Automatic Extraction of the Phraseology of a Legal Subdomain

Fabienne Fritzinger, Ulrich Heid, Nadine Siegmund
University of Stuttgart, Germany

Keywords: specialised terminology, dependency parsing, collocation extraction, morphosyntactic variability, lexical variability

1. Introduction

Many domains of expertise are rather fragmented, and even if different subdisciplines share the core of a specialised language, their terminology and their phraseology may differ considerably. This situation can also be found in juridical language. There is a core common to most subdisciplines, but each legal domain also has its terminology and phraseology. If this is accepted for terms, the fact that the same differentiation can be found in specialised phraseology, is much less acknowledged. However, if a jurist is supposed to produce a text on a domain that falls outside his specialisation, he may still know the terminology to be used, but likely much less the specialised phraseology of the ‘foreign’ domain. A good dictionary of juridical language should thus provide clear labels of subdomains wherever possible.

In this paper, we address this issue in the context of the extraction of single and multiword terms and of specialised collocations from corpus data. We intend to answer the following methodological and technical research questions: if we use simple frequency-based extractors on juridical texts from different domains, can these provide the phraseology of a given juridical subdomain by contrasting multiwords from different domains? And can we identify phraseological sequences longer than two items by systematically analysing the context of word pairs? We also wish to contribute to a more detailed description of the use of adverbs in juridical phraseology.

This paper is structured as follows: to describe the background of the work reported here, we discuss the data on which our study is based, as well as the tools used to process these data (Section 2). We then present the principles underlying our comparison of specialised vocabulary and give examples for single word terms and collocations (Section 3). In Section 4, we address variability within larger word groups, e.g. verb+adverb+object noun, as well as
morpho-syntactic preferences in the phraseology of certain subdomains. We conclude in Section 5.

2. **Background**

2.1. **Objectives**

We address the issue of identifying the specialised terminology and phraseology of a given legal subdomain in connection with term candidate extraction at large. Our tools are intended to cater for both, term extraction and variety extraction, in an integrated environment. We thus developed tools to extract specialised terms (e.g. *schuldnerisch*, *Insolvenzplan*, *Erteilung der Restschuldbefreitung*) and specialised phraseology (e.g. *Gläubiger befriedigen*, *Bestellung eines Insolvenzverwalters*, *drohende Zahlungsunfähigkeit*), and to identify which term and phraseology candidates belong to a particular legal domain. We focus on insolvency legislation, contrasting items which are particularly significant for this domain with data from IPR and trademark legislation. We also make an attempt at identifying items which can be classified as general juridical vocabulary, insofar as they are, roughly speaking, equally significant for both areas of specialisation.

2.2. **Data**

In the context of a recent cooperation project, C.H. Beck publishers in Munich provided us with a huge juridical text collection, covering the domain of insolvency legislations as well as the domain of IPR and trademark legislation.

The insolvency subcorpus is of ca. 10 million words; it contains the relevant law texts, a commentary book (the *Münchner Kommentar*), as well as texts from the specialised journal *NZI, Neue Zeitschrift für das Recht der Insolvenz und Sanierung*.

The IPR and trademark legislation corpus also contains law texts, but in particular no less than 60 years of specialised journal articles, from the journal *Gewerblicher Rechtsschutz und Urheberrecht* (GRUR) and its predecessors. The publication period covered is 1946 to 2006, and the subcorpus totals ca. 78 million words.

Even though both are rather opportunistic text collections that cover a wide range of topics – imagine the range of different products that fall under trademark protection – we successfully used GRUR for the extraction of specialised terminology in the past (Heid et al. 2008). What we do not have, however, is a ‘balanced’ juridical corpus, i.e. one that covers very different domains, or a collection of texts that could be seen as ‘general juridical language‘ (if such a construct exists at all).
As we use ‘general language text’ to contrast it with juridical language, there is a need for a corpus that is not juridically biased. For English, one would use BNC, the British National Corpus, as a mostly balanced, general language text. For German, we are not in the possession of a text collection comparable with the BNC\(^1\), and we thus approximate a ‘general language corpus’ by using newspaper data, from *Frankfurter Rundschau* (FR, 1992/1993; available through the ECI and ELSNET). We use the full production of nearly two years (about 40 million words), thereby trying to avoid a bias to, say, articles about economic or political topics.

2.3. **Tools**

2.3.1. **TreeTagger**

For the extraction of single word terms, it is sufficient to part-of-speech tag the corpora. We use TreeTagger (Schmid, 1994), a highly efficient probabilistic tagger that incorporates decision trees. It is freely available\(^2\) for several languages. The German implementation annotates part-of-speech tags of the Stuttgart-Tübingen TagSet (STTS)\(^3\).

2.3.2. **FSPAR**

Due to the variable word order in German, we are often confronted with syntactic relationships over long distances within a sentence. It is therefore essential to run a deep syntactic analysis on the whole corpus in order to reliably extract syntactically homogeneous pairs of, e.g., verbs and their objects, and thus collocations from that corpus.

In the past, we successfully used the finite-state based dependency parser FSPAR (cf. Schiehlen, 2003) for several collocation extraction attempts (cf. Heid et al. 2008, Fritzinger 2008). This parser is highly efficient due to its implementation based on finite-state cascades and the fact that it leaves structural or label ambiguities unresolved in the output.

Besides that, FSPAR relies on a large lexicon, which notably reduces the amount of false analyses, as the analyses contain only very few unknown words. In the following, we describe the parser’s output using the example sentence “*Wenn die Patente in den zwei Staaten derselben Person erteilt worden sind…*” (*If the patents have been granted to the same person in the two states…*).

---

\(^1\) The DWDS Kernkorpus would be of this type. It is an online interrogatable (http://www.dwds.de/woerterbuch), but cannot be held locally (and thus can not be used with our tools).

\(^2\) [http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html](http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html)

\(^3\) [http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html](http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html)
To enhance readability, the dependency analysis of this sentence is given in its tree representation in Figure 1. The corresponding FSPAR output given in Figure 2 is less graphically appealing but the relevant information it contains is easily accessible with PERL scripts. The output is to be read as follows:

- **1st column**: position of a word in the sentence
- **2nd column**: token
- **3rd column**: part of speech (based on the STTS tagset)
- **4th column**: base form (lemma)
- **5th column**: morpho-syntactic information (case, gender, tense, etc.)
- **6th column**: position of a word’s governor
- **7th column**: grammatical function (subject, object, adjunct, etc.)
In order to extract collocations of different syntactic patterns, we implemented a series of PERL scripts. An example for the extraction of a verb+object pair (based on the representation given in Figure 2) is given below: first, the third row of the parsing output, containing the part-of-speech information, is scanned for nouns (NN). Then, for each noun found, the dependency relation(s) in the second last row of the output is/are checked: the noun has to be dependent (dashed line in Figure 2) on a full verb (identifiable by its part-of-speech tag “VV.+”). Furthermore, the noun must be the direct object of this verb (which is indicated by “NP:8” in the last row of the noun entry in Figure 6). Only if both criteria are fulfilled, a verb+object collocation candidate is found and extracted (in this case: “Patente erteilt”, patents issued). Note that in the example given in Figure 2 the noun refers to a verb cluster (9/10/11, “erteilt, worden, sind”). However, we focus on full verbs (having POS-tags starting with “VV”), and thus leave out the auxiliary verbs (tagged with “VA.+”).

Besides the lemmas, we also extract morpho-syntactic information (e.g. article use, number of the noun) about the collocations, in order to determine a collocation’s morpho-syntactic fixedness (see Section 4 below). As can be seen from Figure 2, the parsing output’s 5th column contains the morphological description of the word.

3. Extraction of Domain-Specific Vocabulary

In this section, we describe our methodology to extract subdomain-specific single word terms and collocations by comparing the different corpora (Section 3.1). Then, we present some results of subdomain vocabulary: single word terms (Section 3.2.), general juridical vocabulary (Section 3.3.) and collocations (Section 3.4.).

3.1. Methodology of Comparison

In order to extract the vocabulary of a given subdomain (here: insolvency legislation), we compare term frequencies from the subdomain corpus (Nzi) with term frequencies from another juridical corpus (GRUR) as well as general language term frequencies (FR). Due to the fact, that the three corpora we have at our disposal differ considerably in size (Nzi: 10 million, GRUR: 80 million, FR: 40 million words), we use relative frequencies instead of absolute frequencies (see also Ahmad et al. 1992, ‘notion of weirdness’) for the comparison. The relative frequency of a term is calculated by simply dividing its absolute frequency by the corpus size.
The underlying hypothesis when comparing relative frequencies is the following: a term that is significantly more frequent in the specialised text than in the general text is likely to be a specific term of the domain covered by the specialised text.

Figure 3 contains an architectural sketch of our comparison pipeline: we compare single word terms separately for each of the word classes *nouns*, *verbs*, *adjectives*, which we identified by part-of-speech tagging. The relative frequency of a term is determined for each of the corpora. If the term occurs in all of the corpora, the relative term frequency from the specialised text (here: NZI) is divided (i) by its relative frequency in the general language text (here: FR) and (ii) by its relative frequency in the other juridical text (here: GRUR). From these divisions, we yield quotients: the higher these quotients are, the more likely the term at hand is a specialised term of NZI (named ‘primarily NZI’ in Figure 3). Those terms that do not occur in either FR or GRUR (named ‘exclusively NZI’ in Figure 3) are supposed to be either highly domain specific, or typing errors. In order to exclude erroneous words from our results, we set a frequency threshold of 5 for these terms.

3.2. Single Word Terms

In the present study, we focus on the extraction of terminology from the insolvency legislation domain. To do so, the insolvency legislation terms are compared with terms from general language text (FR) on the one hand and terms from a different juridical domain (GRUR) on the other hand, as described in the previous Section 3.1. (see also Figure 3 above).

This comparison yields four different groups of insolvency legislation terms: i) those occurring exclusively in NZI, ii) those occurring in both juridical corpora (NZI and GRUR), iii) those occurring in NZI and general language (GENLA) and iv) those occurring in all three text types.
In Figure 4, each (inter)section of the yellow ellipsis represents one of the four groups just described. In the following, we give example terms of each of them.

(a) Exclusively NZI

- Restschuldbefreiung: 6,794
- Insolvenzplan: 3,628
- Masseverbindlichkeit: 2,628
- Masseunzulänglichkeit: 2,329
- Verbraucherinsolvenzverfahren: 2,036
- Eröffnungsverfahren: 1,773
- Handelsgesetzbuch: 1,722
- Schuldnerbereinigungsplan: 1,530
- Regierungsentwurf: 1,380
- Sozialgesetzbuch: 1,351

(b) NZI and GRUR

- Insolvenzgericht: 8,550
- Gläubigerbenachteiligung: 1,459
- Insolvenzgläubiger: 1,084
- Massegläubiger: 956
- Insolvenzschuldner: 4,343
- Concursus: 588
- Insolvenzeröffnung: 2,852
- Gläubigerbefriedigung: 462
- Insolvenzforderung: 435

(c) NZI and FR

- Wohlverhaltensperiode: 730
- Arbeitnehmerbeitrag: 290
- Insolvenzgrund: 283
- Überschuldungsstatus: 201
- Rangklasse: 154
- Abstimmungstermin: 282
- Kreditforderung: 106
- Bund-Länder-Arbeitsgruppe: 203
- Kapitalersatz: 100
- Schuldnerbank: 94

(d) All

- Insolvenzverfahren: 25,786
- Schuldnerin: 8,950
- Verfahrenseröffnung: 4,946
- Insolvenzrecht: 4,353
- Insolvenzordnung: 2,066
- Anfechtbarkeit: 1,167
- Konkursgericht: 936
- Glaubhaftmachung: 909
- Eröffnungsbeschluss: 1,697
- Sequester: 761

Figure 5. Example results of different intersections.

Figure 5 (a) contains terms that exclusively occurred in NZI, ordered by their absolute frequency. As these terms do not occur in any of the two other texts, they are supposed to be highly domain-specific. This hypothesis is true for most of the terms (e.g. Restschuldenbefreiung, Insolvenzplan), but there are also some counter examples (indicated in red, e.g. Handelsgesetzbuch, Regierungsentwurf, Sozialgesetzbuch) that do not seem to be...
highly specific to the domain of insolvency legislation. The use of law texts in the insolvency text collection (see Section 2.2. above) might be a reason for their absence in the other text collections (GRUR and Fr), which are solely composed of journal and newspaper texts.

Not surprisingly, the intersection of NZI and GRUR (Figure 5(b)) contains only juridical terms. As this list is ordered in decreasing order of the NZI/GRUR quotient (Section 3.1), insolvency-oriented terms appear at its top. How to get a more balanced list of general juridical terms is discussed in Section 3.3. below. Note that in Figure 5 (b), the highest listed terms are not always the most frequent ones.

In Figure 5 (c), we show the results of intersecting NZI with the general language text Fr. Even this group contains terms relevant to the domain of insolvency legislation, as e.g. (Wohlverhaltensperiode, Überschuldungsstatus), but they are generally less frequent than the items in the other lists considered so far (Fig. 5 (a)+(b)). Furthermore, this list also contains terms that do not seem to be highly associated to insolvency legislation, such as Arbeitnehmerbeitrag or Rangklasse.

Finally, we present some results from the intersection of all investigated text types: Figure 5 (d) shows, that even this list contains a few relevant terms that one would like to include in a dictionary of insolvency legislation, at least in the top frequency range.

### 3.3. General Juridical Language

In this section, we try to answer the question as to whether it is possible to extract general juridical terms from the intersection of two juridical texts belonging to different juridical subdomains. We considered terms that occurred in both, NZI and GRUR, but not in Fr. For these, we calculated the quotient of relative frequency, as described in Section 3.1. above.

<table>
<thead>
<tr>
<th>Term</th>
<th>f_{NZI}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schadensersatzpflicht</td>
<td>188</td>
</tr>
<tr>
<td>Erwägungsgrund</td>
<td>171</td>
</tr>
<tr>
<td>Sachvortrag</td>
<td>148</td>
</tr>
<tr>
<td>Klageabweisung</td>
<td>93</td>
</tr>
<tr>
<td>Rechtsvorgänger</td>
<td>89</td>
</tr>
<tr>
<td>Streithelferin</td>
<td>82</td>
</tr>
<tr>
<td>Zivilprozeßrecht</td>
<td>80</td>
</tr>
<tr>
<td>Berufungsurteil</td>
<td>76</td>
</tr>
<tr>
<td>Beschwerdebegründung</td>
<td>74</td>
</tr>
<tr>
<td>Ausnahmetatbestand</td>
<td>74</td>
</tr>
</tbody>
</table>

**Figure 6. Terms of general juridical language.**

Terms of general juridical language are supposed to occur with similar frequency in either juridical subdomain. In the present experiment, we extracted those terms from the intersection
list, whose NZI/GRUR quotient was 1 (meaning that the term is equally frequent in both subdomains).

This list was then ordered by descending frequency of occurrence in Nzi. An extract of the result can be found in Figure 6. As far as we can see, these terms have no preference for either one of the two subdomains. The experiment can thus be considered as successful. However, we would like to further investigate this issue with texts of even more different juridical subdomains in the future.

3.4. Collocations

In order to extract domain specific collocations, we performed the same comparison procedures (see Section 3.1+3.2. above) as for single word term extraction. In the present work, we focus on three different collocation patterns: adjectives+nouns (abgewiesene Klage), verb+object (Klage abweisen) and noun+genitive attribute (Abweisung der/einer Klage). As we use parsed text, these collocations can be extracted regardless whether they occur adjacently or not (see also Section 2.3.2. above). In parallel to the single word results presented in Section 3.2. above (Fig. 5), Figure 7(a)-(d) contain examples for some collocation patters across all intersections. Interestingly, some judicial concepts appear in different surface forms, i.e. in different collocation patterns. An example is the verb+object collocation Restschuldbefreiung versagen vs. the noun+genitive attribute Versagung der Restschuldbefreiung, given in Figure 5 (a). We investigated this phenomenon of different surface realisations of the same concept in a separate study (Fritzinger/Heid 2009, in press).

3.4. Collocations

In order to extract domain specific collocations, we performed the same comparison procedures (see Section 3.1+3.2. above) as for single word term extraction. In the present work, we focus on three different collocation patterns: adjectives+nouns (abgewiesene Klage), verb+object (Klage abweisen) and noun+genitive attribute (Abweisung der/einer Klage). As we use parsed text, these collocations can be extracted regardless whether they occur adjacently or not (see also Section 2.3.2. above). In parallel to the single word results presented in Section 3.2. above (Fig. 5), Figure 7(a)-(d) contain examples for some collocation patters across all intersections. Interestingly, some judicial concepts appear in different surface forms, i.e. in different collocation patterns. An example is the verb+object collocation Restschuldbefreiung versagen vs. the noun+genitive attribute Versagung der Restschuldbefreiung, given in Figure 5 (a). We investigated this phenomenon of different surface realisations of the same concept in a separate study (Fritzinger/Heid 2009, in press).
4. Variability: Beyond Word Pairs

4.1. Morpho-Syntactic Behaviour

The specialised phraseology of a subdomain is not only characterised by specific collocational word pairs, but also, it seems, by specific uses of certain word combinations. Such uses manifest themselves in particular in the form of morpho-syntactic preferences and of combinations of more than two words (see below, Sections 4.2. and 4.3.).

Most morpho-syntactic preferences do not show up in isolation. We thus analysed number (singular vs. plural), article use (no article, definite, indefinite, etc.) and modifiability together. We group noun modification (by adjectives: *schriftliche Beschwerde einlegen*) and verb modification (adverbs: *Beschwerde schriftlich einlegen*) together (feature 'mod' in Table 1), even though we are aware of the differences (Storrer 2006).

For each subcorpus (NZI, GRUR, FR) we checked the frequency of the combined preferences, for each verb+object collocation. We could thus identify the most prominent combined preferences. For those which are prominent in NZI, we compared the respective figures from GRUR and FR.

The combination *Interesse+haben* can serve as an example, here (see also Table 1): in both NZI and GRUR, there is a clear preference for *ein Interesse (daran) haben (zu...)* over *Interesse haben*. In general language, however, the distribution is inversed: most occurrences have *Interesse haben* and not *ein Interesse (an...) haben*. It is not obvious whether the two expressions should be considered as syntactic variants, or whether they are semantically different enough to be considered as two different items (as is the case with *in Gang kommen* – ‘be set in motion’ vs. *in die Gänge kommen* – ‘get organised’).
Table 1. Preferences in morpho-syntactic behaviour of *Interesse haben*.

<table>
<thead>
<tr>
<th>Domain</th>
<th>art</th>
<th>num</th>
<th>mod</th>
<th>distrib</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZI</td>
<td>indef</td>
<td>sg</td>
<td>no</td>
<td>35.26%</td>
</tr>
<tr>
<td></td>
<td>indef</td>
<td>sg</td>
<td>yes</td>
<td>33.68%</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>sg</td>
<td>yes</td>
<td>15.79%</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>sg</td>
<td>no</td>
<td>9.47%</td>
</tr>
<tr>
<td>GRUR</td>
<td>indef</td>
<td>sg</td>
<td>no</td>
<td>39.38%</td>
</tr>
<tr>
<td></td>
<td>indef</td>
<td>sg</td>
<td>yes</td>
<td>32.80%</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>sg</td>
<td>yes</td>
<td>10.38%</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>sg</td>
<td>no</td>
<td>9.56%</td>
</tr>
<tr>
<td>FR</td>
<td>no</td>
<td>sg</td>
<td>no</td>
<td>47.31%</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>sg</td>
<td>yes</td>
<td>30.24%</td>
</tr>
<tr>
<td></td>
<td>indef</td>
<td>sg</td>
<td>yes</td>
<td>8.16%</td>
</tr>
<tr>
<td></td>
<td>indef</td>
<td>sg</td>
<td>no</td>
<td>8.16%</td>
</tr>
</tbody>
</table>

Figure 8 shows a similar example for the collocation *Vertrag+erfüllen*, where adverbial modification is much more prominent in juridical language than in general language, and where the triple *Vertrag vollständig erfüllen* is particularly prominent in insolvency texts (51.16%). This combination is not prominent in general language and not very much in our IPR texts. It seems to be a phraseologism of the insolvency law domain.

[def] Vertrag [sg] [adj] erfüllen
NZI: 51.16%, e.g. *den Vertrag vollständig erfüllen*
GRUR: 33.65%, e.g. *den Vertrag ordnungsgemäß erfüllen*
FR: 8.7%, e.g. *den Vertrag termingerecht erfüllen*

Figure 8. Combinatory preferences of *Vertrag+erfüllen*.

4.2. Lexical Variability I: Adjectives/Adverbs

Following on from the above example (*Vertrag vollständig erfüllen*), more triples of noun+verb+modifier (adjective or adverb) were analysed.

In order to avoid data sparseness issues in the general language data, we used more newspaper material than in the experiments reported above: Table 2 shows the portions of news texts we combined into a virtual subcorpus GENLA (for GENeral LAnguage). An analysis of noun+verb-collocations modified by adjectives or adverbs shows in many cases clear differences between the language of IPR and trademark and that of insolvency.
A case in point is Angabe(n)+machen: in trademark legislation, producers of products describe their products, and objectivity is prominent, as is the fact that such descriptions (Angaben) are non-ambiguous (nicht irreführend) and not part of advertisements (nicht werbend). In the field of insolvency legislation, Angabe(n)+machen denotes (financial) statements by companies and/or contractual partners. For these, obviously correctness is also an issue, but less discussed than the issue of Vorsatz vs. Fahrlässigkeit of incorrect or insufficient such indications, see Table 3.

<table>
<thead>
<tr>
<th></th>
<th>NZI</th>
<th>GRUR</th>
<th>GENLA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>vorsätzlich (25%)</td>
<td>irreführend (7%)</td>
<td>mutwillig (6%)</td>
</tr>
<tr>
<td></td>
<td>schriftlich (24%)</td>
<td>objektiv (5%)</td>
<td>bewußt (6%)</td>
</tr>
<tr>
<td></td>
<td>fahrlässig (21%)</td>
<td>verpflichtend (5%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>grob fahrlässig (6%)</td>
<td>wissentlich (5%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>werbend (5%)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Adjective/Adverb variability for Angabe machen.

4.3. Calculation of Lexical Variability

The degree of diversity a collocation exhibits in allowing for different adjectives/adverbs can be measured by using the semantic entropy measure. The following formula is adapted from (Melamed, 1997):

\[
H(C_{\text{domain}}) = - \sum_{a \in C_{\text{domain}}} a \log a
\]

where “C_{domain}” is a collocation of the respective domain and “a’s” are intervening adjectives or adverbs. The underlying hypothesis here is: the less variance, the lower the entropy value and the more likely the expression at hand is fixed.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Vertrag erfüllen</th>
<th>Rechnung tragen</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZI</td>
<td>1.301</td>
<td>1.619</td>
</tr>
<tr>
<td>GRUR</td>
<td>1.354</td>
<td>2.844</td>
</tr>
<tr>
<td>GENLA</td>
<td>2.528</td>
<td>3.202</td>
</tr>
</tbody>
</table>

Table 5. Semantic entropy values across different corpora.

In Table 5, we give the semantic entropy values measuring the degree of adjective/adverb variance of the collocations Vertrag erfüllen and Rechnung tragen, where high values indicate
high variance. Obviously, both expressions allow for more variance in the general language texts (GENLA) than in the juridical texts. In the case of *Rechnung tragen*, however, the degree of variance in GRUR (2.844) is almost as high as in GENLA (3.202), whereas there is a significant difference to NZI (1.619). Figure 9 (a)+(b) shows examples for possible adjective/adverb variances for the two collocations *Vertrag erfüllen* and *Rechnung tragen*.

<table>
<thead>
<tr>
<th>NZI (21)</th>
<th>GRUR (16)</th>
<th>GENLA (15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>vollständig (11) voll (4)</td>
<td>vollständig (6) gehörig (3) ordnungsgemäß (2) teilweise (2) ganz (1) persönlich (1) voll (1)</td>
<td>vollständig (1) voll (1) ordnungsgemäß (1) endlich (1) formal (1) persönlich (1) ernsthaft (1) kurzfristig (1) konsequent (1) wirtschaftlich (1) zuverlässig (1) rechtzeitig (1) selbstverständlich (1) termingerecht (1) korrekt (1)</td>
</tr>
<tr>
<td>beidseitig (3) einseitig (2) vorhanden (1)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Vertrag erfüllen

<table>
<thead>
<tr>
<th>NZI (40)</th>
<th>GRUR (554)</th>
<th>GENLA (146)</th>
</tr>
</thead>
</table>
| hinreichend (15) angemessen (13) umfassend (3) anderweitig (1) verfahrensrechtlich (1) ersichtlich (1) unzureichend (1) stärker (1) grundsätzlich (1) ausführlich (1) gebührend (1) tatsächlich (1) | hinreichend (130) angemessen (46) gebührend (34) ausreichend (32) voll (25) weitgehend (23) gleichzeitig (14) stärker (14) umfassend (8) offensichtlich (7) besser (7) beteiligend (7) ausdrücklich (6) zutreffend (5) verstärkt (4) grundsätzlich (4) | stärker (15) gleichzeitig (9) angemessen (6) weitgehend (5) voll (4) künftig (4) unzureichend (4) versäumt (4) gewiß (3) offensichtlich (3) besser (3) wirklich (3) offensichtlich (2) gebührend (2) umfassend (2) hinreichend (2) ...
| | | |

(b) Rechnung tragen

Figure 9: Adjective/adverb variance of collocations.

The examples in Figure 9 clearly visualise the higher degree of variance in GENLA as opposed to the two juridical text collections. In the case of *Vertrag erfüllen*, 11 of 21 adjective/adverb modifications found in NZI are realised with *vollständig*. Even though *vollständig* is also amongst the adjectives/adverbs that modify *Vertrag erfüllen* in GENLA, there, it is only 1 of 15 different adverbs. Furthermore, there seem to be typical juridical adjective/adverb modifiers,
as e.g. hinreichend and angemessen in Figure 9 (b), that occur significantly more prominently in the uses of Rechnung tragen in juridical texts as in GENLA.

5. Conclusions and Future Work

In this paper we have shown simple and generic tools for corpus analysis and for word frequency relationships can be used to identify term candidates and phraseology that is specific of a given subdomain of juridical language. We compared text from the domain of IPR and trademark legislation with documents from insolvency legislation, and with newspaper texts (used to simulate a general language corpus comparable to the BNC). The analyses of phraseological material showed that the subdomains prefer, at least to some extent, different phraseologisms, both, with respect to their form (and morphosyntactic properties) and the adverbial or adjectival modifiers they contain. This certainly has to do with the conceptual motivation of specialised phraseology: in a given domain, certain topics (and the phraseological ways to speak and write about them) are more important than in others. And terms like Angabe are polysemous with one reading more prominent in one domain and another reading in another domain.

In future work, we intend to cover more subdomains of the German juridical language, to assess possibilities to better identify the terms and the phraseology common to a broad range of legal domains and subsets of juridical discourse. We also plan to improve the recognition and analysis of phraseological multiword items of juridical language: many more combinations, and longer ones than just word pair collocations, seem to play an important role in juridical discourse of different subdomains.

References


