German Verb Patterns and Their Implementation in an Electronic Dictionary

Marc Luder

University of Zürich
Binzmühlestr. 14/16, CH-8050 Zürich
m.luder@psychologie.uzh.ch

Abstract

Here we describe an electronic lexical resource for German and the structure of its lexicon entries, notably the structure of verbal single-word and multi-word entries. The verb as the center of the sentence structure, as held by dependency models, is also a basic principle of the *JAKOB narrative analysis* application, for which the dictionary is the background. Different linguistic layers are combined for construing lexicon entries with a rich set of syntactic and semantic properties, suited to represent the syntactic and semantic behavior of verbal expressions (verb patterns), extracted from transcripts of real discourse, thereby lexicalizing the specific meaning of a specific verb pattern in a specific context. Verb patterns are built by the lexicographer by using a parser analyzing the input of a test clause and generating a machine-readable property string with syntactic characteristics and propositions for semantic characteristics grounded in an ontology. As an example, the German idiomatic expression *an den Karren fahren* (to come down hard on somebody) demonstrates the overall structure of a dictionary entry. The goal is to build unique dictionary entries (verb patterns) with reference to the whole of their properties.

Keywords: electronic dictionary, multi-word units, dependency parsing

1. Introduction

This paper describes an electronic lexical resource for German, especially spoken German, and the structure of verbal dictionary entries as used therein, based on principles of dependency linguistics. Motivation and background for this lexicon is a computer-based text and narrative analysis application. *JAKOB narrative analysis*[1] was conceived and developed at the University of Zurich as a systematic interpretative tool for research, documentation, and practical examination of everyday narratives in psychotherapy (Boothe et al., 2010). One step of the narrative analysis procedure is the coding of the vocabulary with specific psychodynamic categories revealing motives and conflicts of the narrator. The category system is represented in the lexicon as a property of the lexicon entries. Tape-recorded and transcribed discourse of psychotherapy sessions is analyzed sentence by sentence by the JAKOB application, words and expressions are matched with lexicon entries with the goal of coding the text with the appropriate categories. The main emphasis of the lexicon lies on verbs, verbal expressions, and verb patterns. Dependency and valency linguistic models put the verb in the center of the sentence structure; the arguments are *actants* and *circumstants*, described by Tesnière (1976) as "actors in a little drama". This stage metaphor, with the verb in the central position, is also a basic principle of the JAKOB narrative analysis, where a storyteller is presenting a dramaturgically constructed plot with different actions (verbs) by different actors playing different roles in the story (Boothe, 2011). Semantic and syntactic valency and dependency are most important, as are also the relations between the two.

[1] JAKOB narrative analysis (Erzählanalyse JAKOB): http://www.jakob.uzh.ch

2. Background

Semantic valency and dependency models are tightly connected to concepts of frame semantics. When describing properties of verb complements, extra linguistic factors like discourse context, situation, real-world knowledge, and knowledge about social practices are equally important and constrain the potential word combinations.

In the literature, there are different possibilities for the description of semantic valency. Semantic roles, also known as deep cases or theta-roles were introduced in *case grammar* (Fillmore, 2003). Number and properties of these roles vary widely between different studies about the concept. Authors in the tradition of *generative grammar* describe semantic components as selectional restrictions that specify the possible semantic properties of the arguments of a verb. In *Corpus Pattern Analysis* (CPA) (Hanks, 2008), semantic types are organized in a large hierarchically organized ontology (Hanks and Pustejovsky, 2005). Examples from the different hierarchical levels are *Anything, Physical Object, Human, Eventuality, Abstract*. A subset of this ontology defines the semantic types of nouns in the present lexicon project.

The idea of representing meaning in syntactic and semantic dependency structures is grounded in the insight that meaning is expressed by phrases and constructions rather than by single words (Hanks, 2010). Lexicon entries are therefore conceived as constructions in the sense of *construction grammar* - pairings of form and meaning (Goldberg, 1995). In the current lexicon project, we are combining syntactic and semantic properties of verb patterns, for one, explicating the semantic potential to the human reader, and for another, providing a machine-readable representation of the dependency structure for NLP applications. In the following section, we will describe the most relevant properties of the lexicon in more detail; a broader overview of the dictionary structure, the theoretical background, and the lexicon

building based on corpus data is provided by Luder (2011).

3. Data

The vocabulary of the JAKOB lexicon was developed mainly on the basis of spoken discourse from videotaped and transcribed psychotherapy sessions, recorded at our institute and by the *Ulmer Textbank* (Thomä and Kächele, 2006). One intention of this lexicon project was to integrate theoretical perspectives of conversation analysis and interactional linguistics with perspectives of construction grammar and corpus linguistics. Speakers use a big inventory of prefabricated expressions and patterns in their utterances, ranging from entirely fixed units to more loose phraseological constructs (Moon, 2008). The data background of the lexicon is therefore the language as used in our corpora of therapy conversations, with a complete size of ca. 3.8 million tokens. For the purpose of comparison, publicly available corpora were added to extend the data, including data from the *Archiv für Gesprochenes Deutsch*, *Schweizer-Textkorpus*, and a large Internet-based corpus, *deWaC*, integrated in the *Sketch Engine* (SKE) - a Corpus Query System incorporating automatic, corpus-derived summary of a word's grammatical and collocational behavior (Kilgarriff et al., 2004).

The exploration of linguistic research questions requires text corpora featuring linguistic information like morphological and syntactic information and lemmatization. The processing of the therapy scripts was done in the Sketch Engine, thus enhancing the corpora with these features.

In its current state, the JAKOB lexicon has the character of a prototype. There are a total of 7,000 lexicon entries, of which 1,200 entries are verb-based patterns, fully developed with syntactic and semantic characteristics as described above. Here some numerical data about the lexicon (February 2012):

- Total entries: 7,000
- multi-word units (MWU): 905
- verbs including a verb pattern: 1,273
- verbs: 3,927
- nouns: 2,048

4. Implementation

The lexicon implementation is based on the *OLIF format* (Open Lexicon Interchange Format)[2]. Table (1) shows the basic properties of an OLIF lexicon verbal entry – *Angst haben vor* (to fear something) – as used in the project (there are by far more OLIF properties, e.g., for morphology, translation, administration, etc.).

Syntactic properties are part of speech and syntactic type (e.g., main verb, modal verb, function verb). The OLIF property *syntactic frame* (synFrame) is replaced by the value of the sentence patterns described in the German dictionary *Der Kleine Wahrig* (Wahrig, 2007), because it provides a more precise syntactic categorization of the specific German verb valencies. Verbs and their different readings in different syntactic contexts are described by "Satzmuster" [sentence patterns] (example: 500 is the Satzmuster for a transitive verb: verb + direct object). These sentence patterns represent 76 different verb patterns with three-digit numbers. There is also an OLIF property *semType*, defining the semantic type of the entry, and a property *subjField*, describing the thematic context of the entry (domain or subject).

It is difficult to assign semantic and pragmatic properties to dictionary entries (namely, single-word entries) without contextual information; therefore it is necessary to build new formalisms for the representation of (verbal) dictionary entries (Luder and Clematide, 2010). We use the concept of *constructions* as pairings of form and meaning. The semantic type of a verb and particularly the semantic types of its arguments (constituting the verb pattern) define the meaning of the sentence (i.e., the nouns co-occurring with this verb). So we conceived additional properties for the lexicon entry representing dependency structures for verb patterns. They are summarized for the verb *lesen* (to read) in Table (2) and described below (these are selected properties; the overall structure of a lexicon entry can be reviewed online)[3].

4.1. Properties: natclause and bauplan

The properties *natclause* and *bauplan* were introduced to determine the syntactic structure of a verb pattern by means of a *parser*, whereby the property natclause is the input box for the parser *Pro3GresDe*[4] (Schneider, 2008; Sennrich et al., 2009). The output of the parser is automatically written into the input box of the property bauplan, where it can be edited and finally serves as a preset value for the property *pattern*.

The input string for *natclause* is a simple German sentence formulated by the lexicographer, representing the syntactic behavior of the verb in question, including the mandatory valency positions. For the variable argument positions, predefined placeholders are inserted into the sentence, in the current test phase restricted to infinite pronouns in different cases (*someone* and *something*, in German abbreviated as *jmd, jmdn, jmdm, etwas, etwasn, etwasm*). The parser was slightly adapted for this purpose. These placeholders are replaced after the parsing procedure by the semantic types *Human* (someone) and *Anything* (something).

Table (2) demonstrates the database values for the transitive verb "lesen" (to read). The property *natclause* contains the phrase "someone is reading something" (in German: "jmd liest etwasn"); the mentioned placeholders for subject and direct object are changed during postprocessing by coarse-grained semantic markers. The categories of the semantic types are taken from the shallow ontology also used in the CPA (Hanks and Pustejovsky, 2005).

The value of the property *bauplan* in Table (2) demonstrates the parser output including postprocessing, in simple serialized text format. The placeholders *someone* and *something* (jmd/etwas) are replaced respectively by semantic types from the ontology: someone is replaced by *Hu-*

[2] http://www.olif.net

[3] http://www.jakob.uzh.ch/lexikon/

[4] Pro3GresDe – a Dependency Parser for German: https://files.ifi.uzh.ch/cl/gschneid/parser/

OLIF property	Value	Description
canForm	haben Angst vor	canonical form
ptOfSpeech	verb	part of speech (of the head of MWU)
head	haben	verb as head for verb patterns
phraseType	set-phrase	type of MWU
synFrame	550 - verb + AkkO + PrepO	Satzmuster (Wahrig, 2007)
synType	function verb	syntactic behavior
semType	emotion	semantic type (OLIF)
definition	Angst verspüren vor etwas, etwas fürchten	free textual definition
subjField	general (therapy discourse)	domain, genre

Table 1: OLIF properties (Example: *Angst haben vor etwas* (to fear something))

Attribute	Value	Description
canForm	lesen (to read)	canonical form
natclause	jmd liest etwasn	prototypical clause, placeholders for arguments
bauplan	500: 1 - - L2 subj Human 2 lesen VVFIN L0 root n/a 3 - - L2 obja Document 4 ber APPR L2 pp n/a 5 - - L4 pn Anything	machine-generated pattern string
pattern	[[Human]] lesen ([[Document]]) ({über [[Anything]]})	human readable verb pattern (Hanks, 2008)

Table 2: Additional Lexical Properties for Verb Entries (Example: *lesen* (to read))

man, and something is replaced by *Anything*. The lexicographer then has to decide about the appropriateness of the semantic types and normally has to edit and refine them by choosing a more convenient type. On the head of the string *bauplan*, Wahrig's (2007) three-digit number of the Satzmuster gives information about the syntactic properties of the verb pattern. The string *bauplan* is composed of the three-digit Wahrig number, followed by 6 values for each word of the natclause string:

1. sequential number
2. canForm (lemma)
3. part of speech
4. link to root position (L0)
5. syntactic information
6. semantic type

4.2. Property: pattern

The property *bauplan* should provide all necessary information required by NLP applications, but the format of this property is confusing and not easy to read. So we implemented a property *pattern*, corresponding to the English verb patterns proposed by Hanks (2008) and presenting the verb and its arguments and possible semantic types in human-readable format. *Verb patterns* represent the semantic properties of all the elements of a construction, and the meaning of the pattern as a whole is composed from the semantic types of the single elements. Therefore, the description of the pattern meaning is more accurate than the OLIF property *semType*, which describes the semantic type of the verb expression as a whole.

4.3. Example lexicon entry

An elaborated example of a multi-word lexicon entry is presented in Table (3) for the German idiomatic expression "an den Karren fahren" (to come down hard on someone).
Besides the mentioned new semantic and syntactic features of the lexicon entries, we decided to introduce extended pragmatic properties. The original OLIF property *subjField* represents the knowledge domain to which the lexical entry is assigned, according to the language used in different domains, e.g. *agriculture*, *audiovisual*, *aviation*. As this property does not cover the required spectrum of pragmatic descriptions suited for disambiguating polysemous expressions, we extended the OLIF structure with the pragmatic/functional properties *textType*, *register* and *topic*. The property *subjField* marks the larger context of a situation (external, nonlinguistic criteria), whereas *textType* (also discourse type, communication type) represents internal linguistic criteria like *narrative*, *description* or *argumentation*. The values for *register* relate to local patterns of style and social situation, e.g. *formal*, *informal*, *ironic*, *colloquial*. The values for *topic* eventually are very important for local meaning constitution, together with *subjField* they denote the local subject of the current discourse unit with heterogeneous items like *cooking*, *job*, *sports*, as collected by the analysis of the transcripts. Unfortunately, these pragmatic features are very difficult to assign to the entries without the context of discourse, so their contribution to meaning construction is limited.
The lexicon entry "an den Karren fahren" was modeled on the occurrence of this idiomatic expression in the psychotherapy corpus. In the 2.5 million corpus of psychotherapy sessions of one specific patient, the expression was used 13 times, always in the meaning of *communication with aggressive self-assertion*. This expression can be used as a flexible formulation for different purposes (Table 3).

Attribute	Value	Description
canForm	fahren an den Karren	canonical form
head	fahren	head of MWU
ptOfSpeech	verb (head of verbal MWU)	part of speech
single/multi	phr: idiom	phrase and phrase type for MWU
Semantics and Pragmatics:		
subjField, textType,	general, undefined	unspecific without context
register, topic	umg (colloquial), undefined	colloquial style, topic nonspecific
polarity	NEG (1.0)	negative polarity, default score
function/definition	communication with aggressive self-assertion	characteristics found in transcript
semType	act (unspecific activity)	semantic type (whole expression)
natclause	jmd fährt jmdm an den Karren	prototypical clause
bauplan	605: 1 - - L2 subj Human 2 fahren VVFIN L0 root n/a 3 - - L2 objd Human 4 an APPR L2 pp n/a 5 die ART L6 det n/a 6 Karren NN L4 pn n/a	
pattern	[[Human]] fahren [[Human]] {an den Karren}	verb pattern
Frame	ATTACK, angreifen	categorization from FrameNET
Dornseiff Sachgruppen	9.61 Beschädigen	(Dornseiff, 2004)
JAKOB Code	SIG-ATT (speak and attack)	Coding for JAKOB narrative analysis
Cross References	an den Karren pinkeln, ins Gehege geraten	different types of cross references (OLIF)
Syntax:		
part of speech	verb	
synType, transtype	main-verb, intransitive	syntactic behavior
synFrame	650 - DatO + PrepO	Satzmuster (Wahrig, 2007)
preposition	an	required preposition
auxType	sein	auxiliary (to be)

Table 3: Elaborated Example: *an den Karren fahren* (to come down hard on somebody)

For the purpose of comparison, Table (4) demonstrates all the different constructions with the element *Karren* (cart) and their respective frequencies in the mentioned corpora[5].

The OLIF property *cross reference* allows for a multitude of cross reference types, like synonymy, near-synonymy, and diverse association and connotation types, e.g. for constructions with similar or complementary meanings in active and passive forms, like *ins Gärtchen trampen* (to tread on so.'s toes), or *ins Gehege geraten* (to get in each other's way).

The lexicographic work is also comprehending the task of analyzing the corpus data for related meanings and collocations. The three most salient collocation candidates for the nominal expression *an den Karren* in the deWAC-corpus are:

- pinkeln (4 / 5.8), (to pee)
- fahren (65 / 2.5), (to drive)
- spannen (4 / 2.49)[6], (to put to, to yoke)
- (Total 92)

[5]Corpora WIL, AMA, GUS = Psychotherapy discourse, AGD = Archiv für Gesprochenes Deutsch, CHTK = Schweizer Textkorpus, deWAC = Internet based corpus

[6]Frequency and association score from the *Sketch Engine* (logDice score).

5. Conclusion

We present a new dictionary entry structure, combining syntactic, semantic, and pragmatic properties in verb entries for single-word and multi-word units in the context of their patterns of usage.

The verb patterns allow for a more accurate syntactic and semantic description of verbal expressions in the dictionary, therefore reducing the ambiguity between different yet related entries. In the JAKOB narrative analysis application, the same parser used for building the lexicon is also used for parsing the text input from transcripts. The structures of the analyzed text sentences and the structures of lexical verb entries are compared, and patterns of text sentences are matched with patterns of the dictionary.

This matching process is functional as long as there are appropriate patterns in the lexicon and as long as these patterns discriminate between different usages of verbal expressions. The current state and size of the lexicon has the character of a prototype. The building of the lexicon is a time-consuming task, and the further development will be rather slow, because we add new entries only provided that we can find them in real discourse. The final size of the lexicon will therefore grow with the applications using it. Besides the JAKOB narrative analysis, these could be different text mining tools, i.e., for sentiment analysis (Klenner, 2009; Klenner et al., 2012), polarity rating, opinion mining, or, with appropriate extensions, any application based on a priori categorization of linguistic terms.

Construction / Corpora	WIL	AMA	GUS	AGD	CHTK	deWAC
an den Karren fahren	13	0	0	0	0	62
den Karren ziehen	1	0	0	0	3	33
den Karren aus dem Dreck ziehen	0	0	0	0	1	75
den Karren schleppen	1	0	0	0	0	4
sich vor den Karren spannen lassen	0	0	0	0	0	129
vor den Karren spannen	0	0	0	0	0	98
den Karren in den Dreck fahren	0	0	0	0	0	44
Karren (simple noun)	15	5	0	52	125	3598

Table 4: Frequencies of Constructions with *Karren* (cart)

6. References

Brigitte Boothe, Geneviève Grimm, Marie-Luise Hermann, and Marc Luder. 2010. JAKOB Narrative Analysis: The psychodynamic conflict as a narrative model. *Psychotherapy Research*, 20(5):511–525.

Brigitte Boothe. 2011. *Das Narrativ: Biografisches Erzählen im psychotherapeutischen Prozess*. Schattauer, Stuttgart.

Franz Dornseiff. 2004. *Der deutsche Wortschatz nach Sachgruppen*. de Gruyter, Berlin, 8 edition.

Charles J. Fillmore, editor. 2003. *Form and Meaning in Language: Papers on semantic roles*, volume 1 of *CSLI lecture notes*. CSLI Publications, Leland and CA.

Adele E. Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. Univ. of Chicago Press, Chicago.

Patrick Hanks and James Pustejovsky. 2005. A Pattern Dictionary for Natural Language Processing. *Revue Francaise de Langue Appliquée*, 10(2):1–19.

Patrick Hanks. 2008. Lexical Patterns: From Hornby to Hunston and beyond. In Elisenda Bernal and Janet DeCesaris, editors, *Proceedings of the XIII. Euralex International Congress*, pages 89–129, Barcelona.

Patrick Hanks. 2010. How People Use Words to Make Meanings. *NLPCS 2010 Proceedings*, pages 1–11.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. *Proc EURALEX 2004, Lorient, France*.

Manfred Klenner, Stefanos Petrakis, Simon Clematide, and Marc Luder. 2012. Compositional Syntax-based Phrase-level Polarity Annotation for German. *Linguistic Issues in Language Technology - LiLT*, 7(15).

Manfred Klenner. 2009. Süsse Beklommenheit, schmerzvolle Ekstase: Automatische Sentimentanalyse in den Werken von Eduard von Keyserling. In Christian Chiarcos, Richard Eckart de Castilho, and Manfred Stede, editors, *Von der Form zur Bedeutung: Texte automatisch verarbeiten*, pages 91–97. Gunter Narr, Tübingen.

Marc Luder and Simon Clematide. 2010. Constructing a Constructional MWE Lexicon for psycho-conceptual Annotation: Evaluation of CPA and DuELME for Lexicographic Description. In Anne Dykstra and Tanneke Schoonheim, editors, *Proceedings of the XIV Euralex International Congress*, pages 152–153, Leeuwarden.

Marc Luder. 2011. *Konstruktionen im Lexikon - Konstruktionen in der Erzählanalyse*. BOD, Hamburg.

Rosamund Moon. 2008. Sinclair, Phraseology, and Lexicography. *International Journal of Lexicography*, 21(3):243–254.

Gerold Schneider. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Ph.D. thesis, Institute of Computational Linguistics, University of Zurich.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A New Hybrid Dependency Parser for German. In Christian Chiarcos, Richard Eckart de Castilho, and Manfred Stede, editors, *Von der Form zur Bedeutung: Texte automatisch verarbeiten*, pages 115–124. Gunter Narr, Tübingen.

Lucien Tesnière. 1976. *Eléments de syntaxe structurale*. Klincksieck, Paris, 2nd edition.

Helmut Thomä and Horst Kächele. 2006. *Psychoanalytische Therapie: Praxis*, volume 2. Springer, Heidelberg, 3rd edition.

Gerhard Wahrig. 2007. *Der kleine Wahrig: Wörterbuch der deutschen Sprache*. Bertelsmann, München, 4th edition.