Extracting terms and their relations from German texts: NLP tools for the preparation of raw material for specialized e-dictionaries

Ina Rösiger[1], Johannes Schäfer[1], Tanja George[1], Simon Tannert[1], Ulrich Heid[2], Michael Dorna[3]

[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]University of Hildesheim, Germany
[3]Robert Bosch GmbH, Germany
`[roesigia|schaefjs|georgeta|tannersn]@ims.uni-stuttgart.de`,
`heidul@uni-hildesheim.de`, `michael.dorna@de.bosch.com`

Abstract

We report on ongoing experiments in data extraction from German texts in the domain of do-it-yourself (DIY) instructions, where the objective is (i) to extract nominal term candidates with high quality; (ii) to extract predicate-argument structures involving the term candidates, and (iii) to relate German word formation products with syntactic paraphrases: we focus on the analysis of compounds and on relating them with their syntactic paraphrases, in order to provide evidence for the (semantic) relationship between compound heads and non-heads (*Holzbohrer* (wood drill) $\leftrightarrow$ $Holz_{Object}$ *bohren* ([to] drill wood)). The extracted material is collected in order to provide structured data input for the creation of specialized dictionaries that are richer than standard terminological glossaries. For the creation of taxonomic knowledge (*Bandsäge -is-a* $\rightarrow$ *Säge* (bandsaw $\rightarrow$ saw)), we analyze subtypes of compounds.

Keywords: terminology extraction; raw material for specialized dictionary creation; lexical resources; German language; parsing

1. Introduction

There is a growing need for tools to extract terminology and relational data from text of specialized domains. Relational data involve verbal or adjectival predicates, their subjects, objects, complements, or preferred adjuncts; together with (mostly nominal) term candidates, they serve as a basis for ontology building and for the creation of raw material for dictionaries of the language of specialized domains.

The objective of the work described in this paper is the collection of German terminological data from heterogeneous corpora from the domain of do-it-yourself instructions. We use standard corpus linguistic technology for terminology extraction, as well as additional procedures for collecting and grouping related data with a view to the creation of a specialized lexical resource. The procedures are based on automatic word formation analysis and on dependency parsing. While the use of parsing for term extraction is not new, dependency parsing for German of an appropriate quality has only been available for five years (Bohnet, 2010).

The remainder of this paper is structured as follows: Section 2 describes the specialized and general-language corpora used as a text basis for the extraction of term candidates. Section 3 presents the NLP tools and methods involved, and Section 4 gives an overview of the approaches designed to link the extracted term candidates, in order to collect raw material for a dictionary of specialized vocabulary.

2. Corpus data

Since our term extraction procedures rely, among other factors, on the comparison of specialized and "general language" texts, we work with corpora of both kinds.

As a domain-specific corpus, we use a corpus containing both expert and user-generated German texts from the DIY domain, which is composed, among other things, of manuals, practical tips, marketing texts and DIY project descriptions. The basic version of the corpus contains ca. 2.7 M tokens; in the course of this work, the corpus has been extended to 17.9 M tokens (see Tables 1 and 2 for details). The current versions of the corpus are not yet publicly available.

Text type:	**# tokens:**	**authors:**
DIY manual	62,131	experts
DIY encyclopedia	6,868	experts
DIY practical "tricks"	15,104	experts
Marketing texts	35,302	experts
DIY project descriptions	2,160,008	UGC
FAQs (forum)	5,150	UGC
Wiki content	444,381	UGC
Total	2,728,944	

Table 1: DIY corpus

Text type:	**# tokens:**	**authors:**
DIY manual	62,131	experts
DIY encyclopedia	6,868	experts
DIY practical "tricks"	15,104	experts
Marketing texts	35,302	experts
DIY project descriptions	4,479,437	UGC
FAQs (forum)	128,906	UGC
Wiki content	896,267	UGC
DIY articles	2,807,487	experts
Test descriptions	239,238	experts
DIY web encyclopedia	21,562	experts
Forum articles	296,242	UGC
DIY forum posts	7,873,115	UGC
Builders' diaries	22,715	UGC
Video descriptions	2,280	UGC
Tool manuals	69,123	experts
Keyword lists	15,940	experts
Varia (no metadata)	961,236	-
Total	17,932,953	

Table 2: Extended DIY corpus

Our corpora are heterogeneous, as far as authorship and intended readership, text types and the level of specificity of the texts are concerned: while the manuals and the "tips and tricks" documents are written by experts (mostly for semi-experts or lay persons), a large portion of the texts comes from user-generated content (UGC) available in forums and thus likely authored by semi-experts and/or lay persons. The corpus is intended to be a sample of the domain-related material available on the internet with a ratio of roughly 1:4 of expert vs. user generated content. In future work, we intend to separately analyze forum data and texts authored by experts, to assess specificities of each subcorpus.

As for the general-language corpus, we rely on the SdeWaC corpus (cf. (Faaß and Eckart, 2013)), a web corpus covering a wide range of topics and text styles, that contains around 880 M words. SdeWaC is a subset of deWaC (Baroni and Kilgarriff, 2006); it only contains sentences that can be parsed by the rule-based dependency parser FSPar (Schiehlen, 2003).

3. Computational linguistic technology used

The procedures used in our experiments are based on existing generic tools:

- A hybrid term extractor based on the prototype designed in the EU project TTC (*Terminology Extraction, Translation Tools and Comparable Corpora*, FP-7, STREP 248005, (Gojun et al., 2012a), (Gojun et al., 2012b) cf. Section 3.1);
- the dependency parser included in the *mate* tools (Bohnet, 2010), (Björkelund et al., 2010), as well as a tool that annotates syntactic phrases (and their boundaries, implicitly), cf. Section 3.2 and 3.3;
- the compound splitting tool CompoST (Cap, 2014), cf. Section 3.4.

We intend to combine the output of the tools in such a way as to be able to accumulate, from the corpus, the raw material for lexical entries that cater for term variation, partial taxonomies and the description of other, non-taxonomic relationships between concepts denoted by terms of the domain.

In the following, we briefly describe the three types of computational linguistic tools mentioned above.

3.1 Term extraction tools

The term extractor used in our work is a prototype based on a tool for German developed in the TTC project (Gojun et al., 2012b). It is a hybrid tool combining linguistic corpus preprocessing with statistical domain specificity ranking. Figure 1 schematizes the main steps of the tool pipeline.

The pipeline involves the following components:

- Preprocessing:

Figure 1: Steps in term candidate extraction: overview

– Tokenization: sentence and word form delimitation and markup;
– word class tagging and preliminary lemmatization: annotation by means of the RF-Tagger (Schmid and Laws, 2008), including an annotation as "unknown" of word forms absent from the tagger lexicon;
– lemmatization: specific treatment of the word forms absent from the tagger lexicon, with a view to guessing their lemma, by use of word form similarity, inflection-based rules and compound splitting; this component provides lemma forms for most of the "unknowns" which remained after the first lemmatization step.

The preprocessing steps of POS-tagging and lemmatization involve a simple form of domain adaptation: as the tagger used in the first run marks which word forms are not contained in its dictionary ("unknowns", with respect to the data acquired in standard training from newspaper texts), these can be handled in the above mentioned specific lemmatization step which uses morphological knowledge and similarity data to guess lemma values. In future work, this set of procedures will be combined with Named Entity Recognition tools to make it more robust to new domains.
The preprocessing annotations are stored in a one word per line format.

- Pattern-based term candidate extraction:
  use of simple as well as extended POS-based patterns to identify term candidates; typical basic patterns are simple nouns, adjective+noun groups and nouns followed by genitive or prepositional modifiers. For verbal term extraction, patterns based on dependency parses are used, cf. Section 3.2.
- Ranking:
  sorting of the candidate lists produced by the preceding step, according to different measures: a basic approach uses (Ahmad et al., 1992)'s "weirdness ratio" (quotient of relative domain corpus frequency by relative general-language corpus frequency), while more advanced versions involve further measures, such as the C-Value measure ((Frantzi and Ananiadou, 1999); cf. (Schäfer, 2015) for details).

The output of the above steps are term candidate lists by patterns; examples of each pattern are given below:

N	*Bohrmaschine* (drill)
Adj+N	*oszillierende Säge* (oscillating saw)
N+Det+$N_{genitive}$	*Kopf einer Schraube* (head of a screw)
N+Prep+N	*Handkreissäge mit Führungsschiene* (skill saw with guide rail)

In addition to the basic patterns, and in line with Daille's notion of term variants (Daille, 2007), more complex patterns are processed in the same way. The set of extended patterns is described by the regular expressions given below:

– ((Adv)? (Adj)? Adj)? N
– (N Det)? ((Adv)? Adj)? N Prep (Det)? ((Adv)? Adj)? N
– ((Adv)? Adj)? N Det ((Adv)? Adj)? $N_{genitive}$

3.2 Extracting verb object pairs from dependency parsed text

Standard term candidate extraction typically focuses on nouns and nominal phrases as they cover the objects of the domain (see patterns above). For the extraction of relational knowledge and to put the domain objects into context, verbally expressed relations are needed as well. We thus want to apply a variant of the above mentioned term extraction pipeline, i.e. the selection of candidates via linguistic preprocessing combined with a statistical ranking, also to verbal term candidates. The problem that arises is that the POS-based tool has no information about syntactic phrases and their boundaries, such that a part-of-speech-based approach is not sufficient, particularly for a language like German that has three models of verb placement and allows flexible word order.

For the verbal candidate extraction, pre-processing thus includes a separate dependency parsing step, followed by a script that extracts verb object (or subject verb) pairs which are then processed by the statistical filtering step. This treatment leads to local information which can be considered as a combination of dependency syntactic and constituent structural knowledge; it is thus richer than mere dependency annotations as provided, for example by Constraint Grammar.

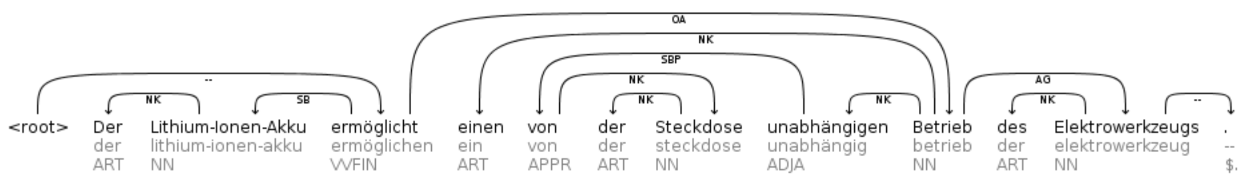
To find suitable verb candidates and their corresponding subjects and objects, we use the dependency parser contained in the *mate* tool package (Bohnet, 2010), (Björkelund et al., 2010) to annotate the texts with dependency syntactic analyses; the parser is trained on a dependency version of the TiGer treebank (Brants et al., 2004), (Seeker and Kuhn, 2012) which contains newspaper texts; there is no domain-specific treebank available. However, the tool profits from the domain adaptation of the pre-processing steps, i.e. lemmatization and POS-tagging. We are currently investigating ways to adapt the dependency parser to the domain without the rather expensive creation of manual gold data.

As we are interested in verb+object (or subject+verb) pairs irrespective of whether the pair occurred in the active or passive voice, we apply an approach that annotates passive sentences with grammatical functions that correspond to the active voice version so that all corpus sentences can be handled in the same way in the pattern-based term extraction step.

For example, *Holz wird gesägt* (wood is sawn) is mapped to the verb object pair *Holz sägen* (saw wood). Active and passive is not explicitly annotated in the dependency parses, but it can be determined by a set of syntactic rules.

The head of an object (*OA* in dependency graph in Figure 2) or of a subject phrase (*SB* in graph) is marked so that one can specify whether the whole phrase should be extracted or just the syntactic head (which helps avoid data sparsity issues).

Figure 2 shows a graphical representation of the dependency parser output and the mapped annotations that are used as the basis for the extraction of candidates. The mappings are stored in a separate column in our one word per line format, distinguish subject (*SUBJ*) from object (*OBJ*) phrases and mark the syntactic head with the ending *-Head*. All other parts of the respective phrase end with *-Embedded*. Verbs are marked, as well as the information whether they occurred in a passive or active sentence (*VERB-Active*).



0	Der	SUBJ-Embedded	The
1	Lithium-Ionen-Akku	SUBJ-Head	lithium ion accumulator
2	ermöglicht	VERB-Active	enables
3	einen	OBJ-Embedded	a
4	von	OBJ-Embedded	from
5	der	OBJ-Embedded	the
6	Steckdose	OBJ-Embedded	socket
7	unabhängigen	OBJ-Embedded	independent
8	Betrieb	OBJ-Head	operation
9	des	OBJ-Embedded	of the
10	Elektrowerkzeugs	OBJ-Embedded	power tool
11	.	NULL	.

Figure 2: Dependency graph and mapped representation for *The lithium ion accumulator enables a socket-independent operation of the power tool.*

To be able to handle queries about verb phrases and their arguments, the term extractor had to be slightly adjusted. Apart from the standard sequence-based patterns it can now handle structure-based patterns and the respective queries. After the extraction of potential term candidates, we apply the same statistical measures that were used in the nominal term extraction.

3.3 Annotation of syntactic boundaries

The dependency parser can also be used to improve nominal term extraction by making sure that noun phrase candidates are syntactically valid. Term candidates covering excessively long spans typically occur in NPs followed by a PP, when part of the extracted candidate is actually attached to the verbal phrase, e.g. in (1) and (2). The invalid term candidates are underlined and marked with an asterisk. In these cases a phrase boundary ([NP][PP]) is found within the extracted string, and the (terminological) NP and the subsequent PPs are sisters. Valid term candidates would consist of a complex NP where the PP is embedded. We filter the output of the POS-pattern based extraction by using *mate* to find start and end points of NPs.[1]

(1) die *Vorlage mit Sprühkleber besprühen (spray the *template with paint)

(2) ein *Loch in die Wand bohren (drill a *hole into the wall)

The boundary violation filter works as follows: if one or more words of the selected term candidate go beyond the phrase boundary, the candidate is not counted as a valid occurrence of this particular lemma sequence. The candidate sequence is not removed from the list of possible candidate terms, as other occurrences might not violate syntactic boundaries. The filter is thus a "soft" one as it only affects the frequency of the lexeme combination candidate. We also experiment with a "hard" filter, where the lexeme combination candidate is removed altogether as soon as an invalid candidate occurrence is found.

3.4 Compound splitting

For compound splitting we use CompoST (Compound Splitting Tool, (Cap, 2014)), a compound splitter which combines the use of a rule-based morphology system (SMOR, (Schmid et al., 2004)) with subword (i.e. morpheme) verification in corpus data, thereby extending and improving on the approach proposed by (Koehn and Knight, 2003) for statistical machine translation: for all components of a compound, including those which are complex themselves, the tool verifies the presence and number of occurrences in a (set of) texts; in our application, the do-it-yourself corpus is used as a knowledge source for this check, in addition to a (newspaper-based) general language corpus. Splits that involve implausible or rare components are dispreferred.

[1] In current experiments only for NPs in subject or object position; work towards covering all relevant construction types is ongoing. We are aware that *mate* has not been optimized to solve the PP attachment problem.

For specialized terms, taking a domain corpus as the basis for the computation of probable splits often has the effect that wrong splits based on general-language frequencies (*Betonverbinder (concrete connector)* split into *Beton(concrete)|verb(verb)|inder(indian)*) are avoided and the right splits are produced (*Beton(concrete)|verbinder(connector))*. The tool allows a set of parameters, such as to show all possible splits or just the most probable one, and to decide whether the output should contain surface forms or lemmatized forms, to name only a few.

3.5 Quality of the term candidate extraction

The performance of the basic pipeline (cf. Section 3.1) has been evaluated on a gold standard data collection created from the 2.7 M words corpus described above in Section 2.

The gold standard (GS) was annotated manually by three independent experts; only term candidates with a minimum frequency of four and pertaining to one of the basic patterns (Section 3.1) were annotated, following predefined guidelines (cf. (George, 2014)). The candidates based on the extended patterns and the verbal candidates have not yet been evaluated against a gold standard.

We obtained a strict and a liberal version of the gold standard, where the strict GS only contains items for which full agreement on their term status was found. The total GS contains 4,238 single-word terms and 859 multi-word terms. The strict GS contains 2,777 terms, while the liberal GS includes additional 2,320 term candidates. The inter-annotator agreement ranges between moderate and substantial agreement (Landis and Koch, 1977), cf. Table 3.

annotators:	κ of N+*"von"*+N:	κ of N+Det+N_{gen}:	κ of N:	κ of Adj+N:	κ of N+Prep+N:
A1&A2	0.69	0.47	0.50	0.55	0.63
A2&A3	0.65	0.60	0.54	0.54	0.65
A3&A1	0.71	0.48	0.48	0.52	0.60
A1, A2&A3	0.68	0.52	0.51	0.54	0.63

Table 3: Inter-annotator agreement for the gold standard data. Interpretation of the kappa values: 0.41 – 0.6 = moderate agreement; 0.61 – 0.8 = substantial agreement.

We automatically evaluated the output of our pipeline computing precision, recall and f-measure for each of the basic patterns. Table 4 contains the results obtained on the liberal gold standard.

We furthermore compared the term candidates extracted from our corpus with a commercial tool (SDL MultiTerm Extract, version May 2014[2]) which is based exclusively on statistical

[2] http://www.sdl.com/de/cxc/language/terminology-management/multiterm/extract.html

	N+"von"+N	N+Det+N_{gen}	N	Adj+N	N+Prep+N
Precision	72%	65%	52%	38%	55%
Recall	84%	91%	85%	55%	73%
F-measure	78%	76%	65%	45%	63%

Table 4: Precision, recall and f-measure values for the basic patterns compared with the liberal gold standard

procedures; while that tool is applicable to many languages without any need for language-specific knowledge, it is clearly outperformed on the German data by our prototype (George, 2014).

So far, no extensive GS-based evaluation of the effect of the phrase boundary check has been performed. However, tendencies can be observed: for the 107 terms of the GS which show the POS pattern "Noun+Preposition+Noun", an improvement in precision is found both with the "soft" and with the "hard" filter. For the term candidates extracted on the basis of the extended patterns, we also checked the top-500 candidates that contained a preposition, and we determined whether the removal from the candidate list which was suggested by the filter was justified: it achieved, on that sample, 83% precision. This means in four out of five cases the removed candidate was indeed violating syntactic boundaries.

4. Collecting raw material for a dictionary of specialized vocabulary

In this section we show how the corpus data and the above mentioned processing tools can be used to relate the term candidates extracted, with a view to the provision of a maximal amount of structured raw data for subsequent (manual) lexicographic work.

We do not aim to automate the creation of a specialized dictionary, but we intend to provide rich input for the lexicographic process. The focus in this paper is on term variants (in the sense of (Daille, 2007)) and on partial taxonomies. We explain different procedures used for this purpose, and we give examples of the output of each one. As we report on ongoing work, no quantitative evaluation of these procedures is yet available.

4.1 Analyzing variation in multi-word terms

As discussed in Section 3.1, we use basic POS patterns for the extraction of multi-word term candidates as well as extended ones which we relate in a meaningful way to the basic patterns, as suggested by (Daille, 2012). We consider a term candidate with an extended pattern to be a variant of a term candidate with a basic pattern if it contains the tokens of the basic one (in the same order). The term candidates with basic patterns are in turn retrieved by seeding the extractor with the nouns from our gold standard.

The relationships observed in the data can be subdivided into the following three types:

(1) Variation:
  - Example:
    Verkleidung aus Rigipsplatten (cladding made of plasterboard) $\leftrightarrow$
    Gipskartonplatten als Verkleidung (plasterboard as cladding)
(2) Subtype relations:
  - Example: Adj N $\rightarrow$ Adv Adj N:
    weiße Farbe (white paint) $\leftrightarrow$
    matt weiße Farbe, normal weiße Wandfarbe, weißlich durchsichtige Farbe
    (flat white paint, normal white wall paint, whitish sheer paint)
  - Example: N $\rightarrow$ Adj N:
    Schraube (screw) $\rightarrow$
    spezielle Schraube, passende Schraube, kleine Schraube, lange Schraube
    (particular screw, appropriate screw, small screw, long screw)
(3) Relations of non-taxonomic type, e.g. focusing on aspects of an item:
  - Examples:
    * $\text{Adj}_1\ \text{N}_1 \rightarrow \text{N}_2\ ((\text{Det}_1)\ \text{Adj}_1\ \text{N}_1)_{genitive}$:
      bodengleiche Dusche (walk-in shower) $\rightarrow$ *Aufbau einer bodengleichen Dusche* (construction of a walk-in-shower)
    * $\text{Adj}_1\ \text{N}_1 \rightarrow \text{N}_2\ \text{Prep}\ ((\text{Det}_1)\ \text{Adj}_1\ \text{N}_1)$:
      bodengleiche Dusche (walk-in shower) $\rightarrow$ *Anschluss an die bodengleiche Dusche* (connection to the walk-in-shower)

4.2 Analyzing compounds for the creation of taxonomic knowledge

Many specialized compounds are transparent, compositional determinative compounds and thus their head denotes their hypernym: *Kreissäge* (buzzsaw) “is-a” *Säge* (saw). On this (simplistic) assumption, compound splitting and the identification of heads allow for a grouping of items according to subtype relations. For example, starting from a simplex term (e.g. *Säge*, saw), all compounds could be identified that have this term as a head (e.g. *Bandsäge* (bandsaw), *Kreissäge* (buzzsaw), etc.), and a subtype relation could be assigned. This strategy could be applied recursively to create a partial hierarchy from more general to more specific terms (such as, e.g. *Säge* $\rightarrow$ *Bandsäge* $\rightarrow$ *Horizontalbandsäge* (horizontal bandsaw)).

The implementation differs from this principle, in order to correctly cover multimorphemic non-head elements: it takes a compound, splits it into morphemes, removes the first one and tries to find occurrences of the remaining part in the corpus. If, for example, it starts from *Eigenbaubandsäge* (self-made bandsaw) (split as *Eigen·bau·band·säge*), it will check the corpus for $^{??}$*Baubandsäge*, and it will not find any occurrence. It then skips the element *-bau-*

and checks for *Bandsäge*, where a sufficient number of occurrences are found. As we work on compounds from the domain, not finding an item in the corpus will most often mean that this item does not exist (as the hypothetic form ??*Baubandsäge*); obviously, a few cases may also be due to data sparsity. The full set of subtypes of *Bandsäge* (bandsaw), as found in our data, is summarized in Table 5. An exemplary hierarchy for the term *Säge* (saw) is given in Figure 3.

Eigenbaubandsäge (self-made bandsaw)	Eigen|Bau|Band|Säge
Elektro-Bandsäge (electric bandsaw)	Elektro|Band|Säge
Hand-Bandsäge (hand bandsaw)	Hand|Band|Säge
Horizontalbandsäge (horizontal bandsaw)	Horizontal|Band|Säge
Vertikalbandsäge (vertical bandsaw)	Vertikal|Band|Säge
Metallbandsäge (metal bandsaw)	Metall|Band|Säge
Minibandsäge (mini bandsaw)	Mini|Band|Säge
Bandsäge (bandsaw)	Band|Säge

Table 5: Subtypes of *Bandsäge* (bandsaw) in the corpus

For the term *Säge* (saw) we gathered and manually verified the partial ontology constructed from the compounds analyzed in this way. Of 213 compound candidates, 36 candidates are not found in the corpus, because the compounds do not exist in German or because the forms used as an input to the procedures contain typographic errors.

4.3 Analyzing syntactic paraphrases of compounds

We use the parsed version of the corpora to identify potential syntactic paraphrases of German noun compounds; examples include nouns with genitive attributes (*Holzmaserung – Maserung des Holzes* (grain of wood)) and nominals with PPs (*Wasserkontakt, Kontakt mit Wasser* (contact with water)) as well as verb+object collocations (*Temperaturerhöhung – Temperatur+erhöhen* (increase (in) temperature)).

4.3.1. Compounds with nominal heads

We acquire paraphrases for compounds with nominal heads by querying noun+preposition+ +noun or noun+determiner+noun (in genitive case) patterns in the 17.9 M corpus. Searching for syntactic paraphrases (synt) of nominal compounds (cmpd) serves two different purposes of lexicographic relevance:

(i) quantitative aspects: to find more instances of an item, by grouping term variants together:

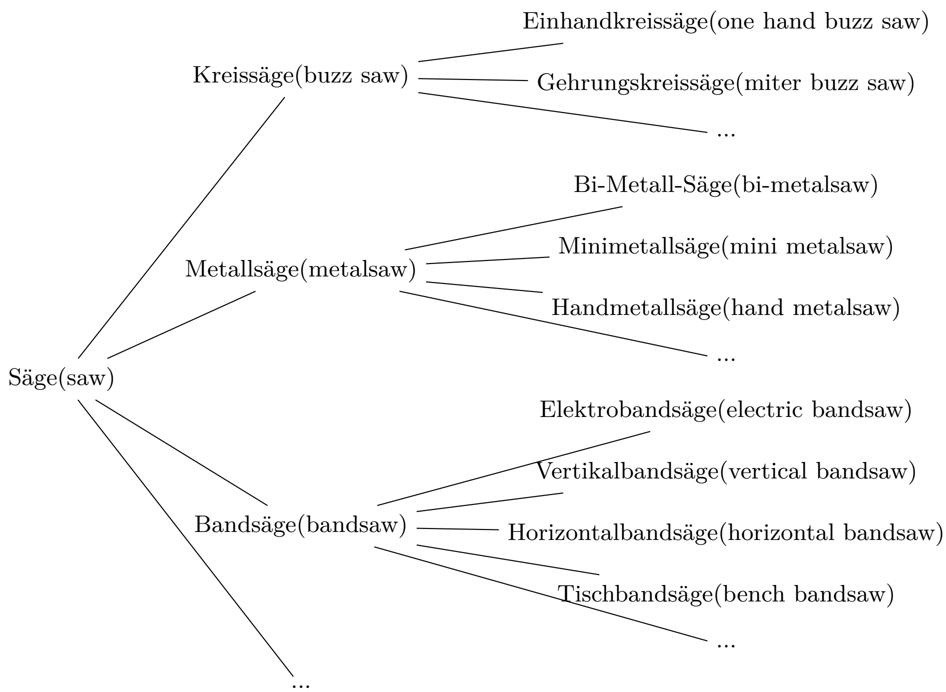


Figure 3: Sample of a partial hierarchy of the term candidate *Säge* (saw)

			f_{cmpd}	f_{synt}	$\sum$
– *Schraubenloch* (screw+hole)	↔	*Loch für Schraube* (hole for screw)	441	15	456
– *Raummitte* (room+centre)	↔	*Mitte des Raumes* (centre of the room)	37	57	94
– *Holzmaserung* (wood+grain)	↔	*Maserung des Holzes* (grain of the wood)	136	56	192
– *Brettkante* (board+edge)	↔	*Kante des Brettes* (edge of the board)	79	41	120

(ii) to derive the semantic relation existing between the compound head and the non-head:

	f_{cmpd}	f_{synt}	$\sum$
– *location: Fliesenfuge* (slab+joint) ↔ *Fuge zwischen Fliesen* (joint between slabs)	110	17	127
– *material: Teakmöbel, Teakholzmöbel* (teak(wood)+furniture) ↔ *Möbel aus Teak* (furniture made of teak)	7(+8)	21	28
– *material: Beton-Fundament, Betonfundament* (concrete+basement) ↔ *Fundament aus Beton* (basement made of concrete)	127(+22)	21	148

With respect to the first objective, a simple case is the collection of all possible "genitive" forms: next to the rare item *Loch bohren* (drill a hole) (f = 7), we find *Bohren des Lochs* (drilling of the hole) (103), *Bohren eines Lochs* (drilling of a hole) (6), *Bohren von Löchern* (drilling of holes) (8). These procedures allow us to collect all morphosyntactic variants of a collocation, i.e. verb+object (*Temperatur erhöhen* (increase temperature)), nominalisation of the verb+genitive (*Erhöhung der Temperatur*), compound (*Temperaturerhöhung*) and, if the lexicographer regards this as a separate type, attributive participle (*erhöhte Temperatur*). We are aware that these "variants" are not necessarily fully synonymous. Specialized languages in addition tend to be highly selective with respect to the choice among these variants as shown by (Fritzinger and Heid, 2009) for a subdomain of juridical language.

A more difficult task is that of relating compounds with appropriate noun+PP paraphrases.

While some compounds only have one paraphrase, or only one statistically prominent paraphrase, others have several potential paraphrases, especially those which are truly polysemous. An example of this last case is *Holzfarbe* (wood+colour): it is polysemous and denotes (a) the colour of wood or (b) (synthetic) colours designed to paint wood. Both readings show up in our corpus, but the first reading is most prominent in the syntactic paraphrase data. For a disambiguation of the compound occurrences (e.g. to provide example sentences for the lexicographer), we intend to rely on indicator items from the context, e.g. (semantic) types of adjectives preceding *Holzfarbe* (*graue (gray), weiße (white), ...* → colour to paint wood; *originale (original), natürliche (natural), ...* → colour of wood).

The taxonomy of compounds with a specific head noun (as in Figure 3) can now be enriched with the semantic relations acquired from the noun+PP paraphrases, which makes it possible to group the subtype items. Table 6 presents an excerpt from a detailed analysis of compounds of the noun *Schraube* (screw) and their paraphrases where the compounds are grouped by the semantic relation between the compound head and the non-head.

material:		preposition: *aus* (made of)	
	Stahlschraube	↔ *Schraube aus Stahl*	(steel screw)
	Edelstahlschraube	↔ *Schraube aus Edelstahl*	(stainless steel screw)
	Kupferschraube	↔ *Schraube aus Kupfer*	(copper screw)
application:		preposition: *für* (for)	
	Rigips-Schraube	↔ *Schraube für Rigips*	(screw for plasterboard)
type:		preposition: *mit* (with)	
	Senkkopf-Schraube	↔ *Schraube mit Senkkopf*	(countersunk head screw)
purpose:		preposition: *als/zu* (as/to)	
	Führungsschraube	↔ *Schraube als Führung*	(screw as a guide)
	Befestigungsschraube	↔ *Schraube zu Befestigung*	(screw as a fixing)

Table 6: Compounds with the head *Schraube* (screw) and their paraphrases

Finally, there are cases where the compound is not paraphrased adequately in the corpus; equally, more work needs to be done to remove spurious paraphrase candidates:

- *Treppenraum (stairwell)* ↮ *Raum unter der Treppe (room under stairs),*
  ↮ *Raum zwischen Treppe und Wand (room between stairs and wall)*

Overall, the simple procedures sketched above produce relatively good results; a precision evaluation of a sample is planned.

Compound	**Object +**	**Verb**
Temperaturerhöhung (temperature rise)	Temperatur (temperature)	to rise (erhöhen)
Temperaturmessung (temperature measurement)	Temperatur	messen (to measure)
Temperaturregelung (temperature control)	Temperatur	regeln (to control)
Temperaturüberwachung (temperature monitoring)	Temperatur	überwachen (to monitor)
Dübellochbohrer (dowel hole drill)	Dübelloch (dowel hole)	bohren (to drill)
Fliesenbohrer (tile drill)	Fliesen (tile)	bohren
Holzbohrer (wood drill)	Holz (wood)	bohren
Kreisbohrer (circle cutter)	Kreis (circle)	bohren
Kunststoffbohrer (plastic drill)	Kunststoff (plastic)	bohren
Langlochbohrer (deep-hole drill)	Langloch (deep hole)	bohren
Maschinenbohrer (machine drill)	??Maschinen (machine)	bohren
Nagelbohrer (nail drill)	??Nagel (nail)	bohren
Pfostenbohrer (jamb drill)	??Pfosten (jamb)	bohren
Diamantbohrer (diamond drill)	NOT: *Diamant (diamond)	bohren

Table 7: Deverbal compounds and their syntactic paraphrases for *Temperatur (temperature)* and *Bohrer (drill)*

4.3.2. Compounds with verbal heads

For deverbal compounds, we aim to distinguish different relations between the head and the non-head by analyzing the presence (or absence) of certain syntactic paraphrases, e.g. verb object pairs. The following section describes our experiments on linking deverbal compounds and their corresponding verb object pairs. In the future, we also plan to investigate subject verb pairs or other constructions that put the involved term candidates into context, such as predicative expressions.

For deverbal heads and their respective non-heads, there is a variety of possible relations between the two. If we take *Bohrer* (drill), for example, we can find a number of different semantic relations: *Diamantbohrer* (diamond drill) exemplifies an `is-made-of` relation where the non-head describes the material of which the drill is made, whereas a *Holzbohrer* (wood drill) is used to drill wood. Here, the non-head specifies the object to be drilled.

Thus, in our ongoing work, we first extract all deverbal compounds and the corresponding verb (a total of 8,750 compound types with verbal head and nominal non-head are present in our corpus) and then look for the respective verb object pairs in the dependency parses where the object equals the non-head of the compound. We then sort the extracted paraphrases by the nominal non-head (as in the first example in Table 7) and find events involving the noun, or we can sort by the deverbal head (as in the second example in Table 7) and find typical objects of the verb.

Table 7 shows the compounds and their matching paraphrases for two examples, *Temperature* (temperature) as a non-head and *Bohrer* (drill) as a head. When we find a verb object pair for a certain compound, e.g. *Kunststoffbohrer* (plastic drill), we now know that it is used to *drill plastic*. For *Diamantbohrer* (diamond drill) we do not find such a paraphrase. This confirms our claim that the relation between the head and the non-head in this case is a different one, i.e. a `is-made-of` relation. In some cases, Noun+PP-evidence confirms this classification, cf. *Hartmetallbohrer* (tungsten carbide drill) ↔ *Bohrer aus Hartmetall* (drill made of carbide).

While a quantitative analysis of this automatic linking approach has not yet been performed, we have found a total of 7,411 occurrences of verb object pairs for our 8,750 compound types (1,381 unique verb object pairs). The reported links have been created on the basis of the 2.7 M corpus. We are currently performing experiments on the 17.9 M corpus, which will increase the coverage of matching paraphrases for the candidate terms extracted by the term extractor. We think that the number of links found is large enough to be beneficial for the creation of a specialized dictionary.

4.4 Lexicographic use of the collected data

The procedures discussed in section 4 of this paper are all meant to support human lexicographers in the preparation of entries of an online dictionary. The targeted dictionary is meant to be both a resource for human use and a knowledge source of automatic or semi-automatic tools, e.g. for e-mail routing, knowledge extraction from texts, as well as passage retrieval.

A possible interactive version of the dictionary would be characterized, among other factors, by the following properties: (i) it is a monolingual specialized dictionary allowing both semasiological and onomasiological access (the latter through the (partial) taxonomies constructed according to the procedures described in section 4.2); (ii) it goes beyond the structure and descriptive programme of terminological databases, insofar as it has not only nouns, but also verbs as lemmata and because it relates action-denoting verb+object pairs with terms; (iii) we foresee the possibility to add other languages to the dictionary.

The raw material gathered by means of the devices discussed in section 4 will serve the lexicographers as an input: it is not intended to create the lexicographic product fully automatically. The objective is to combine all evidence gathered for a given nominal or verbal element and to present this synthetically to the lexicographer. Furthermore, we intend to experiment with possibilities to propose collocation candidates on the assumptions (i) that most compounds in the domain are compositional and transparent and (ii) that in such cases compounds "inherit" collocational preferences from the heads of their bases: thus, as we have *Schraubenloch* (screw+hole) and *Loch für Schraube* (hole for screw) (section 4.3.1), as well as *Loch bohren* (drill a hole) and *Bohren des Lochs* (drilling of a hole), we provide *Schraubenloch bohren* and *Bohren des Schraubenlochs* as candidates, even though these are not covered by our current corpora, but may well be found in other corpora of the domain.

As of the summer of 2015, we are in the process of enhancing the tools; while experimental lexicographic work is going on to assess the usefulness of the tools, no large-scale lexicographic activity has yet been carried out.

5. Conclusion and future work

In this paper we presented tools and procedures for the extraction of term candidates from German specialized language texts, and for grouping the extracted data in a meaningful way, in order to provide raw material for the interactive construction of specialized dictionaries.

Since we intend these dictionaries to be used especially for semi-automatic document classification in the context of electronic communication between experts and lay persons or semi-experts, as well as for text production, we based our extraction procedures on both expert and user-generated text.

We consider that term variants, taxonomic relations, as well as other relations, such as purpose or material are crucial. To provide hints at such semantic relations, we use different morphological, morphosyntactic and syntactic extraction tools and relate their results. The setup is similar to that of the *Sketch Engine* (Kilgarriff et al., 2004), in so far as we extract syntagmatic data by means of pattern-based search, we are able to combine the results to make relations between the elements of German compounds explicit. We can go beyond the functions of *Sketch Engine* by exploiting nominal compounds and their syntactic paraphrases, and by interpreting e.g. noun+PP co-occurrences semantically.

The use of existing semantic lexicons, such as WordNet (Fellbaum, 1998)[3], to seed the semantic classification, as well as the use of domain-specific hierarchies (e.g. provided by relevant manufacturers) is being investigated; a first inspection of WordNet data for the types of drills discussed in Table 7 showed mixed results: at an abstract level, "diamond" and "wood" are both materials, and disambiguation on WordNet data alone seems less powerful than the paraphrase-based approach discussed.

Future work will include broader coverage experimentation on the 17.9 M words corpus, the use of domain-specific taxonomic data from manufacturers, more paraphrase-based interpretation rules and quantitative evaluations of subsets of the data produced. Furthermore, the extraction procedures themselves will be fine-tuned, and experiments into low-cost domain-adaptation will be made.

6. Acknowledgements

The work reported in this paper has been carried out in the framework of the collaborative research project "Terminologieextraktion und Ontologieaufbau" financed by the corporate research department of Robert Bosch GmbH. We gratefully acknowledge this support.

[3] `http://wordnetweb.princeton.edu/perl/webwn`

7. References

Ahmad, K., Davies, A., Fulford, H. & Rogers, M. (1992). What is a term?—the semi-automatic extraction of terms from text. In *Translation Studies – An Interdiscipline*, pp. 267–278.

Baroni, M. and Kilgarriff, A. (2006). Large linguistically-processed web corpora for multiple languages. In *Processing of EACL Conference 2006*.

Björkelund, A., Bohnet, B., Love, H. & Pierre, N. (2010). A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, Beijing, China. Coling 2010 Organizing Committee, pp. 33–36.

Bohnet, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China*, Association for Computational Linguistics, pp. 89–97.

Brants, S., Dipper, S., Eisenberg, P., König, E., Lezius, W., Rohrer, C., Smith, G. & Uszkoreit, H. (2004). Tiger: Linguistic interpretation of a german corpus. *Journal of Language and Computation*, 2, pp. 597–620.

Cap, F. (2014). Morphological processing of compounds for statistical machine translation. Dissertation, Institute for Natural Language Processing (IMS), University of Stuttgart.

Daille, B. (2007). Variations and application-oriented terminology engineering, pp. 163–177.

Daille, B. (2012). Building bilingual terminologies from comparable corpora: The ttc termsuite. In *Proceedings, 5th Workshop on Building and Using Comparable Corpora with special topic "Language Resources for Machine Translation in Less-Resourced Languages and Domains", co-located with LREC 2012*, Istanbul, Turkey.

Faaß, G. and Eckart, K. (2013). Sdewac – a corpus of parsable sentences from the web. In *Language Processing and Knowledge in the Web: 25th International Conference, GSCL 2013, Darmstadt, Germany, September 25–27, Proceedings*, pp. 61–68.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database.* Bradford Books.

Frantzi, K. and Ananiadou, S. (1999). Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal of Digital Libraries*, 6, pp. 145–179.

Fritzinger, F. and Heid, U. (2009). Automatic grouping of morphologically related collocations. In *Proceedings of the Corpus Linguistics 2009 Conference*, Liverpool/UK.

George, T. (2014). Comparing a commercial term extraction tool with a research prototype: an evaluation study on DIY instruction texts. Bachelor thesis, Institute for Natural Language Processing (IMS), University of Stuttgart.

Gojun, A., Heid, U., Blancafort, H., Loginova, E., Guégan, M. & Gornostay, T. (2012a). Reference lists for the evaluation of term extraction tools. In *Proceedings of Terminology and Knowledge Engeneering Conference*, pp. 651–656.

Gojun, A., Heid, U., Weissbach, B., Loth, C. & Mingers, I. (2012b). Adapting and evaluating a generic term extraction tool. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 651–656.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The sketch engine. *Information Technology*, 105, pp. 116.

Koehn, P. and Knight, K. (2003). Feature-rich statistical translation of noun phrases. In *Proceedings of ACL 2003*.

Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, pp. 159–174.

Schäfer, J. (2015). Statistical and parsing-based approaches to the extraction of multi-word terms from texts: implementation and comparative evaluation. Bachelor thesis, Institute for Natural Language Processing (IMS), University of Stuttgart.

Schiehlen, M. (2003). A Cascaded Finite-State Parser for German. In *Proceedings of EACL 2003*, pp. 163–166.

Schmid, H., Fitschen, A. & Heid, U. (2004). Smor: A german computational morphology covering derivation, composition, and inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, pp. 1263–1266.

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, pp. 777–784.

Seeker, W. and Kuhn, J. (2012). Making ellipses explicit in dependency conversion for a german treebank. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey, pp. 3132–3139.