Constructing a Constructional MWE Lexicon for Psycho-Conceptual Annotation: An Evaluation of CPA and DUELME for Lexicographic Description

Marc Luder, Department of Psychology, University of Zurich, Swiss
Simon Clematide, Institute of Computational Linguistics, University of Zurich, Swiss

The German JAKOB lexicon provides a basis for the coding of patient narratives and is currently extended in the direction of a phraseological and construction-grammar resource. For this purpose, we will compare two formalisms for the representation of multiword expressions (MWE): The Dutch Electronic Lexicon of Multiword Expressions (DuELME, Grégoire 2009) and the verb patterns from Corpus Pattern Analysis (CPA, Hanks 2008). We are looking for a representation format which is human-readable, and equally adapted for natural language processing (NLP). The JAKOB lexicon is implemented in the OLIF format and currently contains 7000 entries. The MWEs investigated are verbal phraseologisms and originate from the corpora of three different clients, consisting of a total of more than 400 transcribed sessions.

The narrative analysis method JAKOB is a tool for investigating everyday stories from psychotherapy transcripts (Boothe 2004). Stories are annotated on the basis of our predefined psycho-conceptual coding system represented in the lexicon. JAKOB allows formulating hypotheses about the client's conflicts, the analysis of the discourse being one component thereof.

DuELME is an NLP lexicon project which encodes MWE descriptions in a theory- and implementation-independent way. Every MWE is an instance of a construction class with elements including morpho-syntactic parameters. CPA patterns represent semantic properties for the elements of a (verbal) construction, whereas syntactic properties are represented in the JAKOB lexicon by the subcategorization frames (Satzmuster) of Wahrig (2007). We are implementing an additional lexicon property 'bauplan' which is formally constructed as a combination of the DuELME component list, the Wahrig subcategorization frame and semantic information out of the CPA-pattern. Because this structure is difficult to read for the lexicographer, it is generated automatically and can be hidden from the user, but is available for NLP tasks.

1. Introduction

The German JAKOB[1] lexicon provides a basis for the coding of patient narratives and is currently extended from an inventory of single words with contextually different senses in the direction of a phraseological and construction-grammar resource. In this paper, we will discuss lexical information, which is useful and essential from a phraseological point of view, and which is required from the point of view of the automatic coding procedure. For this purpose, we will compare two formalisms for the representation of multiword expressions (MWE) with regard to German MWE: the Dutch Electronic Lexicon of Multiword Expressions (DuELME, see Grégoire 2006; 2007; 2009) and the verb patterns from Corpus Pattern Analysis (CPA, see Hanks & Pustejovsky 2005; Hanks 2008).

The narrative analysis method JAKOB is a tool for investigating everyday stories from psychotherapy transcripts (Boothe 2004), these stories are dramaturgically constructed and enacted narrative episodes, by which clients are staging themselves and displaying wishes, fears, and defenses. Narratives are extracted from transcripts and manually segmented into simple sentences (subject-predicate units), words are lemmatized and POS-tagged automatically, and each segment is divided into syntactic units (slots) based on the simple pattern 'who does what how?' (*wer tut was wie*) or 'what happens to whom and how?'. The story vocabulary is then annotated on the basis of our predefined psycho-conceptual coding system represented in the lexicon. Finally, the narrative analysis according to JAKOB allows formulating hypotheses about the client's conflicts, the analysis of the discourse being one component thereof.

[1] See http://www.jakob.uzh.ch/lexikon for online access to the lexicon.

The JAKOB lexicon is implemented in the OLIF format[2] and currently contains 7000 entries (6000 single-word entries, 1000 multi-word expressions). The MWEs are conceived as constructions, i.e. pairings of form and meaning, based on the notion that meaning arises from the collocational context, not by summarizing the meanings of single words (Croft 2001). The MWEs investigated in this paper are verbal phraseologisms (idioms, collocations, set phrases) and originate from the corpora of three different clients, consisting of a total of more than 400 transcribed sessions (approx. 5 million tokens). The lexicon content is therefore based on spoken language, transcribed from psychotherapy sessions (German, Swiss German dialect variation).

The paper is designed as follows: In section 2 we give a short introduction to an OLIF lexicon entry with the properties proposed by the OLIF standard. Section 3 describes the pattern formalisms of DuELME and CPA and the application of these patterns for sample lexicon entries. The assets and drawbacks (pros and cons) of DuELME and CPA for lexical purposes are discussed in section 4. Finally, we propose a possible combination of the two formalisms and show the opportunities for lexicographical tasks (section 5).

2. OLIF lexicon entry: status quo

Table 1 shows the basic properties of an OLIF lexicon entry as used in our project (there are more OLIF properties, e.g. for morphology, translation, administration, etc.). Note, that the syntactic frame contains all argument positions, i.e. the ones which are internal and external to an MWE.

OLIF Property	Value	Description
canForm	haben Angst vor	canonical form
crossRef	fürchten; near synonym	cross references, types *e.g.* synonym, antonym, *etc.*
ptOfSpeech	Verb	part of speech (head of MWE)
head	Haben	
phraseType	set-phrase	type of MWE
synFrame	550 (verb + AkkO + PräpO)	"Satzmuster" (Wahrig 2007)
synType	function verb	syntactic behavior
semType	Emotion	semantic type (OLIF)
definition	Angst verspüren vor etwas, etwas fürchten.	free text definition
subjField	general (therapy discourse)	domain, genre

Table 1. Sample OLIF entry

3. MWE and their formal description

The need for lexicographic descriptions of speech units spanning more than one word arises if they involve mutual morphological, syntactic, or semantic idiosyncrasies (Moszczynski 2007). Therefore, restrictions on admissible or forbidden modifications, lexical variability, and allowed syntactic variations (e.g. passive transformation, negation, relative clauses) have

[2] URL: http://www.olif.net. Details on our implementation are described in (Luder, Clematide & Distl 2008).

to be stated. Similar to the descriptions in treebanks[3] and to their query languages, we can encode the syntactic structure of MWEs through the relations of labeled dominance, dependency (syntactic functions), and linear precedence. We may even recycle their relatively theory-neutral categories and labels.

We are looking for a representation format which satisfies all these needs, is human-readable, and equally adapted for natural language processing (NLP).

Although the OLIF standard allows for multi-word entries, it has no recommended formalism to specify the constituent structure of the MWEs and their idiosyncratic properties. The description level of OLIF does not offer much more than a multilingual terminological database.

For German, PhraseManager (Pedrazzini 1994) and Phraseo-Lex (Keil 1997) are two dedicated academic software solutions in the sense of a complex lexicographic MWE workbench. Because they are closed systems, it's difficult to integrate them into our lexicon.
Due to their compact and textual descriptions, DuELME and CPA are two possible candidates for MWE representations. We will present their concepts in the following sections.

3.1. DuELME

DuELME is an NLP lexicon project which tries to encode MWE descriptions in a fairly theory- and implementation-independent way. Similar as in PhraseManager (Pedrazzini 1994), a strictly class-based approach called ECM (equivalence class method) is used. Therefore, every MWE is an instance of a construction class. If these classes are fine-grained, the danger of inconsistent lexicographic descriptions increases, as the lexicographer may lose oversight of hundreds of classes. For this reason, the formalism is enhanced by introducing parameters in order to specify variable morpho-syntactic constraints on the expression level (cf. Grégoire 2006).

The linguistic description of the classes and the parameters (called 'patterns') expresses the constituency and dependency structure (including information on modifiability) and contains numbered slots for the actual lexical components, which, of course, need to be specified on the individual expression level. This is called the 'component list' (CL). To be able to fill the pattern slots with the numbered referents of the component elements, a canonical serialization of the MWE is crucial. The CL differs from a traditional lexicographic head word (which is also present under the label 'expression') with respect to explicitly expressed linguistic features.

In table 2, two original Dutch verbal patterns in bracketed notation from the DuELME lexicon[4] are shown. Figure 1 shows the corresponding tree structures. The example pattern for the entry 'angst voor' *(be afraid of sth.)* in table 2 means: 'angst' is a modifiable direct object of the verb 'hebben' (this is specified in a separate list), where the NP contained in the PP is not restricted ('var'). The example 'angst aanjagen' *(to scare sb.)* encodes the information that the direct object has to be in singular with an empty determiner ('EMP') and that the verb is a particle verb ('[part]'). The parameters specifying an element of the CL are written in separate brackets after each CL.

[3] For German, e.g. see the TIGER treebank: http://www.ims.uni-stuttgart.de/projekte/TIGER.

[4] URL: http://duelme.inl.nl.

Modifiability is expressed by special head categories. E.g., 'N' allows nominal heads to be modified, whereas 'N1' forbids any modification of the head. Further material not present in the CL, e.g. common modifiers or a list of allowed verbs, can be added in separate database fields ('List').

Expression:	angst aanjagen
CL:	EMP angst[sg] aan_jagen[part]
Pattern:	[.VP [.obj2:NP (var)] [.obj1:NP [.det:D (1)] [.hd:N (2)]] [.hd:V (3)]]
Description:	Expressions headed by a verb, taking (1) a variable indirect object and (2) a direct object consisting of a fixed determiner and an unmodifiable noun.
Expression:	angst voor
CL:	angst voor
List:	hebben
Pattern:	[.VP [.obj1:NP [.hd:N1 (1)]] [.hd:V (list)] [.pc:PP [.hd:P (2)] [.obj1:NP (var)]]]
Description:	Expressions headed by a verb, taking (1) a direct object consisting of a modifiable noun, and (2) a PP-argument consisting of fixed preposition and a variable complement (list).

Table 2. Sample MWE entries from DuELME

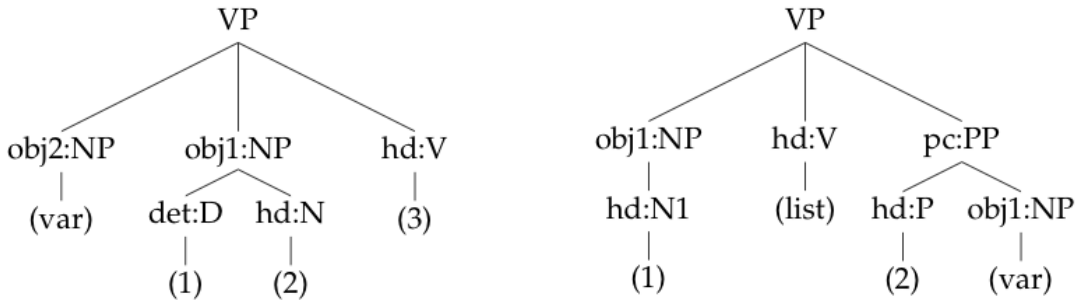

Figure 1. Tree structure of the two original DuELME examples 'EMP angst aanjagen' and 'angst voor'

It is important to note that the basic syntactic category of every component of a MWE is explicitly specified through the part-of-speech tags of the component list. Although sometimes parts of fixed MWEs have lost their specific syntactic category, in most cases the syntactic function is still transparent.

The main tasks of the lexicographer are (1) to identify the correct pattern for a candidate MWE entry, (2) to determine the parameters and lists needed for the pattern, (3) to construct the correct CL. Grégoire (2006) showed for Dutch that 11 parameterized equivalence classes cover 90% of the verbal constructions with 3 up to 4 words from an idiom dictionary.

3.2. CPA

Corpus Pattern Analysis (CPA) is a means to build a Corpus Pattern Dictionary (Hanks & Pustejovsky 2005; Hanks 2008) to associate meanings to verb patterns. The Pattern Dictionary provides prototypical syntagmatic structures occurring with English verbs. Meaning is associated with patterns rather than with single words and results from word use in specific phrasal, syntactic and semantic contexts. Verb patterns are linguistically and statistically significant collocations extracted from corpora; not every word co-occurrence is a pattern. The theory of *Norms and Exploitations* (TNE, see Hanks 2004) postulates that each verb pattern in normal use has its individual and exclusive meaning. Verb patterns are related to argument and valency structures and semantic frames. This is a promising approach for word sense disambiguation.

We present an original example entry from the English Pattern Dictionary[5]:

[[Human]] plug {hole} {(up)} {(with [[Stuff]]) | (with [[Physical Object]])}.

The meaning of the pattern is defined as a paraphrase created by the lexicographer (implicature): [[Human]] closes {hole} by filling it with [[Stuff | Physical Object]].

The concepts in double squared brackets represent semantic types and originate from the CPA ontology, a shallow semantic ontology (Pustejovsky, Hanks & Rumshisky 2004). Syntactic constituents are displayed in curly brackets, optional parts in parentheses. Simple squared brackets designate phrasal and adverbial categories (Hanks 2008). The overt representation of subject fillers in the canonical form is characteristic of CPA.

A simple grammar formalism defines the rules of pattern building. As an illustration we present four basic grammar rule snippets (cf. Pustejovsky et al. 2004, adapted by the authors):

- The sequence of constituents in the pattern is fixed (subject – verb – objects – complements – adverbials) (SPOCA)[6].
- *Pattern -> Segment verb Segment | verb Segment | Segment verb* (Grammatical rewriting rule: A pattern consists of different segments around the verb).
- *Segment -> Element | Segment Segment | '{'Segment'}' | '('Segment')'*. A segment is a single element or it consists of multiple segments, a segment can be a constituent, and it can be optional.
- An element can be a whole phrase, a semantic type, or an individual word (Hanks 2008).

3.3. Suitability and adaption of CPA and DuELME

As an example for our investigation we chose the pattern ‘Angst haben’ (*to be scared, to be afraid (of)).* Constructions including this expression occur 144 times in the mentioned corpora, as in the example ‘Herrgott nochmal, man kann einfach zu viel Angst haben vor Sachen, nicht?’ *(For God's sake, one can be simply too much afraid of things, can’t one?).*

The following verb patterns were found in the corpora:

a) ‘Angst haben’ *(to be afraid)*
b) ‘furchtbar Angst haben’ *(to be terribly scared)*
c) ‘Angst haben vor etwas’ *(to be afraid of sth.)*
d) ‘Angst haben um jemanden’ *(to worry about sb.)*

These verbal expressions may be represented as different CPA patterns as follows:

a) ‘(grosse) Angst haben’: [[Human | Animate]] haben {([ADJ]) {Angst}}. Subject of this pattern are humans or generally animate creatures, the word ‘Angst’ is used without determiner, but can be modified optionally by an adjective.
b) ‘(furchtbar) Angst haben’: [[Human | Animate]] haben {Angst} {furchtbar}. This pattern has an adverbial modifier.

[5] URL: http://nlp.fi.muni.cz/projekty/cpa/.

[6] SPOCA: see Hanks (2008: 94): ‘Complement is a clause role that is co-referential with either the subject or the object of the clause.’ Example: the adjective *happy* in *he seems happy.*

c) ‘Angst haben vor etwas’: [[Human]] haben {([ADJ]) {Angst}} {vor [[Anything]]}. This pattern has an additional object with the preposition ‘vor’. One can be afraid of anything, from humans to animals, situations, or facts.
d) ‘Angst haben um jemanden’: [[Human]] haben {([ADJ]) Angst} {um [[Anything]]}. This pattern has an additional object with the preposition ‘um’.

For us, the semantic types (e.g. [[Human]]) function as prototypical fillers, and not as obliging restrictions. Thus, prototypical categories can be passed over (overwritten) by exploitations of the semantic core meaning. Generally speaking, CPA has a more semantic flavor. The patterns give no explicit characterization of the syntactic category of the components of the MWE.

The main task for the adaption of DuELME is to conceive a set of equivalence classes which build upon our canonical forms that are formulated according to the OLIF guidelines. Table 2 shows an attempt to encode our example phrases replacing the DuELME labels stemming from a Dutch Treebank by labels from the TIGER corpus[7].

The lack of overt subject positions in DuELME prevents the use of the patterns themselves as subcategorization frames. However, in our lexicon we use the subcategorization classification codes from Wahrig (2007), which also serve as basis for our pattern classes.

Expression (a): CL: Pattern 500: Description:	haben Angst haben Angst[sg][uncountable] [.VP [.HD:V (1)] [.OA:NP [.HD:NN (2)]]] Expressions headed (1) by a verb, taking (2) a direct object consisting of a modifiable noun.
Expression (b): CL: Pattern 513: Description:	haben furchtbar Angst haben Angst[sg] furchtbar [.VP [.HD:V (1)] [.OA:NN (2)] [.MO:ADV (3)]] Expressions headed (1) by a verb, taking (2) a direct object consisting of a bare noun and (3) a fixed adverbial modifier.
Expression (c): Expression (d): CL: Pattern 550: Description:	haben Angst vor haben Angst um haben Angst[sg][uncountable] vor haben Angst[sg][uncountable] um [.VP [.HD:V (1)] [.OA:NP [.HD:NN (2)]] [.OP:PP [.HD:APPR (3)]]] Expressions headed (1) by a verb, taking (2) a direct object consisting of a modifiable noun and (3) a PP-argument containing a fixed preposition.

Table 3. Adaption to DuELME

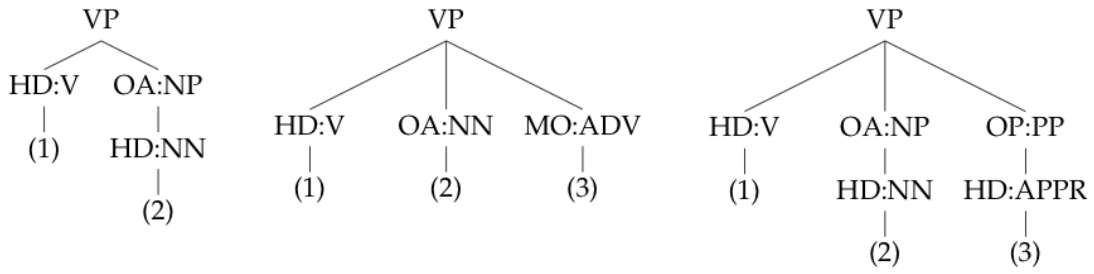

Figure 2. Tree structures of the German sample patterns.

[7] We deviate from the part-of-speech tagset in the case of verbs. As using the standard tags for verbs which include morphological information (finite or base form) does not make sense in a lexicon we simply notate ‘V’.

4. Discussion

The representation of the sentence patterns according to CPA has several advantages: The formalism is rather simple; the patterns are well-readable by human users. The verb patterns are widely compatible with the four 'slots' of the text analysis application (and also to the grammatical structure of subject – predicator – object – complement – adverbial). Matching the semantic types of the CPA ontology with the semantic types from OLIF is easy to do. The CPA patterns add a more semantic perspective to the syntactical valency pattern of the lexicon entry.

There are also some shortcomings: CPA was developed for English verb patterns; in German the constituent order is rather free. Therefore, it would be a good extension to German verb patterns to include grammatical information about syntactic functions or cases. Another weak point: Because the pattern structure is not strictly fixed by a class system, the lexicographer is free to build individual patterns for an entry. Automatic processing and parsing of the patterns will be rather difficult because of the somewhat ambiguous and non-systematic grammar rules. DuELME offers a more fine-grained, but less readable lexicographic specification language. Although in principle, the mechanism of parameters allows the injection of any entry-specific information needed in combination with the pattern, there are quite a lot of additional database fields (e.g. concerning subject or object realization) regulating the behavior of an entry in the original DuELME.[8] This complicates the adaption of DuELME for our lexicon, but probably this is just a shortcoming of the current implementation. The class-based approach allows for a systematic and simple integration into NLP applications. Semantic constraints as found in CPA are not used in DuELME. However, it's uncomplicated to integrate semantic parameters into the component list. Constraints on subjects could be inserted as parameters into the head of a MWE. For the example expression (a), this could be formulated as 'haben[sb-animate] Angst[sg][uncountable]'. The parameter '[sb-animate]' would express that the subject of the verbal head is an animate entity.

The formal patterns need a language description model for German. One choice would be an adapted set from the OLIF standard, another choice, as used in table 2, is the TIGER annotation language. The OLIF categories are not particularly well-adapted to German. On the other hand, the TIGER set of phrases and functions may be somewhat too specific. Unfortunately, the development of standardized and widely accepted data categories as proposed in the framework of ISO 12620 is still in a very provisional state.[9]

The aim of the current project is to find the appropriate pattern design for disambiguating lexical entries. For evaluation purposes we plan to extend a tenth of our 1000 MWEs with CPA and DuELME patterns. Possible solutions could on the one hand be to extend CPA patterns with constituent information and maybe to classify them, on the other hand to extend the DuELME formalism with subject information and semantic classes.

[8] One may speculate whether this was a design decision from the beginning, or motivated through the evolving practical needs.

[9] See for example the categories which have been entered into the ISO data category repository for ISO 12620 accessible from http://www.isocat.org.

Example 1: CPA pattern with constituent information:
[[Human]]$_{SB}$ haben {([ADJ]) {Angst}}$_{OA}$ {vor [[Anything]]}$_{OP}$ [10]

Example 2: DuELME component list with semantic information:
haben[sb-animate] Angst[sg][uncountable] vor[anything]

A third proposal will follow in the next section.

4.1. Integration of DuELME, CPA, and Wahrig subcategorization frames

Verb patterns represent semantic properties for all the elements of a (verbal) construction, which is very important, because pattern meaning often depends on the semantic types of the arguments (Hanks 2008: 115). The OLIF property 'semanticType' in contrast refers to the head of the entry, for our examples the verb. Syntactic properties are represented in the JAKOB lexicon by the subcategorization frames (Satzmuster) of Wahrig (2007) and are already included in the lexicon structure. The Wahrig frame for 'Angst haben' is no. 550 (verb + direct object + prepositional object).

Based on the formerly discussed findings, we are planning to implement an additional lexicon property 'bauplan' which is formally constructed as a combination of the pattern identifier and the DuELME component list. It is assembled from the Wahrig subcategorization frames and semantic information out of the CPA-pattern. Because this structure is difficult to read for the lexicographer, it is generated automatically and can be hidden from the user, but is available for NLP tasks.

Example 3: Property 'bauplan': 550:
haben[sb-animate] Angst[sg][uncountable] vor[anything]

5. Outlook: Support for lexicographic tasks

The use of MWE patterns has several advantages for the lexicographer. As Grégoire (2006) showed, it is possible to automatically assign lexical entries to a pattern class if we analyze the entry syntactically. Another important help for lexicographers is corpus investigation. We have already imported our corpus into the SketchEngine (Kilgarriff, Rychly, Smrz & Tugwell 2004). Given a CL of a lexical entry and a pattern assignment, it's feasible to automatically compute a corresponding corpus query. One further idea to assess the quality of the MWE descriptions is the use of a specialized generation grammar that produces example sentences for an entry and is guided by the parameters and the patterns: Missed phenomena concerning restrictions of modification, passivization, etc. will then show up.

[10] SB = subject, OA = accusative object, OP = prepositional object (TIGER annotation).

References

Boothe, B. (2004). *Der Patient als Erzähler in der Psychotherapie*. 2nd ed. Giessen: Psychosozial-Verlag.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Gregoire, N. (2006). 'Elaborating the parameterized Equivalence Class Method for Dutch'. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. 1894-1899.

Grégoire, N. (2007). *MWE Lexicon for Dutch Encoding protocol*. University of Utrecht.

Grégoire, N. (2009). 'DuELME: a Dutch electronic lexicon of multiword expressions'. In *Language Resources and Evaluation* (online first).

Hanks, P. (2004). 'The Syntagmatics of Metaphor and Idiom'. In *International Journal of Lexicography* 17 (3). 245-274.

Hanks, P. (2008). 'Lexical Patterns: From Hornby to Hunston and beyond'. In Bernal, E; DeCesaris, J. (eds.). *Proceedings of the XIII. Euralex International Congress, Barcelona*. 89–129.

Hanks, P.; Pustejovsky, J. (2005). 'A Pattern Dictionary for Natural Language Processing'. In *Revue Francaise de Langue Appliquée* 10 (2). 1-19.

Keil, M. (1997). *Wort für Wort: Repräsentation und Verarbeitung verbaler Phraseologismen (Phraseo-Lex)*. Tübingen: Niemeyer.

Luder, M.; Clematide, S.; Distl, B. (2008). 'Ein elektronisches Lexikon im OLIF-Format für die Erzählanalyse'. In Bernal, E.; DeCesaris, J. (eds.). *Proceedings of the XIII. Euralex International Congress, Barcelona*. 729–735.

Kilgarriff, A.; Rychly, P.; Smrz, P.; Tugwell, D. (2004). 'The Sketch Engine'. In *Proceedings EURALEX 2004, Lorient*.

Moszczyński, R. (2007). 'A practical classification of multiword expressions'. In *Proceedings of the ACL 2007 Student Research Workshop*. 19-24.

Pedrazzini, S. (1994). *Phrase manager: A system for phrasal and idiomatic dictionaries*. Hildesheim: Olms.

Pustejovsky, J.; Hanks, P.; Rumshisky, A. (2004). 'Automated induction of sense in context'. In *COLING 2004, Geneva*. 1-7.

Wahrig, G. (2007). *Der kleine Wahrig: Wörterbuch der deutschen Sprache*. München: Bertelsmann.