoc composition, produced in the domain of the subculture music scene, is not lexicalized, though understandable in the given context. Accessing only the base noun Anekdote 'anecdote' may capture the more relevant meaning component, but the contribution of the first noun and the modus of combination will be lost. Thus, some morphological analysis would be very useful.

Feedback on missing literals or meanings in GermaNet is provided. Once the first tag-set is defined, the corpus training phase can be started.

4. Conclusion

This paper has presented the architecture of GermaNet, emphasizing its approach to artificial concepts and verb representation, and outlines our basic perspectives concerning the extension and use of the database for important tasks within NLP like sense-tagging and the acquisition of selectional preferences. Extending and improving GermaNet may support the respective applications, which, on the other hand, hint at deficiencies and inconsistencies of our resource. We assume that both lines of actions, database extension as well as test applications, will mutually benefit from one another.

5. References

- Abe, N. & H. Li (1996). Learning Word Association Norms Using Tree Cut Pair Model. In *Proceedings of* 13th International Conference on Machine Learning.
- Buitelaar, P. (1999). Concepts in Multilingual Information Retrieval. In *Proceedings of EUROLAN'99*, 252-262).
 4th European Summer School on Human Language Technology. Iasi, Romania. July, 1999.
- Burnage, G. (1995). *The CELEX Lexical Database, Release 2*. Max Planck Institute for Psycholinguistics: Nijmegen, The Netherlands.
- Cruse, D.A. (1986). *Lexical semantics*. Cambridge: Cambridge University Press.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.
- Kunze, C. (1999). Semantics of Verbs within GermaNet and EuroWordNet. In: Kordoni, E. (ed.), *Lexical Semantics and Linking in Constraint-Based Theories*. Workshop Proceedings of the 11th European Summer School in Logic, Language and Information, 189-200.
- Mädche, A. & S. Staab (2000). Semi-Automatic Engineering of Ontologies from Texts. In Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering (SEKE'2000).
- McCarthy, D. & A. Korhonen (1998). Detecting Verbal Participation in Diathesis Alternations. In *Proceedings* of the 36th Annual Meeting of the ACL. Vol. 2, 1493-1495. Montreal, Canada.
- Miller, G. & R. Beckwith & C. Fellbaum & D. Gross & K. Miller (1993). *Five Papers on WordNet*. CSL Report, Vol. 43. Cognitive Science Laboratory, Princeton University.
- PAROLE, German Resources (1998). Produced in LE 2-4017. Mannheim: Institut für deutsche Sprache.
- Peters, W. & I. Peters & P. Vossen (1998). The Reduction of Semantic Ambiguity in Linguistic Resources. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*. Granada, Spain.
- Resnik, P.S. (1993). Selection and Information: A Class-Based Approach to Lexical Relationships. PhD thesis, University of Pennsylvania.
- Vossen, P. (1999). EuroWordNet. Building a Multilingual Database with Lexical-Semantic Networks for the European Languages. In *Proceedings of EUROLAN'99*, 262-272). 4th European Summer School on Human Language Technology. Iasi, Romania. July, 1999.
- Wagner, A. & C. Kunze (1999). Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche. To appear in: Schröder, B. et al. (eds.), Probleme und Perspektiven computergestützter Lexikographie. Tübingen:Niemeyer.

¹² *Nahrungsmittel* corresponds to the same level of abstraction like a EuroWordNet base concept.

¹³ This application is being carried out by P. Buitelaar and his co-workers at the DFKI, Saarbrücken.