Freie Universität Berlin

DISTINGUISHING SPECIALISED DISCOURSE

The Example of Juridical Texts on Industrial Property Rights and Trade-mark Legislation

Fabienne Cap and Ulrich Heid
December 2010

Berliner Arbeitspapier zur Europäischen Integration Nr. 16

Berlin Working Paper on European Integration No. 16

This paper is part of a series of method papers on the use of advanced text analytical methods in the political and social sciences. The papers where presented at the workshop "How to Analyse Millions of Texts and Still Be Home for Tea? Innovations in Textual Analysis in the Social Sciences" conducted at Freie Universität Berlin in May 2010. The workshop was organised by Amelie Kutter and Cathleen Kantner and it was co-financed by the FP6 Integrated Project "RECON – Reconstituting Democracy in Europe" and the Jean Monnet Centre of Excellence "The EU and its Citizens" at Freie Universität Berlin.

Arbeitsstelle Europäische
Jean Monnet Lehrstuhl Integration

Content

Die Berliner Arbeitspapiere zur Europäischen Integration werden von Prof. Dr. Tanja A. Börzel, Jean Monnet Lehrstuhl für Europäische Integration und Leiterin der Arbeitsstelle Europäische Integration am Otto-Suhr-Institut für Politikwissenschaft der Freien Universität Berlin, veröffentlicht. Die Arbeitspapiere sind auf der gemeinsamen Internetseite von Jean Monnet Lehrstuhl und Arbeitsstelle verfügbar: http://www.fu-berlin.de/europa

The Berlin Working Papers on European Integration are published by Prof. Dr. Tanja A. Börzel, Jean Monnet Chair for European Integration and Director of the Centre for European Integration of the Otto Suhr-Institute for Political Science at the Freie Universität Berlin. The Working Papers are available on the joint website of the Jean Monnet Chair and the Centre: http://www.fu-berlin.de/europa

Die Autoren / The Authors

Fabienne Cap is a computational linguist currently doing her Ph.D. in the field of Statistical Machine Translation at the University of Stuttgart. Besides that she is interested in the automatic acquisition of Multiword Expressions from corpora using cross-lingual knowledge. In 2010, she worked in the EU-project Terminology Extraction, Translation Tools and Comparable Corpora (TTC), where she focused on domain specific terminology extraction from texts. Contact: fritzife@ims.uni-stuttgart.de

Ulrich Heid teaches computational linguistics at the Universities of Stuttgart and (since 2008) Hildesheim. His major research interests are in tools and methods for corpus analysis and in data extraction from text corpora; these are also the topic of Heid's project in the Stuttgart basic research cluster SFB-732. He also works on electronic dictionaries and on terminology extraction (e.g. in the EU-project TTC). Contact: heid@ims.uni-stuttgart.de

Abstract

In this paper, we have given an overview of computational linguistic tools available to us, which can be used to produce raw material for the lexicographic description of a specialised language. The underlying idea of our method is the following: what is significantly more frequent in a domain-specific text than in a general language reference text may be a term (or collocation) of the domain. In the near future, our tools will be integrated in a web-based environment in order to make them available for text-based research, e.g. in the humanities, whenever needed. The researcher interested in term or phraseology candidate extraction of a certain domain would identify and upload texts to be searched, and the tools would be running on servers of e.g. computational linguistics centres. The researcher would select tools to be applied and receive the analysis results over the network.

Keywords: extraction of terminology, collocations, specialised phraseology

1. Introduction

In this paper, we present and discuss computational linguistic tools for the extraction of linguistic data from texts. Other than in Information Retrieval or Information Extraction, our focus is not on extracting factual data, but on identifying the linguistic form of discourses, e.g. by extracting their (specialised) vocabulary and phraseology.

It has become clear over the past two decades that domain-specified discourses are to some extent characterised by variation; technicians speak differently about a product (say, a car or a washing machine) than marketing people. And even technicians of one company use technical vocabulary not used in another company (*corporate language*). In political discourse, variation is often due to political convictions of the speaker or writer: in debates of the German Bundestag of the years 1994 and 1995, we have annotated, wherever possible, the name of the speakers and the political party which they are a member of[1]. Searching for words starting with the elements *'Kernkraft.*'* ('nuclear energy') or *'Atomkraft.*'* ('atomic energy'), respectively, gives an interesting distribution: compounds with *'Kernkraft-'* are used by all parties, with a slight underuse in the ecologist party. Compounds with *'Atomkraft-'*, however, are not used at all by members of CDU/CSU, the governing conservative party of the period in question. We find, however, massive use of such terms in discourses pronounced by ecologists. At this time, critical views on nuclear energy were expressed using the term *'Atomkraft'* (cf. *'Atomkraft – nein danke!*' ('atomic energy – no thank you!')).

These few examples may illustrate the interest of a detailed analysis of lexical material in specialised texts. Other fields closely related with the issues mentioned above are the identification of formulaic (recurring) expressions, and sentiment analysis. In all cases, there is a need to identify the lexical items and the word combinations (= phraseology) used by authors of texts from specific domains.

To be able to correlate the linguistic phenomena observed with external factors (such as the party a member of parliament belongs to), the texts under analysis need to be annotated with metadata:

[1] Cf. the demonstration on the following URL:
http://www.ims.uni-stuttgart.de/projekte/CQPDemos/Bundestag/frames-cqp.html
By using the *distribution*-button, frequency distributions over months of the session period and over parties can be obtained.

These are data describing the text, its author(s), the date of publication, the medium, etc. Depending on the research questions we want to put to the texts, we may need different kinds of metadata. For example, for a project on the (potential) impact of terrorism on legislation (are new laws motivated by the danger of terrorism?[2], the date of proposals for new laws, the party of the proposers, their role in political decision making etc. may need to be annotated and correlated with their text production.

Even though our tools were primarily designed for lexicographic purposes (i.e. to provide raw material for dictionary making), we think that the procedures discussed in this paper can prove useful also for tasks like those evoked above: much of the facts or assumptions underlying certain actions (which manifest themselves in texts) can be uncovered by an analysis of the lexical material used by the author: single words, word combinations and multiword expressions.

2. Background

It is well-known that a large part of any language's specialised vocabulary is used to denote rather complex objects, properties and states of affairs. Thus, not next to single word terms, there are large quantities of multiword terms and of typical word groups related with terms, e.g. to express actions carried out with objects denoted by terms. This field of specialised multiword items includes multiword terms in the strict sense, as well as the phraseology of a specialised language. It is only in the course of the last 15 years that the phraseology of specialised languages has been analysed to some extent. As it is lexicographically relevant (i.e. needs to be included in a specialised dictionary) and important for a detailed text-based analysis of certain domains of knowledge (e.g. political sciences, sociology, etc.), we concentrate, here, on specialised phraseology.

We assume that the basic descriptive categories of general language phraseology carry over to specialised language, and we thus use terms like 'collocation[3]' very much the same way as general language lexicographers do. In the remainder of this section, we outline our view of (specialised and general language) collocations and then describe the data and the tools we work with.

[2] Cf. the ongoing project ``European Legislative Responses to International Terrorism'' (ELIT) at University of Mannheim. http://www2.sowi.uni-mannheim.de/lspol2/06forschung01.html.
[3] A *collocation* is a sequence of words that co-occur more often than would be expected by chance.

2.1 Notion of collocation

The term *collocation* has been used to denote a range of different phenomena: it has often been used synonymously with *co-occurrence* or *multiword expression*. We share the lexicographic view formulated by (Bartsch 2004): 'Collocations are lexically and/or pragmatically constrained recurrent cooccurrences of at least two items which are in a direct syntactic relation with each other'. This definition relies on criteria of lexical co-selection (a base selects its collocates), statistical significance of cooccurrence, and syntactic patterns.

2.2 Data for terminology extraction

Bergenholtz and Tarp (1995) suggested that a text collection (*corpus*) of one million words should be sufficient for the identification of the core terminology of a scientific domain. However, this holds only for a text collection that is relevant and central for the domain under analysis, and such a collection is not always directly accessible. Furthermore, when using statistical measures to identify domain terminology, it is preferable to have maybe less balanced but quantitatively more data at hand (in the range of 10 to 100 million words), as the quantity is assumed to level out deficiencies with regard to the composition of the corpus. For work on the language of a domain, group of persons, political party etc., obviously, it is important to ensure that the texts have been produced by authors from the respective group.

In our experiments, we use a juridical text collection which C.H. Beck publishers in Munich provided us within a recent cooperation. This collection covers the juridical sub-domain of Industrial Property Rights and trademark legislation: it is composed of the German juridical journal *'Gewerblicher Rechtsschutz und Urheberrecht'* (henceforth: GRUR) and amounts in total to ca. 78 million words (1946 to 2006). For more details about the GRUR corpus, see also Heid et al. (2008).

Obviously, the content of the GRUR text corpus is rather opportunistic than balanced – imagine the variety of different products that fall under trademark protection (e.g. the yellow colour used by the German postal services). However, we assume that the long period of publication levels out local terminological bursts in individual articles. Furthermore, a huge corpus is particularly relevant for the extraction of phraseology:

according to Evert (2004), only word combinations that occur at least 5 times in the corpus under analysis should be taken into account.

In order to automatically extract domain-specific terminology, there is a need for a corpus which the words and phraseological units of GRUR can be contrasted to. This comparison corpus should be unbiased, especially not biased to the juridical domain. We used a collection of different newspaper corpora that were available to us. In total, this collection (henceforth named GENLA for 'general language') amounts to roughly 200 million words; the composition of GENLA is given in Table 1. The considerable difference in corpus size of GRUR (78 million) and GENLA (198 million) is irrelevant, as the extraction algorithm incorporates relative frequencies instead of absolute frequencies when contrasting the two corpora.

Table 1: Composition of the general language text collection (GENLA).

name	newspaper	years	size (words)
FAZ	Frankfurter Allgemeine Zeitung	1996-98	70 million
FR	Frankfurter Rundschau	1992-93	40 million
STZ	Stuttgarter Zeitung	1992-93	36 million
ZEIT	Die Zeit	1995-01	51 million
GENLA			198 million

2.3 Description of our computational linguistic tools

A number of different tools are required to automatically preprocess the corpora under investigation in order to be able to extract terminologically relevant material. As a first step, the tools and procedures presented here aim at extracting all term and collocation candidates from the texts they are applied to (in this case from both, GRUR and GENLA). Then, in a second step, as described in Section 3, the term candidate lists are filtered in order to retain only domain-relevant items.

In case of single word term candidates, nouns, adjectives and verbs are most relevant. A *part-of-speech tagger* accounts for the automatic assignment of word classes (Section 2.3.1). For a languages with a fairly rigid word order, such as English, the information provided by a tagger is sufficient even for the extraction of collocations, as grammatical functions are typically encoded in such languages by positions: the first noun phrase to the right of the finite verb tends to be this verb's object. In contrast, positional criteria and case are often ambiguous in German (Ivanova et al. 2008). For

example, in the German phrase *'Lehrer fragen Schüler'* ('teacher ask pupils') it is not clear who asks whom, as both, teachers and pupils could be either subject or direct object. In contrast, this ambiguity does not arise in the English translation of the sentence. Furthermore, the variable word order of German allows the words of a collocation to occur not always adjacently: e.g. *'im Raum stehen'* in *'Also **steht** das Gerücht weiter **im Raum**.'* ('Thus, the rumour is still to be dealt with'); a pattern based extraction routine on tagged text would miss such instances if they fall outside the window of N tags under consideration. Seretan (2008) reported that deep syntactic analysis (henceforth called *parsing*) has a positive impact on the precision of collocation extraction and Heid et al. (2008) found that (dependency) parsing improves recall considerably. We thus use a dependency parser in our collocation extraction work (see Section 2.3.2 for a description).

Finally, in order to identify collocations that appear as compounds (e.g. *'Patenterteilung'* – *'Patent erteilen'* ('to grant a patent')) and to be able to group morphologically related collocations together (e.g. *'Patent erteilen*' and *'erteiltes Patent'* ('granted patent')), a morphological analysis is required to access the inherent structure of the words involved. The reduction of words to their stems (called *stemming*) would not be sufficient, as unrelated but formally similar words might be grouped together (e.g. *'Beton' (*'concrete') vs. the verb stem *'beton$_V$-*' ('emphasize'), *'Betonung'* ('emphasis')). A description of a morphological analyser which provides a detailed morpheme analysis is given in Section 2.3.3. Details of how to group morphologically related collocations together are presented in Section 4 below.

Table 2: Example analysis of the POS-tagger TREETAGGER.

token	POS	lemma	POS glosses
<s>			
Das	PDS	die	substituting demonstrative pronoun
so	ADV	so	adverb
eingerichtete	ADJA	eingerichtet	attributively used adjective
System	NN	System	noun
war	VAFIN	sein	finite form of auxiliary verb
indessen	ADV	indessen	adverb
nicht	PTKNEG	nicht	negation particle
erfolgreich	ADJD	erfolgreich	predicatively used adjective
.	$.	.	
</s>			

2.3.1 Part-of-speech tagger

TREETAGGER is a freely available highly efficient tagger for German (Schmid 1994). It is widely used in *Natural Language Processing* (NLP) research. TREETAGGER annotates part-of-speech tags (*word class categories*) of the Stuttgart-Tübingen TagSet (STTS[4]) to text, where word boundaries are identified. As a by-product of tagging, the sentence borders are indicated and the base form (*lemma*) of the words that are contained in the tagger's lexicon are also provided.

Domain-specific texts may contain words that are unknown to the lexicon of TREETAGGER. In such cases, the POS-tag is guessed from the context and <*unknown*> is provided in place of the lemma. Note however that the tagger lexicon can be enriched with domain specific terminology with a moderate manual effort, to further enhance tagging quality. An example analysis of TREETAGGER for the sentence *'Das so eingerichtete System war indessen nicht erfolgreich.'* ('The system arranged in this way was however not successful.') is given in Table 2.

2.3.2 Dependency parser

FSPAR is a broad coverage dependency parser for German (Schiehlen 2003). We successfully used it for several differentcollocation extraction tasks in the past (see e.g. Heid et al. (2008), Fritzinger (2009)). FSPAR leaves both, structural ambiguities and label ambiguities unresolved, thus enhancing the probability of the correct analysis being among the results. Structural ambiguities often arise when the attachment of a prepositional phrase is not clear (cf. 'he saw the man with the telescope'). An example of a label ambiguity is the case ambiguity in *'Lehrer fragen Schüler.'* ('teacher ask pupils').

For the task of collocation extraction, the number of undesired analyses is not an obstacle: we simply extract all possible collocations, assuming that correct collocations are recurring more often than wrong combinations and that the latter ones are thus filtered out by our statistical procedures. FSPAR is a fast and highly efficient parser. It takes about 30 minutes to parse 10 million words. Figure 1 shows the FSPAR dependency analysis for the sentence *'Es gibt Länder, deren geltendes Recht die*

[4] //www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html

Patentierung von Pflanzen ausschließt.' ('There are countries, in which the applicable law excludes the patenting of plants.'). The output format[5]is to be read as follows:

Figure 1: Example analysis of the dependency parser FSPAR.

A column: position of a word in the sentence
B column: word as it occurs in the text (token)
C column: part of speech category (based on STTS)
D column: base form of the word (lemma)
E column: morpho-syntactic information (case, gender, number, tense, person,e tc.)
F column: position of a word's governor
G column: grammatical function of the word in this sentence (subject, object, adjunct, etc.)

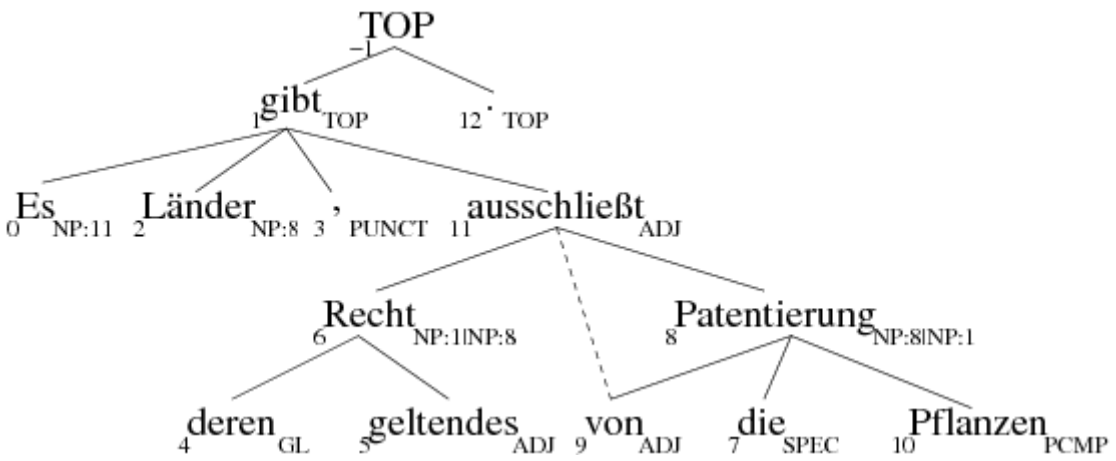
A	B	C	D	E	F	G
nr.	token	POS	lemma	morph.description	dep.	function
0	Es	PPER	es	Nom:N:Sg	1	NP:11
1	gibt	VVFIN	geben	3:Sg:Pres:Ind	-1	TOP
2	Länder	NN	Länder	Akk	1	NP:8
3	,	$,	,	|	1	PUNCT
4	deren	PRELAT	d	Gen:F:Sg|Gen:Pl	6	GL
5	geltendes	ADJA	gelten	|	6	ADJ
6	Recht	NN	Recht	Nom:N:Sg|Akk:N:Sg	11	NP:1|NP:8
7	die	ART	d	|	8	SPEC
8	Patentierung	NN	Patentierung	Nom:F:Sg|Akk:F:Sg	11	NP:8|NP:1
9	von	APPR	von	Dat	11|8	ADJ
10	Pflanzen	NN	Pflanze	Dat:F:Pl	9	PCMP
11	ausschließt	VVFIN	ausschließen	3:Sg:Pres:Ind*	1	ADJ
12	.	$.	.	|	-1	TOP

In order to enhance intelligibility of the example analysis, a dependency tree representation of the sentence is given in Figure 2. Note however that this tree representation is not directly provided by FSPAR, but can be drawn based on the

[5] **Explanation of POS categories in column C of Figure 1**: PPER: personal pronoun, VVFIN: finite main verb, NN: noun, PRELAT: relative pronoun, ADJA: adjective used attributively, ART: article, APPR: preposition
Explanation of morpho-syntactic descriptions in column E of Figure 1: Nom: nominative, Akk: accusative, Dat: dative,Gen: genitive, F: feminine, N: neutrum, Sg: singular, Pl: plural, 3: third person, Pres: present tense, Ind: indicative
Explanation of grammatical functions in column G of Figure 1: NP11: expletive subject, NP:1: subject, NP:8: accusative object, TOP: root node, PUNCT: punctuation, ADJ: adjunct, SPEC: specifier, PCMP: prepositional complement
Other explanations concerning Figure 1: #: morpheme boundary, |: label ambiguity, ||: structural ambiguity.

analysis result given in Figure 1 above. Basically, each node of the tree[6] consists of three parts, see e.g. the node *'Patentierung'* ('patenting'): the left subscript *8* refers to the word's position in the sentence, cf. column A in Figure 1. The middle part of the node contains the word as it appeared in the sentence (here: *'Patentierung*'), cf. column B 'token' in Figure 1. And finally, the right subscript denotes the grammatical function of the node in the sentence (cf. column G 'function' in Figure 1). Note that in FSPAR's internal notion, *NP:1* denotes subjects, while *NP:8* denotes direct objects. The edges of the tree are a visualisation of the dependency structure encoded in column F of Figure 1. The node *'Patentierung'*, for example, is dependent of the node at sentence position *11*, which is *'ausschließt'* ('excludes').

Figure 2: Tree representation corresponding to the dependency structure in Figure 1.



From the example sentence, we extract the following collocations: *'geltendes+Recht'* (adjective+noun), *'Patentierung ausschließen'* (verb+object) and *'Patentierung von Pflanzen'* (noun+von-PP, replacing a genitive attribute)[7].

It can be seen from the dependency tree in Figure 2, that even though *'Patentierung ausschließen'* does not occur adjacently in the original sentence, a verb-object relation between the two words can be identified. Collocations are extracted from the parsing output (as given in Figure 1) using PERL scripts that take into account the part-of-speech of the words (column C), the morpho-syntactic information (column E) and the governor information in column F of the parsing output. To give an example

[6] Each word of a sentence is represented as a node in a parse tree.
[7] In principle, collocations of any length can be extracted from the parsing output. However, as the statistical measures we use for contrasting are designed for pairs, we restrict the collocations to word pairs here.

for the extraction of e.g. verb+accusative-objectcollocations, consider the row of the word *'Patentierung'*: the POS-column indicates that it is a noun (cf. NN tag); the morpho-syntactic description says that it can be either nominative or accusative and its governor is the word at position 11. If this word is a main verb, a verb+accusative-object pair is found and extracted: in this case it is *'ausschließen'*. In order to accumulate data for collocation types, rather than instances, only the lemmas of the elements of the collocations are extracted, not their inflected forms.

2.3.3 Morphological analyzer

SMOR is a computational morphology system developed by (Schmid et al. 2004). It covers inflection and the productive word formation processes of German, namely derivation, transposition and compounding. It relies on a number of word formation rules and has a large lexicon (in total ca. 40,000 stems), thus providing good coverage. Figure 3 contains the SMOR analyses[8]of the words *'verkennt'* ('misconcieves'), *'anwendbares*' ('applicable') and *'Patenterteilungen'* ('grants of the patent(s)').

Figure 3: Example analyses of the computational morpholgy SMOR.

```
analyze> verkennt
verkennen<+V><2><Pl><Pres><Ind>
verkennen<+V><3><Sg><Pres><Ind>
verkennen<+V><Imp><Pl>

analyze> anwendbares
an<VPART>wenden<V>bar<SUFF><+ADJ><Pos><Neut><Acc><Sg><St>
an<VPART>wenden<V>bar<SUFF><+ADJ><Pos><Neut><Nom><Sg><St>

analyze> Patenterteilungen
Patent<NN>erteilen<V>ung<SUFF><+NN><Fem><Acc><Pl>
Patent<NN>erteilen<V>ung<SUFF><+NN><Fem><Gen><Pl>
Patent<NN>erteilen<V>ung<SUFF><+NN><Fem><Nom><Pl>
Patent<NN>erteilen<V>ung<SUFF><+NN><Fem><Dat><Pl>
```

[8] The tags of the Smor example analyses are explained in the following: **word class tags**: <+ADJ> (adjective), <NN> (noun), <+V> (verb); **word part tags**: <VPART> (verb particle), <SUFF> (suffix); **person tags**: <2>, <3>; **tense tags**: <Pres> (present); **mood tags**: <Ind> (indicative); **comparison tags**: <Pos> (positive); **gender tags**: <Neut> (neutrum), <Fem> (feminine); **case tags**: <Nom> (nominative), <Gen> (genitive), <Dat> (dative), <Acc> (accusative), **number tags**: <Sg>(singular), <Pl> (plural).

3. Extraction of domain-specific vocabulary

3.1 Single word term candidates

The underlying hypothesis of our methodology is the following: what is significantly more frequent in a domain-specific text than in a general language reference text, may be a term (or collocation) of the domain. This goes back to the approach by Ahmad et al. 1992, where relative frequencies of items from the domain-specific text are compared with the relative frequencies of the same lemmas in text not biased to a given domain We use the formulas given in Figure 4 to first calculate the relative frequencies of term candidates and then the quotient of their occurrence in specialised vs. general language text. The absolute frequency of the term in the domain-specific text (here: GRUR) is referred to as f_{spec}, and in the general language text (here: GENLA) f_{gen}, respectively.

Figure 4: Formulas for the calculation of relative frequencies and quotient of occurrence.

$$r_{spec} = \frac{f_{spec}}{\#corpus-size} \qquad q_{spec/gen} = \frac{r_{spec}}{r_{gen}}$$

$$r_{gen} = \frac{f_{gen}}{\#corpus-size}$$

(relative term frequency) (quotient of occurrence)

Only term candidates that are either *adjectives*, *nouns* or *verbs* are included in the comparison. For each word class, a separate comparison is performed. We use Perl scripts on word class annotated text (cf. Section 2.3.1 for details) to extract adjectives, nouns and verbs and to compute relative frequencies and the quotient of occurrence. The output of the comparison consists of two files: (i) words found exclusively in GRUR and (ii), words found in both texts, but which are considerably more frequent in GRUR. This procedure is obviously domain- and language-independent. There is a smooth transition from domain-relevant to irrelevant terms in the latter group. A brief manual

inspection of the data is necessary to draw the line between the top of the list, with a high density of domain-relevant terms, and the body of the list consisting of less relevant terms (e.g. general juridical terms) and general language items. Figure 5 illustrates the whole comparison procedure for single word terms.

Figure 5: Methodology of comparison.

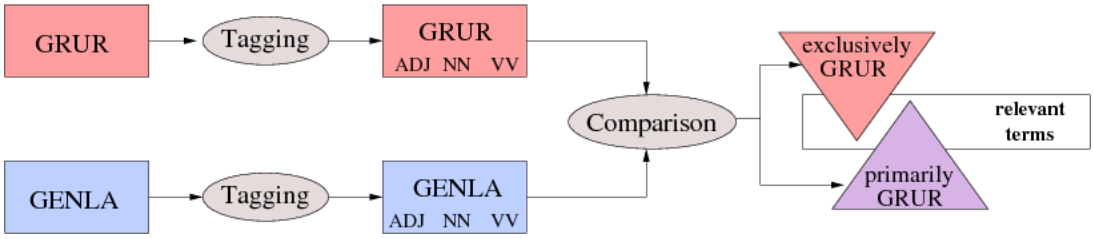


Table 6 (a) contains the most frequent nouns extracted from GRUR before the filtering. As can be seen, there are many general juridical terms, such as *'Beklagte'* ('defendant') or *'Recht*' ('law'), other terms belong to the subdomain of trademark legislation (e.g. *'Marke*' ('trademark') or *'Schutz'*('protection')). However, all of these high-frequent nouns are rather general and need to be filtered in order to assess more specific terms of the domain. If a corpus of general juridical texts (not biased to one juridical subdomain) was available, such filtering could be carried out by comparison with its contents.

Nouns that occur exclusively in GRUR (with no occurrences in GENLA) are given in Table 6 (b). Note that some of them were unknown to the tagger lexicon, but instead of extracting the <*unknown*>-tag, we used the surface form of the word (e.g. the genitives *'Anmelders*', *'Streitpatents'*), alternatively, we could generate lemma hypotheses. The nouns in Table 6 (b) are sorted by frequency. Obviously, all of them are highly specific terms, such as e.g. *'Verkehrsgeltung'*('validity') or *'Kennzeichnungskraft'* ('distinctiveness of a trademark').

Table 6 (c) contains nouns that occurred both in GRUR and GENLA. Note that this list is sorted by the quotient $q_{spec/gen}$ as calculated using the formula introduced in Figure 4 above. The results in Table 6 (c) show, that *more frequent* does not automatically mean *more relevant for the domain*: consider e.g. *'Warenzeichenrecht'* ('trademark legislation') which occurred 7,711 times in the 78 million word corpus GRUR, but only once in the 198 million word corpus GENLA, resulting in a very high occurrence quotient. In contrast, *'Unterscheidungskraft'* ('distinctive character')

occurred 13,095 times in GRUR, but also 4 times in GENLA, thus yielding a lower quotient than *'Warenzeichenrecht'*.

Table 3: Extracted noun term candidates from GRUR.
(a)+(b) are sorted by frequency, (c) is sorted by quotient of occurrence.

(a) most frequent in GRUR

noun	freq.
Beklagte	208,117
Recht	132,065
Entscheidung	121,389
Marke	117,453
Frage	114,084
Fall	103,309
Schutz	100,260
Ware	97,903
Gesetz	84,659
Erfindung	84,214

(b) exclusively in GRUR

noun	freq.
Verkehrsgeltung	6,755
Kennzeichnungskraft	6,444
Anmelders	5,979
Verbandsübereinkunft	5,306
Streitpatents	3,832
Patentanspruchs	3,725
Ausführungsform	3,602
Zeicheninhaber	3,550
Klagepatents	3,364
Beschwerdesenat	3,177

(c) primarily in GRUR

noun	quot.	freq.
Warenzeichenrecht	19,574	7,711
Patentfähigkeit	15,550	6,126
Prüfungsstelle	12,776	4,911
Anmelderin	9,410	11,122
Nichtigkeitsverfahren	9,382	3,696
Unterscheidungskraft	8,310	13,095
BT-Druck	7,689	3,029
Patentanspruch	7,611	14,993
Verlagsrecht	7,267	2,863
Diensterfindung	7,244	2,854

The same comparison procedure was applied to filter adjectives and verbs. However, there were less useful candidates found exclusively in GRUR, as this group contained a lot of wrongly tagged material (such as foreign language items or abbreviations). Examples for domain-specific verbs primarily found in GRUR include *'unterfallen*' ('to be categorised (as)'), *'abbedingen'* ('to waive sth.'), *'derogieren*' ('to derogate sth.'), examples for specific adjectives comprise e.g. *'neuheitsschädlich*' ('prejudicial to novelty') or *'streitgegenständlich'* ('litigious').

3.2 Collocations candidates

The following collocation patterns were considered interesting from a lexicographical point of view and thus extracted: *adjective+noun*, *noun+genitive attribute* and *verb+object*. They may also be relevant for a general impression of the phenomena dealt with in the text.

In order to extract domain-specific collocations, the relative frequencies of the collocation candidates extracted from GRUR are compared to their frequencies in GENLA, a procedure straigthforward to the one we applied to extract domain-specific single word terms in the previous section. The results of this comparison are given in Table 4a. It can be seen that this procedure yields many subdomain-specific terms, such as e.g. *'Warenzeichen benutzen'* ('make use of a trademark') or *'Patentanmeldung einreichen*' ('to file a patent application'). Obviously, the candidate list still contains a few trivial combinations (e.g. *'Revision+rügen'*, ('to find fault with'+ 'appeal')) and artifacts of the analysis (*'Anmerkung+sehen'*, from *'siehe Anmerkung N'* ('cf. note N').

Table 4: Collocation candidates extracted from GRUR after comparison to GENLA.

(a) exclusively GRUR (sorted by frequency)

object	verb	freq.
Priorität	nehmen	925
Revision	rügen	677
Patent	erklären	670
Warenzeichen	benutzen	578
Verwechslungsgefahr	bejahen	552
Verkehrsgeltung	erlangen	529
Erfindung	offenbaren	490
Unterscheidungskraft	haben	465
Eintragung	versagen	460
Verwechslung	hervorrufen	460

(b) primarily GRUR (sorted by quotient of occurrence)

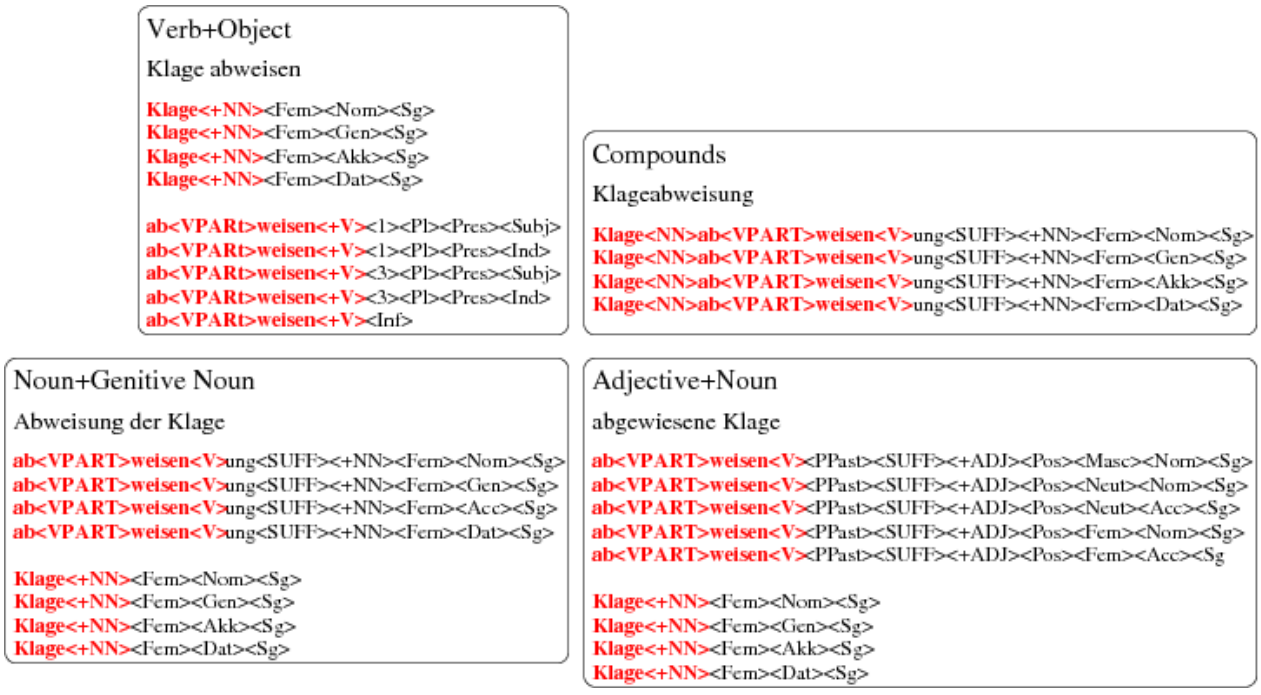
object	verb	quot.	freq.
Anmeldung	einreichen	3,086	1,216
Anmeldung	zurückweisen	3,025	1,192
Anmerkung	sehen	2,596	1,023
Ware	unterscheiden	2,066	814
Zeichen	eintragen	1,972	777
Erfindung	benutzen	1,898	748
Rechtsfehler	lassen	1,652	651
Patentanmeldung	einreichen	1,477	582
Zeichen	benutzen	1,432	1,693
Verwechslungsgefahr	verneinen	1,421	560

Besides verb+object collocations, we analogously extracted domain-relevant adjective+noun collocations, such as e.g. *'geistig+Eigentum'* ('intellectual property'), *'gewerblich+Schutzrecht*' ('industrial property right') and noun+genitive attribute collocations, e.g. *'Schutz des Eigentums*' ('protection of property') or *'Eintragung der Marke'* ('registration of trademark').

4. Identification of collocation groups

We found that there are a number of lexical concepts that appear across different syntactic collocation patterns, e.g. *'Gebrauchsmuster anmelden*' ('to register a utility model'), that was also found as adjective+noun, *'angemeldetes Gebrauchsmuster'* and noun+genitive attribute, *'Anmeldung eines Gebrauchsmusters*'. For lexicography, it is useful to group such items, such that the dictionary authors can provide information on variants and preferences. But such grouping also allows us to find more examples of the same *idea*: the variants all go back to some verb+object pair, but show up in different forms in the text. Therefore, the verb+object collocations are used as a basis for grouping, as nominalisations and adjectival participles are morphologically derived from verbs.

Figure 6: Grouping of morphologically related collocations.



In order to relate nominalisations and adjectival participles to their underlying verbal concepts, it is required to know about the internal structure of (complex) words. We use a computational morphology system (SMOR, see Section 2.3.3) to automatically produce a detailed analysis of the words. Figure 6 contains the morphological analyses of all collocational surface forms found for *'Klage abweisen*' ('to dismiss a charge').

Obviously, all realisations share the same root lexemes, which are represented identically in the morphological analysis (cf. **red** font in Figure 6). It is thus sufficient to run a simple PERL-script on the morphologically analysed collocations to find related collocations and group them together.

We found that the vast majority of collocations occurred in only one or two categories. However, about 20,000 collocations occurred in the three surface forms adjective+noun, noun+genitive attribute and verb+ object, and about 1,000 were found to occur in all of the four investigated surface forms (including compounds)[9] Some examples of these two latter groups, along with their distribution over the three or four surface forms, are given in Table 5.

There are cases where one pattern is (more or less clearly) prominent (e.g. adj+nn for *'Marke+eintragen'*: *'eingetragene Marke'*, nn+gen for *'Patent+vernichten*', compound for *'Warenzeichen+anmelden'*), while for others, the distinction is less clear *'Schutzbereich+einschränken*, *'Nutzungsrecht+einräumen'*).

These analyses show the degree of variation in multiword expressions of a given sublanguage, and the degree to which certain forms are lexicalised. Where there is variation, it is useful to capture all possible variants, in order to achieve a better coverage with respect to a certain *concept* (e.g. when it comes to (automatic) translation: it may be useful to consider the variants together). Where there are clear preferences, these need to be marked in a dictionary. More work on this morphological grouping is certainly necessary, to better assess its usefulness: across domains, and possibly also for work on general language.

Table 5: Distribution of collocation occurrences across different syntactic patterns.

	Marke eintragen	Schutzbereich einschränken	Patent vernichten	Nutzungsrecht einräumen	Warenzeichen anmelden
adj+nn	46.99%	7.69%	14.86%	12.61%	14.20%
nn+gen	28.92%	47.06%	68.78%	32.32%	20.27%
vv+obj	14.19%	43.89%	16.36%	33.25%	9.49%
compound	9.89%	1.36%	-	21.83%	56.04%
glosses	register+ trademark	restrict+domain of protection	destroy+ patent	grant+ right to use sth.	register+ trademark

[9] For more detailed information about the quantitative distribution, see Fritzinger/Heid (2009).

5. Beyond word pairs and lemmas

Domain-specific terminology is not restricted to single word terms or word pair collocations. There are also longer phraseological units that are relevant to the domain. There is a broad range of different syntactic patterns observable, but only few instances of a given pattern are phraseologically relevant. It is thus not efficient to define longer syntactic patterns and to apply the pattern-based extraction approach as described in Section 3.2 above.

Another aspect of specialised phraseology which can be identified in corpora are preferences with respect to the morphological form (e.g. singular vs. plural, definite vs. indefinite vs. null article, etc.). Starting from word pair collocations that are relevant for the domain, we take one step back to the syntactic parsing analysis of the sentences in which they occurred and search for words (e.g. adjectives or adverbs) that are in a direct syntactic relation to the collocation. In this way, the collocation *'Schutzbereich einschränken*' ('to restrict the domain of protection') might be found to be frequently extended by the adverb *'rechtskräftig'* ('legally binding') to *'Schutzbereich rechtskräftig einschränken*'}.

However, the data requires filtering to retain only relevant modifications. We use distributional information on morpho-syntactic variability to do so. The test case will be adverbial modification of verb+object collocations, but any syntactic collocation pattern could be used. Next to the lexical variance with respect to the adverb, we also consider the use of a determiner (none vs. definite or indefinite) and the number of the noun. In German, different such morpho-syntactic features can sometimes make a huge difference in terms of semantics, e.g. in the case of *'in+Gang+kommen'* (singular, no article: 'to be set in motion') and *'in+die+Gänge+kommen'* (plural, definite article: 'to get organised'). All features are automatically collected using PERL scripts on the parsed texts, and in the following counted and grouped to give an overview of their distribution.

The distributions of morpho-syntactic features for the general-juridical collocations (a) *'Recht+geben'* (lit.: 'to give right') and (b) *'Schaden+ersetzen'* ('to make up for a damage/loss') are given in Table 6[10]. It can be seen from Table 6 (a) that *'Recht+geben'* mostly occurs with a definite article in the domain specific text GRUR, while in the general language (GENLA) it is mostly used without an article. This

[10] In the following, we give short descriptions of the columns in Table 6: **Article use (art.)**: no, definite (def.), indefinite (indef.); **number of the noun (num.):** singular (sg), plural (pl); **adverbial modification (mod.)**.

difference encodes two readings of the word group made *'Recht+geben'*: using a definite article as e.g. *'jemandem das Recht geben etwas zu tun'* (lit.: 'to give s.o. the right to do sth.') means 'to entitle someone', while without article, e.g. *'Ich gebe Dir in diesem Punkt Recht'* (lit.: 'I agree with you in this point') is to be read as 'to concede a point to someone'.

Table 6: Distribution of morpho-syntactic features for 'Recht+geben' ('to give s.o. the right to do sth.').

Domain	art.	num.	mod.	distrib.
GRUR	def	sg	no	29.28%
	def	sg	yes	27.10%
	no	sg	yes	14.49%
	no	sg	no	12.17%
GENLA	no	sg	no	31.09%
	no	sg	yes	23.83%
	def	sg	no	18.65%
	def	sg	yes	13.99%

6. Summary and future work

In this paper, we have given an overview of computational linguistic tools available to us, which can be used to produce raw material for the lexicographic description of a specialised language. We showed examples of tools that extract single word and multiword terms, phraseological word groups, as well as illustrative material (e.g. example sentences), from large collections of German juridical texts.

Porting the tools and the approach to other domains and sciences is relatively straightforward; it may require updates of the lexical resources used by the tools, which can be carried out with comparatively little manual effort.

In the medium term, we envisage the tools (or variants thereof) to be usable also outside lexicography; currently, we use them only on our infrastructure to provide services. However, work in the European project CLARIN[11] and its German national counterpart, D-SPIN[12], aims among others at making such tools available via the Internet, as web services. The idea is for text-based research, e.g. in the humanities, to

[11] http://www.clarin.eu
[12] http://www.d-spin.org

be able to rely on computational linguistic tools over the web, whenever needed. The researcher interested in term or phraseology candidate extraction would identify and upload texts to be searched, and the tools would be running on servers of e.g. computational linguistics centres. The researcher would select tools to be applied and receive the analysis results over the network. A proof of concept implementation has been created within the D-SPIN project: it is called WEBLICHT[13] (web-based linguistic chaining tool) and provides several dozen tools from different European computational linguistics centres (Hinrichs et al. 2010); WEBLICHT *knows* which tools can be combined (in the sense of the tool chains described in this paper), and it supports the uploading of texts and the interactive inspection of analysis results.

If the WEBLICHT type of tools is maybe still focused on general language and generic tasks, similar models of interaction between computational linguistic tool providers and users from, e.g. the humanities, are imaginable for specific tasks. Examples include the search for news articles describing particular types of events: in a cooperation with the *Max-Planck-Institute für Ausländisches und Internationales Strafrecht* (MPICC, Freiburg im Breisgau), a few years ago, we extracted articles on homicide-suicide events from a large stream of press articles (by using conventional, non-web techniques). Other examples are focused opinion mining, or the identification of motivation patterns in legislation (which laws or proposals are motivated, e.g. by the financial crisis, or by the danger of terrorism?).

In such specialised setups, obviously the tool components need to be combined each time in an appropriate specific way. In an experimental fashion, we have put together the tools described here into a web service chain of the WEBLICHT type (cf. Fritzinger et al. (2009), Heid et al. (2010)). Other experiments may follow: we think that the tools are mature enough for more experiments to be made.

7. References

Ahmad, K./Davies, A./Fulford, H./Rogers, M., 1992: What is a Term? The semi-automatic extraction of terms from text, in: Snell-Hornby, M. (Ed.): Translation Studies – an Interdiscipline. Amsterdam, Philadelphia.

[13] https://weblicht.sfs.uni-tuebingen.de, restricted access only

Bartsch, S., 2004: Structural and Functional Properties of Collocations in English. A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-Occurrence. Tübingen.

Bergenholtz, H./Tarp, S., 1995: Manual of Specialised Lexicography – The Preparation of Specialised Dictionaries. Amsterdam, Philadelphia.

Evert, S., 2004: The Statistics of Word Cooccurrences: Word Pairs and Collocations. Ph.D. thesis, Universität Stuttgart.

Fritzinger, F., 2009: Using Parallel Text for the Extraction of German Multiword Expressions. Lexis – E-Journal in English Lexicology 4.

Fritzinger, F./Heid, U., 2009: Automatic Grouping of Morphologically Related Collocations, in: Proceedings of the 5th Corpus Linguistics Conference.

Fritzinger, F./Kisselew, M./Heid, U./Madsack, A./Schmid, H., 2009: Werkzeuge zur Extraktion von signifikantenWortpaaren als Web Service, in: GSCL-Symposium Sprachtechnologie und eHumanities. Universität Duisburg Essen.

Heid, U./Fritzinger, F./Hauptmann, S./Weidenkaff, J./Weller, M., 2008: Providing Corpus Data for a Dictionary for German Juridical Phraseology, in: Storrer, A./Geyken, A./Siebert, A./Würzner, K.-M. (Eds.): KONVENS'08: Text Resources and Lexical Knowledge – Selected Papers from the 9th Conference on Natural Language Processing. Berlin, New York.

Heid, U./Fritzinger, F./Hinrichs, E./Hinrichs, M./Zastrow, T., 2010: Term and Collocation Extraction by Means of Complex Linguistic Web Services, in: LREC'10.

Hinrichs, M./Zastrow, T./Hinrichs, E., 2010: WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure, in: LREC'10.

Ivanova, K./Heid, U./Schulte im Walde, S./Kilgarriff, A./Pomik´alek, J., 2008: Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case, in: LREC'08.

Schiehlen, M., 2003: A Cascaded Finite State Parser for German, in: EACL'03.

Schmid, H., 1994: Probabilistic part-of-speech tagging using decision trees, in: Proceedings of the international conference on new methods in language processing.

Schmid, H./Fitschen, A./Heid, U., 2004: SMOR: A German computational morphology covering derivation, composition and inflection, in: LREC '04.

Seretan, V., 2008. Collocation Extraction Based on Syntactic Parsing. Ph.D. thesis, University of Geneva.