SUPPORTING THE LEGAL SUBSUMPTION PROCESS: DETERMINATION OF CONCRETENESS AND ABSTRACTNESS IN GERMAN LAWS USING LEXICAL KNOWLEDGE

Bernhard Waltl[1], Florian Matthes[2]

[1]Research Associate, Technische Universität München, Department of Informatics, Software Engineering for Business Information Systems, Boltzmannstraße 3, 85748 Garching bei München, DE
b.waltl@tum.de; https://wwwmatthes.in.tum.de/
[2]Professor, Technische Universität München, Department of Informatics, Software Engineering for Business Information Systems Boltzmannstraße 3, 85748 Garching bei München, DE
matthes@in.tum.de; https://wwwmatthes.in.tum.de/

Keywords: ***German Laws, Subsumption, Lexical Knowledge, GermaNet***

Abstract: *Determining, whether an act is applicable or not, is a non-trivial task. This is strongly associated with the interpretation of acts and the subsumed objects. Although subsumption is a complex process, it is also well-studied process and a central part of the legal theory and practice. Words are used as a base line during subsumption and allow for taxonomic structuring amongst itself, using hyper- and hyponym relationships. E.g., the word "energy" is a hypernym to "electricity". This paper determines the application scope of acts by accessing real-world knowledge stored in a German lexical knowledge database, called GermaNet. Based on the set of the ten largest German law texts we determine the average level of abstractness over a huge set of norms. Our research shows that words used in German acts are either a very high or a very low abstractness. Furthermore, we compared highly related laws from distinct countries, namely Austria and Germany, namely the act governing the liability for a defective product. We are able to automatically determine differences in the application scope of acts, respectively norms.*

1. Introduction

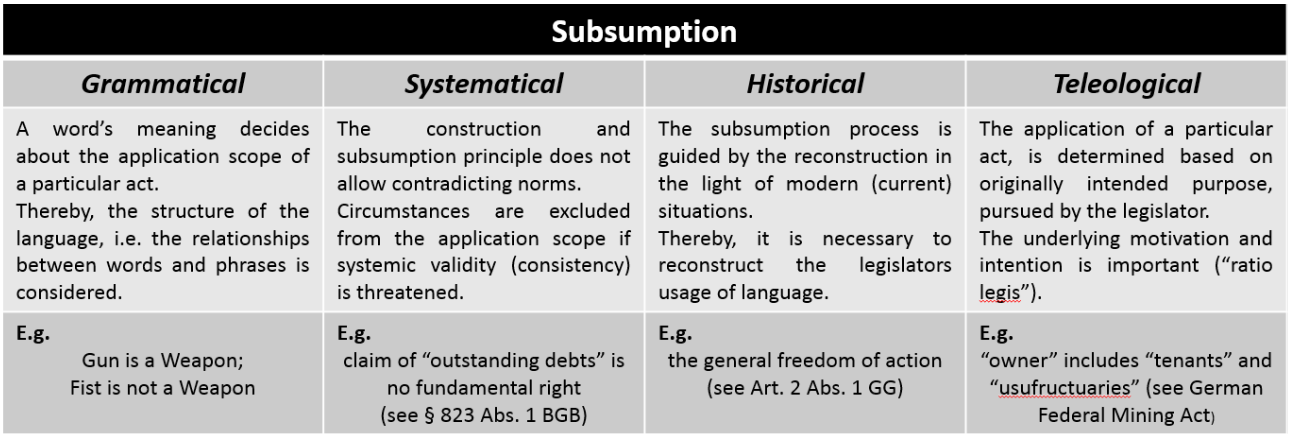
Subsumption is a fundamental process in the legal domain. It is necessary to determine whether case facts are within the scope of a particular legal act or not. The logic reasoning process behind the subsumption process in the legal domain is the well-studied syllogism (Larenz, Canaris 1995; Raabe et al. 2012). Syllogism is one kind of logical argument, which reasoning nature is deductive. The base line for the reasoning are two asserted true propositions. One proposition is the major premise, whereas the second is the minor premise. A famous example is about the mortality of people. Knowing that all people are mortal (major premise), and knowing that men are people (minor premise), the syllogism now allows us to make the logical conclusion, that all men are mortal. The so-called "middle (M)“, namely "people“, is the key, connecting the major and minor premise. Consequently, if a reasoner decides about the mortality of something, he could automatically decide for everything that is subsumed within people is mortal. In order to refine this structure, advanced taxonomies can be defined. E.g., if one would add the premise that a bachelor is a men, the automatically bachelor is mortal. This transitivity in the subsumption process calls for complex structures, representing real-world knowledge and allowing advanced reasoning process.

Obviously, different nouns refer to different real-world objects. It is the very nature of the word itself and of course the usage of the word in a particular context, which determines how many objects of the real world are affected. Natural language offers us mechanisms to address many objects at once, using a common word. E.g. the words “organism“ or “person“ refer to many different real-world objects. The implicit abstractness of those words is the actual key to subsume various objects. Again, real-world knowledge is necessary to determine the implicit relationship if “is-a“ between “person“ and “organism“. However, making this implicit relationship explicit is one of the major challenges to make a next step towards computer-assisted subsumption.

This paper analyzes the subsumption of words and their corresponding objects used in acts (see Section 2). Thereby, machine-readable real-world knowledge as described in Section 3 will be accessed. Section 4 continues by referring to existing work and aligning to related and prior approaches. Section 5 introduces the research objectives, and the used data. An definition about abstractness and concreteness of normative texts will be provided in Section 6. The paper's contribution, namely the automated measurement of abstractness and concreteness in German law texts, is in Section 7. Strength and limitations are critically reflected in Section 0 A usage scenario, namely law comparison, is given in Section 7.3. . Finally, the work concludes and shows further research directions in Section 9.

2. Subsumption in Legal Theory and Legal Practice

Several challenges, mainly addressing the limited expressiveness of natural language and the correspondence to real world situations, exist during the subsumption process of legal norms. However, the problem is well studied in legal theory and the subsumption process allows different techniques to provide solutions to the fundamental problems. The subsumption process, which is essential during the application of a norm has, according to Larenz and Canaris (see Larenz, Canaris 1995), four different dimensions, which are shown in the Figure below.



Subsumption			
Grammatical	***Systematical***	***Historical***	***Teleological***
A word’s meaning decides about the application scope of a particular act. Thereby, the structure of the language, i.e. the relationships between words and phrases is considered.	The construction and subsumption principle does not allow contradicting norms. Circumstances are excluded from the application scope if systemic validity (consistency) is threatened.	The subsumption process is guided by the reconstruction in the light of modern (current) situations. Thereby, it is necessary to reconstruct the legislators usage of language.	The application of a particular act, is determined based on originally intended purpose, pursued by the legislator. The underlying motivation and intention is important (“ratio legis”).
E.g. Gun is a Weapon; Fist is not a Weapon	**E.g.** claim of “outstanding debts” is no fundamental right (see § 823 Abs. 1 BGB)	**E.g.** the general freedom of action (see Art. 2 Abs. 1 GG)	**E.g.** “owner” includes “tenants” and “usufructuaries” (see German Federal Mining Act)

Figure 1: Differentiation of legal subsumption according to Larenz and Canaris

The essential differentiation of the subsumption process in legal theory and practice shows the great potential but also the great challenges for automated or semi-automated algorithmic reasoning approaches. The flexibility, and adaptability of language is its main strength but also its main weakness (Larenz, Canaris 1995). This research aims to support the grammatical subsumption process and evaluates the usage of lexical knowledge as introduced in the next section.

3. Representation of Real-World Knowledge

In order to enable computer-assisted subsumption over objects, an adequate representation, i.e. machine-readable form, of real-world knowledge is required. Thereby different approaches exist, whereas structuring real-world data in taxonomic structures is one of the most promising organizing principles (Hutchison et al. 2005). Enhancing those taxonomies with functionality regarding semantic constraints, lead to ontologies.

From early stages on, scientists and philosophers tried to set up a complete and comprehensive taxonomy, in which every observation, that can be addressed using words, has its unique place. Those approaches were introduced in the domains of nature and life sciences, i.e. biology. The taxonomy thereby is the organizing principle whereas each entity is classified regarding to its properties, so that it can be either distinguished or combined with other entities that are already in the taxonomy. The linguistics answer to the biological classification is of course a taxonomy over words of a language. Two prominent representatives are WordNet, „a large lexical database of English“ (see George 1995), and GermaNet, the German pendant to WordNet (see Hamp, Feldweg 1997). Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.

GermaNet is based on the same structure as WordNet, in which words are the key elements. Those are linked to each other regarding various relationships. Many words are synonyms to other words, such as „shut“ and „close“ or „car“ and „automobile“, therefore they are stored in a common set, namely the synset. Hence, the synonym relationship is expressed using a common set as storage. The most common relationship between those synsets is the super-subordinate relation, which defines hyper- and hyponym or ISA relation between words. Thereby general synsets such as „energy“ are linked to more specific synsets „electricity“ or „heat“. Beside of this relation type other relationships between words are stored in GermaNet, which are not relevant to us in this particular research. Integrating a huge amount of words into this structure leads to a comprehensive tree-like structure, whereas general and specific words can uniquely be identified. Due to the fact, that we solely investigated German and Austrian acts, we limited ourselves to the usage GermaNet (see Section 5.1.).

4. Related and Prior Work

The usage of ontologies is very common at the intersection of law and informatics. The guiding rationale is the creation of domain specific knowledge with proper semantic constraints, which allows the modeling of an excerpt of the real-world with regards to specified problems (Bench-Capon et al. 2012; Sartor et al. 2013). The principle behind ontologies addresses different aspects of modeling. The mentioned lexical knowledge databases WordNet and GermaNet are ontologies integrating lexical information and semantic relationships, such as hypernymity and homonymity.

The creation of ontologies to support legal information retrieval across different language barriers was the objective of LOIS (Lexical Ontologies for Information Sharing) by Tiscornia et al. (Tiscornia 2006). The main idea was to describe the legal domain of six different European languages and link the concepts between them. Words as placeholders for legal concepts are linked with each other and the resulting semantic lexicon supports multi-lingual information retrieval. WordNets architecture of ordering lexical information was the template for the architecture used in LOIS.

Textual representation as the interface between normative regulation through legislation and the effect on real-word problems inevitably calls for linguistics, since it is the science dealing with language in particular. From a linguistic point of view, several properties of legal texts could analytically be investigated (see McNamara et al. 2014; Köhler 2005). The main aspects that the analytical and quantitative linguistic is dealing with, concerns structure, coherence, hierarchy, etc. of text. Using linguistic methods to analyze words in legal texts and acts inevitably leads to an overlap to legal sciences. In legal sciences, the word is the basic information entity to communicate and interpret norms. Consequently, the wording is more crucial in legal sciences than it is in any other discipline. Which did other legal experts and philosophers already express „Law is a profession of words“ (Mellinkoff 2004).

The usage of lexical knowledge as a grounded measure for the abstractness, respectively generality, was also used in the domain of social environments. Thereby, Benz et al. used the measurement of a words position within the taxonomic tree as a comparison to other competing abstract metrics (Benz et al. 2011). During the analysis of generality of error- and noise-prone tags of social information systems, such as folksonomies, the semantic information contained in „well-defined“ semantic repositories, namely WordNet and GermaNet, served as base line and gold standard for the comparison.

4.1. Grammatical Subsumption

The normative character of laws consequently leads to abstract norms and regulations. A general formulation of norms is required not to only describe allowed and prohibited actions on a level of single and isolated actions and tasks, but to provide statements about a set of actions. To determine the application scope of norms several mechanisms exist (Larenz, Canaris 1995; Mellinkoff 2004). Within in this paper we are particularly interested in the processes based on the wording of an act.

The usage of words and nouns, that refer to many objects in the real world. The rationale behind is that words can be more or less concrete, whereas concrete words can be subsumed beneath different, more general nouns. E.g., „electricity“ can be subsumed under „energy“. This arises from straightforward linguistic definitions of abstractness as exemplary given by Brown (Brown 1958). Wording is what Larenz and Canaris call the grammatical subsumption principle and is the starting point of every subsumption principle. Larenz and Canaris argue, that it is obvious because the legislator uses the words in a common sense, so that the citizens and addressees can read it and determine whether they are effected or not (Larenz, Canaris 1995, p. 141).

Raabe et al. also use the subsumption of legal terms starting from the wording argument. They argue, that it is essential during the reconstruction of a legislative term, to start with the wording of a term and then -- if necessary -- progress to the more elaborated subsumption processes (Raabe et al. 2012). Raabe et al. also used during their research ontologies to provide domain specific structure and relationships, such as DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) (Gangemi et al. 2002). According to Raabe et al. the ontological categorization of DOLCE can be extremely helpful in order to extend a words meaning, which could not be captured by its original sense.

The wording mechanism has a strong focus on the usage and analysis of text that is used to express the meaning and intention of norms. Within this work, we will focus on the determination of the abstractness and generality of words within law texts, which is due to the objectivity of text suitable for algorithmic and computer-supported analysis.

5. Research Objectives

The paper summarizes a quantitative and descriptive empirical research. Thereby, we used publicly available German law texts and applied information extracted from a lexical knowledge database, i.e. GermaNet, to them. The resulting values are metric indicators for the textual abstractness and concreteness. The indicators allow for a comprehensive and effective analysis and comparison of acts and their application scope. The quantitative research based on textual information with consideration of lexical knowledge is addressing a few research questions. That aim to determine whether the measurement of abstractness and concreteness of concept used in law is in principle possible or not. The concrete questions are as follows:

1. What does abstractness and concreteness mean in the domain of legal language, i.e. text?
2. How to measure abstractness and concreteness formally and objectively and what is a possible quantification?
3. What is the distribution of abstract and concrete words used in German acts?
4. What are the limitations of the usage of lexical knowledge for quantifying textual properties?

5.1. Data

To perform the proposed analysis our research requires two different dataset, a German law text corpus and a lexical database containing real-world knowledge. The German law corpus was retrieved on the 10th of October 2014 from the platform *www.gesetze-im-internet.de* hosted and maintained by the Federal Ministry of Justice, represented by *Kompetenzzentrum Rechtsinformationssystem* (CC-RIS), which represents „almost the complete and current federal law“ (BMJ 2014). The second dataset is the lexical database GermaNet. GermaNet is the German pendant to the English WordNet. The number of different words distinguished by their meaning is called lexical unit, of which 121 810 are contained within GermaNet.

6. Concreteness and Abstractness in German Laws

The question what makes a law abstractness or concreteness, cannot easily be answered. As we have already stated out in the introduction, this paper addresses legal norms on a textual level, namely the level of grammatical, linguistic representation, i.e. words. There are three major reasons for these decisions and we will shortly summarize them:

1. Text is an objective artefact. As an artefact, it does not contain subjective biases. This certainly changes during the interpretation by a reader (for a discussion about text-reader interaction models see Schendera 2004), nevertheless the text itself remains unchanged.
2. The subsumption process is heavily determined by the words that are used to express a norms application scope. Although there are concept during the interpretation process like teleological reduction or teleological expansion, which influence - from case to case - the scope of a norm, the act and the concepts described with words, i.e. nouns, remain unchanged.
3. Due to its accessibility, the text is suitable for automatic processing. Beside of the fact, that natural language processing is challenging (see Section 0), algorithms allow to process a huge amount of data, which can create useful insights and reliable results.

In the following, we explain in more detail what abstractness and concreteness of nouns can be and we will give constructive definitions how to measure them. Furthermore, it will become clear how the used nouns and their usage within ordinary language and how this can be accessed by algorithmic and automated approaches effect the scope of norms.

6.1. Abstractness

Using a lexical knowledge database, the calculation of the abstractness of a word can be measured in various ways. For our approach, we decided to use a straightforward approach, namely counting the number of child nodes. Figure 2 (right side) visualizes the rationale behind our definition of abstractness.

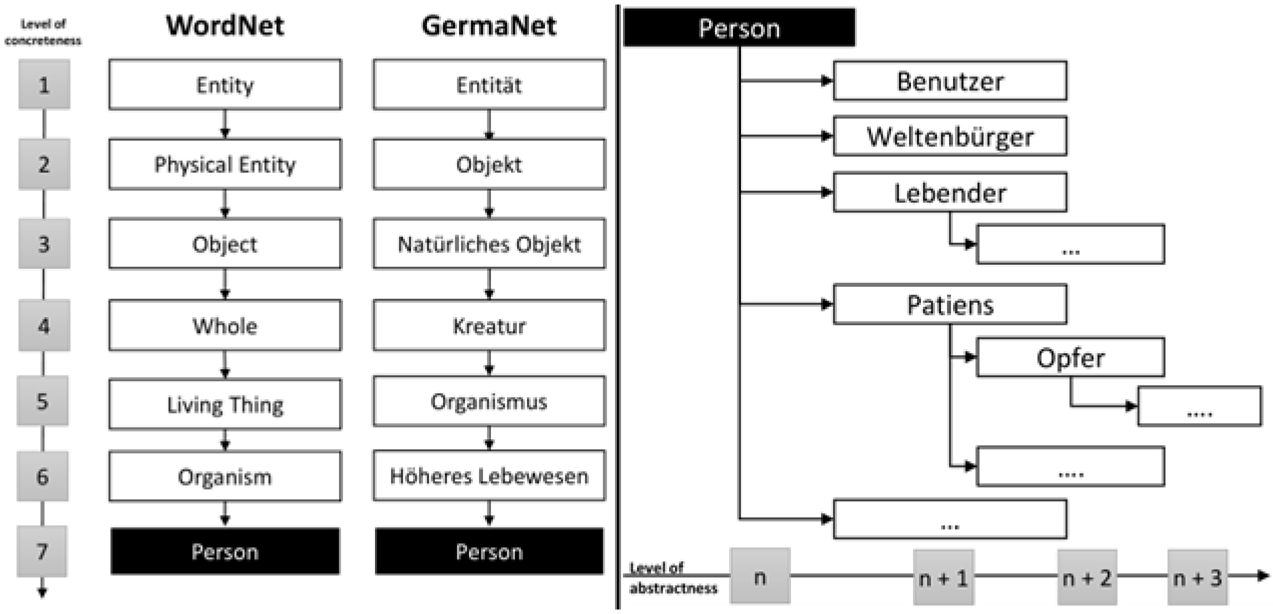


Figure 2: Concreteness (left side) and abstractness (right side) of the noun "Person" in GermaNet

Figure 2 starts with the root node „Person" and determines all the nouns, that have a ISA (hyponym) relation to the noun. In this particular case, this are several words like "Benutzer", "Weltenbürger", "Lebender", etc. However, the abstractness of words -- as we use it in our research -- does not only consider the child nodes but also the child nodes of the child nodes and so forth. Consequently, the whole subtree underneath one node is determined. This of course could cover several levels until the algorithm terminates in the leafs of the tree.

The abstractness of a term is not determined on the number of levels the corresponding node has to its leafs (depth of inheritance), but is the number of child nodes that are below the node. The idea is that the abstractness is determined in a mathematical sense by counting number of possible meanings, respectively words. Therefore, the more child nodes a node has, the more meanings are subsumed underneath the node and therefore, the more abstract the node actual is.

The example given in Figure 2 shows the noun "Person" with some of its child nodes. The lexical database GermaNet knows 49 direct hyponyms for the noun „Person". This means, that on the n+1 level of abstractness 49 different specialized words for „Person" exist, such as "Benutzer", etc. The tree structure of the lexical database can lead to an exponential growth of the number of words while moving down-wards the hyponym relations. Consequently, the following definition emerges:

Definition Abstractness: *The total number of nouns $\alpha_n$which can be reached by a hyponym relation starting from the noun n will be defined as n's abstractness. Reoccurring hyponyms are only counted once.*

Using the definition above, we are now able to determine the abstractness of "Person" explicitly, objectively and quantifiable. Querying GermaNet returns a total number of $\alpha_{Person} = 12\ 628$ distinct nouns, which can be reached by the noun "Person" solely by using hyponym relation.

6.2. Concreteness

In contrast to the abstractness of a noun, the concreteness determines the specificity of a noun, which is semantically connected to the abstractness. It is possible to derive an objective measurement for the noun's concreteness by using a lexical knowledge database, it. Thereby, again the hypernym relation provided by the lexical databases is used. Figure 2 shows an excerpt of WordNet and GermaNet hypernym relations starting from the noun "Person". The Figure furthermore provides a comparison between WordNet and GermaNet regarding the level of concreteness. Following the noun

“Person“ upwards in the hypernym tree, WordNet returns us the nouns “Organism“, “Living Thing“, “Whole“, “Object“, “Physical Entity“ and the least concrete concept “Entity“.

The inheritance tree on the left side was derived from WordNet and GermaNet for the right inheritance tree. Both lexical databases return the same depth of inheritance for the English and German noun „Person“. The depth of inheritance is seven in both cases. This means that seven steps along the hypernym relation are required to reach the root node “entity” (“Entität”).

Using GermaNet or WordNet as a base line to determine the concreteness of a noun has the advantage, that every noun, that is stored in the database has a hypernym relationship, as long as it is not the root itself. In GermaNet, the root is called “GNRoot”. From this root node downward, the nouns are placed hierarchically using the hypernym relationship. “GNRoot” has several child nodes, such as “Zustand“ (state), „Attribut“ (property), etc. “GNRoot” is the most generic concept since it does not have any hypernyms, obviously it is a fictional concept inserted for technical reasons. Based on our prior investigations of the concreteness of a noun, we are now defining the concreteness as an objective measure that can be used in all lexical knowledge databases:

Definition Concreteness: *The total number of hypernyms β_n that exist between the noun and the root node of the lexical database will be defined as concreteness.*

Problematically, the words stored in WordNet and GermaNet, are organized in synsets, in order to enable polysemy. Language allows several meanings for the same word depending on its usage and its context (polysemy). For example, the word „bank“ can have several meanings. Those meanings are represented in so-called synsets, which contain an entry for all the different meanings a word can have. Each of the meanings can have a different concreteness. It might be the case, that the financial institution „bank“ has more hypernyms until the root node is reached than other meanings of the same word. As a possible workaround, we determine the average of all different concreteness measures starting from a particular noun:

Definition Average Concreteness: *The average concreteness of a noun is the average length $\bar{\beta}_n$ of all possible hypernym paths starting from a given noun n.*

To illustrate the difference, we will have another look at the example in Figure 2. Using GermaNet, $\beta_{Person} = 7$ but using the same dataset $\bar{\beta}_n = 7.5$. Based on our definition, this means that the average length of all paths from the noun „Person“ to the root node „Entity“, considering polysemy, consists of 7.5 vertices. Taking into account the polysemy is necessary in order to do not make systemic errors. Furthermore, if an analysis is done on a sufficiently large dataset the error becomes very small.

6.3. Generality as the Synthesis of Concreteness and Abstractness

The determination for the abstractness and the concreteness of a noun in legal texts, using lexical knowledge, are two diverging approaches. Both measurements consider different aspects, namely generality and specificity as a linguistic phenomena. However, in order to fully understand a terms generality, respectively specificity, both measurements have to be considered simultaneously. An integrative indicator, combining both values is the synthesis of two opposing and intrinsic properties of a linguistic term, which can be defined as follows:

Definition Generality: *The generality γ_n of a noun is the noun's abstractness α_n divided by it's average concreteness $\bar{\beta}_n$.*

$$\gamma_n = \frac{\alpha_n}{\bar{\beta}_n}$$

Consequently, the generality for „Person“, as used in the prior Sections, is as follows:

$$\gamma_{Person} = \frac{\alpha_{Person}}{\bar{\beta}_{Person}} = \frac{12\ 628}{7.5} = 1\ 683.73$$

The generality for the noun “Person” is high, since several thousand hyponyms ($\alpha_{Person} = 12\ 628$) are stored in GermaNet, whereas the average concreteness is $\bar{\beta}_{Person} = 7.5$. Analyzing a second example, the German word for “law”, i.e. “Gesetz”, the generality is quite different.

$$\gamma_{Gesetz} = \frac{\alpha_{Gesetz}}{\bar{\beta}_{Gesetz}} = \frac{131}{6.67} = 19.65$$

The two examples show that the combination of abstractness and average concreteness as defined above, combined in a division gives an overview of the nouns generality. Based on the measurement on the dataset of GermaNet, we continue to automatically analyze the application scope of legal norms and legal texts in general (see Section 7.1.). Thereby, the measurement serves as a heuristic to compare different but related legal texts, such as the German and Austrian definition of products given in the act governing the liability for a defective product (see Section 7.3.).

7. Algorithmic Determination of Abstractness and Concreteness

7.1. German Laws

From the introduced dataset of German laws, we selected the ten acts, containing the most words. Based on this selection, we automatically analyzed 1 018 448 words. From this 1 018 448 words, 221 985 are nouns (21,78%). Considering only distinct nouns, we aggregated those without stemming and finally retrieved the number of distinctive nouns for each law (see Table below).

Law	Distinct Nouns	Recognized norm	Recognized stem.	Recognized brute-f.
AMG	2079	0,48	0,66	0,75
BGB	3399	0,46	0,63	0,71
HGB	2429	0,47	0,65	0,74
KAGB	2097	0,46	0,64	0,72
KredWG	2628	0,42	0,60	0,69
SGB 5	4004	0,39	0,56	0,65
SGB 6	2173	0,45	0,61	0,69
StGB	1898	0,56	0,75	0,83
StPO	2101	0,51	0,69	0,76
ZPO	2464	0,47	0,64	0,72
MEAN	2527	0,47	0,64	0,73
SD	636,69	0,04	0,04	0,05

Table 1: GermaNet noun recognition rate

Table 1 shows the ten German laws with the most words ordered regarding their total word count. The first column gives the name of the corresponding act. Due to lack of additional space, the Table only shows abbreviations. The second column refers to the number of distinct nouns contained in the law. This number represents the overall number of nouns, determined by the Stanford POS tagger

(Toutanova et al.). No aggregation regarding noun stem or any other pre-processing is performed. The next three columns show the recognition rates of the different approaches used to determine whether a noun can be found in GermaNet or not. Thereby, we have implemented three different algorithms. The first approach (3rd column) just considers the noun as it appears in the text and compares it to the nouns provided by GermaNet. No stemming or other pre-processing is done. The second approach (4th column) stems the noun - if it is not found as is - using a porter-stemming algorithm and afterwards the retrieved word stem is again compared to GermaNet. The third „brute-force“ approach (5th column) firstly searches for the word as provided from the text. If it cannot be found, it stems the word and searches again in GermaNet. If the search is again without success, the algorithm takes the word-stem and iterates over all nouns in GermaNet (93 631). In case there is noun in GermaNet, that starts with the word-stem, the algorithm terminates and returns the determined noun. This algorithm is vulnerable to errors, because it would reduce the word „Mitteilungen“ to its stem „Mitteil“ which afterwards matches to „Mitteilungsblatt“, which is of course wrong. Consequently, it rather be used as decision support and proof-of-concept, than as exact result.

The recognition rates shown in column 3-5 of Table 1 give an overview of the nouns used throughout German law texts and their correspondence in GermaNet. The recognition rate 0.46 of the German Civil Code (BGB) represents the fact, that 46% of the nouns as used in the law text are contained in GermaNet. Using a stemming algorithm, a recognition rate of 63% is achieved. Furthermore, the vulnerability to errors is also kept low since it mainly removes lexical post-fixes like, -s, -er, -es, -en, etc. The usage of the brute-force algorithm also leads to a higher recognition rate (71%), but the failure rate also increases (see Section 7.2.). The before mentioned problem remains: making the „starts-with“ criteria sufficient for the assignment of nouns is problematically. The sources of possible errors and the error rate are in detailed discussed in the next Section 7.2.

	Nouns	$\boldsymbol{\alpha}$	$\overline{\boldsymbol{\beta}}$	$\boldsymbol{\gamma}$
Normal	1547	260.50	7.37	63.95
Stemming	2141	264.10	7.32	66.47
Brute-Force	2417	245.55	7.35	61.44

Table 2: Abstractness, average concreteness and generality of the German Civil Code (BGB)

Using the noun as is allows to find 1547 nouns (46%) in GermaNet. Based on every single noun, the average concreteness, abstractness and generality as defined above is calculated. The same procedure was done using the stemming and brute-force method to increase the recognition rate of nouns in GermaNet. Table 2 shows the overview of the average measurements performed. The average concreteness, the distance to the GermaNet root node, does not differ significantly, but is at about 7.3. Comparing the different abstractness rates, the number of hyponyms of a noun, the difference is again not very large. The normal method leads to an average noun abstractness of 260.50, the stemming method give 264.10 and the brute-force method results in 245.55 average noun abstractness. Calculating the standard deviation on the concreteness and abstractness offers a greater insight into the distribution of the determined measurements. The standard deviation of the concreteness ranges from 2.14 (normal) to 2.19 (stemming) which is relatively low. Considering the standard deviation from the abstractness measurements the situation is different. The standard deviation ranges from 1530.24 (brute-force) to 1594.56 (stemming). The detailed investigation of the large standard deviation of the abstractness, 6-7 times as large as the average, was done with histograms (see Figure 3**Fehler! Verweisquelle konnte nicht gefunden werden.**), whereby the focus was set on the distribution of the abstractness and concreteness as defined in Section 6. The distribution of the nouns' generality γ_n does not much differ from a qualitative perspective from the abstractness. Therefore, we omitted the visualization in this paper.

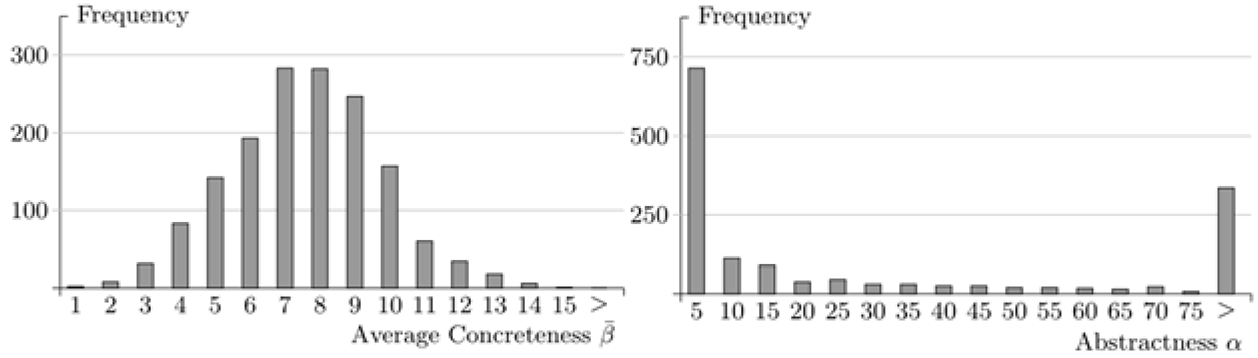


Figure 3: Histogram of nouns' concreteness and abstractness

The distribution of the concreteness of the different nouns in the German Civil Code is shown as histogram in Figure 3. Thereby, the concreteness of those nouns detected using the normal approach, are visualized. Consequently, 1549 different nouns with $\bar{\beta}_n = 7.37$ and standard deviation of 2.14 (see Table 1) were considered. The Figure shows, that the distribution is similar to a Gaussian distribution, although the imaginary bell shaped curve would not fit perfectly on the dataset. The decrease of nouns frequency towards higher concreteness is faster, than to lower concreteness. Mathematically speaking, the distribution has a non-zero and positive skewness (right-tailed). Splitting the measurements into three equal distance clusters, shows, that in the cluster $1 < \bar{\beta} \leq 5$, 17% of all nouns, in the cluster $5 < \bar{\beta} \leq 10$, 75% of all nouns and in the upper third $10 < \bar{\beta}$ 8% of all nouns are contained. Interestingly, the first cluster containing less concrete nouns has two times more nouns than the third cluster.

The situation is different if we analyze the distribution of the nouns' abstractness. We already observed that the average abstractness is 260.50 with a standard deviation of 1548.80. The histogram of the nouns' abstractness measurements is also shown in Figure 3. At a first glance, one can already see the distribution is completely different. The analysis of the abstractness measurements offers, that a noun used in the German Civil Code are either very abstract ($\alpha_n > 75$) or are the quite opposite ($\alpha_n \leq 5$). This also explains the high standard deviation that was measured. Obviously, the calculated mean is not an adequate representative for the determined nouns' abstractness since it does not represent the information about distribution.

7.2. Evaluation

Since NLP is error prone and we also faced some drawbacks using algorithms to automatically determine words and their POS. We identified three different error sources for possible errors during the overall processing the German law texts:

1. NLP techniques: The POS tagger does not always determine right and comprehend results. Some words are determined to be nouns, although they are something different, like adjectives or verbs. The misspelling and orthographic errors of the text can be neglected, since law texts are mostly free from those errors due to their high textual quality.
2. Pre-processing techniques: Due to the lack of processing the determined noun, it cannot be found in GermaNet. In some cases, it is not possible to just stem the word, because the stemming does not always deliver correct and useful results. The introduced brute-force method to boost the recognition rate increases the error rate so that it unusable.
3. Incompleteness of GermaNet: Some nouns used in the law are not contained in the GermaNet since there first usage ever is the law text. The German legislator acts as the creator of new and artificial words like „Leibrentenversprechen“, „Zahlungsauthentifizierungsinstrument“, or „Verfahrensbeteiligter“. Hence, the lexical database lacks of a law specific vocabulary.

To evaluate the usage of NLP technologies, we performed a manual evaluation of the retrieved nouns using the Stanford log-linear POS Tagger. Table 3 summarizes the result of the manual evaluation part of the German Civil Code (BGB). The Stanford POS tagger determined 3399 distinct nouns, from which we checked 1000 randomly selected nouns. The result is, that out of the selection of 1000 different nouns 982 words are nouns, whereas only 18 are no nouns and tagged wrongly. This leads to a precision rate of 98.20%, which is quite high.

We also analysed the recall of the Stanford POS tagger manually. Therefore, we randomly selected 200 nouns from the law text and looked them up in the list of nouns that were determined. The resulting recall was quite surprising: 78.50%. Many words could not be determined as nouns by the POS tagger. From the arbitrary selection of 200 nouns, the algorithm also recognized only 157. One possible explanation of this phenomena is, that the vocabulary contains nouns that are not used in common language like complex composite words, e.g. „Leibrentenversprechen“, etc. A further explanation would be that the German Civil Code, created 1896 and promulgated 1900, uses a vocabulary, which nowadays outdated in some cases. Therefore, current training models for POS tagger might not be able to recognize all those words.

	Amount	**Percentage**
Precision	982 out of 1000	98.20%
Recall	157 out of 200	78.5%

Table 3: Precision and recall for noun recognition in the German Civil Code

We also analyzed the nouns in the German Civil Code, that could not be recognized by GermaNet in a first step and on which either the stemming or the brute-force method leads to a successful identification. The following Table 4 summarizes the precision rates, which were checked manually.

	Amount		**False Positives**	
Normal	1 548	45.54%	0	0.00%
Stemming	592	17.42%	19	3.21%
Brute-Force	276	8.12%	153	55.43%
Not found	983	28.92%	-	-

Table 4: Recognition error rates for the Civil Code

The table above shows the error rates we manually detected after the matching of nouns and GermaNet. Of the 3399 nouns determined using the normal approach of processing, 1548 (45.54%) could be found in GermaNet. Additional pre-processing like stemming and the mentioned brute-force searching increased the number of matched nouns by 17.42%, respectively 8.12%. Nevertheless, 983 (28.92%) of the determined nouns could not be found in GermaNet. On the other hand, pre-processing also increases the vulnerability to errors. Out of the 592 stemmed nouns, 19 were wrongly determined (3.21%). The error rate using the brute-force method was very high. Out of the 276 nouns, 153 (55.43%) were wrongly classified. Obviously, the usage of pre-processing can really boost the recognition rate, but has to be used with care, because it dramatically increases the error rates.

7.3. Act Governing the Liability for a Defective Product: Germany vs. Austria

Due to the political situation, the European Union has an impact on the national legislation, resulting in Council Directives that have to be adopted and promulgated by its member states. A common example is the Council Directive 85/374/EEC which governs the liability for defective products. In the years after its entrance into force, the German and the Austrian legislation also promulgated their versions of the corresponding act. Below, both articles are given:

Austria ProdHaftG §4: *Produkt ist jede bewegliche körperliche Sache, auch wenn sie ein Teil einer anderen beweglichen Sache oder mit einer unbeweglichen Sache verbunden worden ist, einschließlich Energie.*

Germany ProdHaftG §2: *Produkt im Sinne dieses Gesetzes ist jede bewegliche Sache, auch wenn sie einen Teil einer anderen beweglichen Sache oder einer unbeweglichen Sache bildet, sowie Elektrizität.*

Based on this selection of two different but related norms, we did an analysis regarding the concreteness, abstractness and generality of the nouns.

Germany	**Austria**	$\boldsymbol{\alpha}$	$\overline{\boldsymbol{\beta}}$	$\boldsymbol{\gamma}$
Produkt		4224	5.00	844.80
Sache		23573	3.50	6735.14
Teil		11506	4.20	2739.52
	Energie	118	6.50	18.15
Elektrizität		58	8.00	7.25

Table 5: Comparison of the acts' nouns

Table 5 lists the nouns of the Austrian and Germany act governing the liability for a defective product. Both acts share three different nouns, which appear in both acts, namely "Produkt" (product), "Sache" (thing), and „Teil" (part). Additionally, the table provides information about the average concreteness $\overline{\beta}$, abstractness α and generality γ. Interestingly, the two acts differ regarding their application scope. Whereas the Austrian act also includes energy (Energie) as a product, the German act only electricity (Elektrizität). This difference in the application scope can automatically be determined using lexical knowledge. Table 5 holds both nouns and their respective information from GermaNet. As one can clearly see, „Energie" has an abstractness $\alpha_{Energie} = 118$, which means that GermaNet knows 118 different nouns that are specific forms of energy. The electricity as used in the German act, has an abstractness $\alpha_{Elektrizität} = 58$. Furthermore, using lexical knowledge it is also possible to determine the hypernym relationship between both nouns.

The ISA relationship between energy and electricity is mapped in GermaNet. Consequently, the usage of this lexical knowledge allows the subsumption in a restricted way, namely the subsumption according to the words sense (see Section 2). The ISA relationship between "Elektrizität" and "Energie" cannot be determined by solely looking at the corresponding α, β, $\overline{\beta}$, or γ values. This information is a relation between two separate words, i.e. nouns. This information is stored in GermaNet. Nouns, that are subsumed as energy but not as electricity are concepts like "Primärenergie" (engl. primary energy), "Wärmeenergie" (engl. thermal energy), or "Arbeit" (engl. work). The ISA relationship is transitive; consequently, every noun that is a hyponym to electricity is also a (inherited) hyponym of energy. This inheritance and transitivity is a basic principle and enables the subsumption in the sense of word meaning.

8. Challenges and Limitations

As we have shown above, lexical knowledge can be used to support the comparison of laws regarding the application scope of legal norms and acts. The lexical knowledge databases thereby serve as information provider making the implicit semantic relationships between words of a language explicit and accessible to algorithms. This Section summarizes the challenges arising during the processing of texts, accessing a lexical knowledge and objectively measuring concreteness, abstractness and generality.

Lexical knowledge: GermaNet. Although the usage of lexical knowledge allows an extensive analysis of semantic relationships between words, some drawbacks remain. Firstly, GermaNet lacks of comprehensiveness especially with regard to the vocabulary used in the legal domain.. Secondly, the determination of concepts, represented in language as bigrams, such as „natural phenomena“ or „living thing“, harden the problem of determination the semantics. At last, the problem of polysemy detection, as already observed by Gangemi et al. (Gangemi et al. 2002) exists. If a noun has several meanings, such as the noun „bank“ it is unclear, which word sense is the right one.

Natural language processing (NLP). The algorithmic processing of natural language is known to be challenging but promising. Especially in the research domain of legal texts and legal informatics, the usage of NLP technologies is common. Legal texts are usually well written and without orthographical or grammatical mistakes. This positively contributes to the precision rates. Nevertheless, the complex sentence structures and word compositions introduced by the German legislation are major drawbacks.

Domain specificity of legal practice and theory. The applicability of automatically processed legal texts and acts depend on the intended usage. Thereby, the field of application determines the requirements and use cases that decide about the usefulness of information. This variation also effects the words and their meanings. This complex, and to a certain extend social phenomena, challenges the automated determination and usage of a words meaning even more.

We briefly sketched the limitations of the usage of automatically derived information from legal texts as well as their combination with existing lexical real-world knowledge. The paper now proceeds with a conclusion, summarizing the papers' contribution.

9. Conclusion and Outlook

Our paper is a contribution to the investigation of the application scope of legal texts, based on the subsumption principle, which is a complex and well-studied field in legal theory. We restricted the subsumption process on its word-sense, i.e. grammatical, driven process. We analyzed German law texts regarding their nouns and proposed a theory regarding concreteness, abstractness and generality of nouns, using a lexical knowledge database, namely GermaNet, the German pendant to WordNet.

We exhaustively analyzed the nouns of the German Civil Code. Based on our analysis we can draw conclusion regarding the content and the used method. Whereas the concreteness of nouns used in the law text follows a bell-shaped distribution, the abstractness behaves opposite. Nouns are either very abstract $\alpha > 75$ or it not very abstract $\alpha < 5$, but it's unlikely to be in between. The processing of natural language has some drawbacks, contributing to the recognition, precision and recall, which we measured and discussed in detail. Our approach also allows for the comparison of application scopes of different acts and norms. We exemplary showed this on the German and the Austrian version of the act governing the liability for a defective product, with the algorithmically reproducible result, that the Austrian act defines a more abstract ($\alpha_{Energie} > \alpha_{Elektrizität}$) application scope for products than the German version.

These measurements give insights into the structure and usage of words, especially nouns, in the domain of law texts. During the subsumption, this could serve as a base line heuristic for decision support, but also for automated comparison the application scope of acts and norms.

Acknowledgement

This research was sponsored in part by the German Federal Ministry of Education and Research (BMBF) (project “Software Campus (TU München)”, grant no. 01IS12057).

10.Publication bibliography

Bench-Capon, Trevor; Araszkiewicz, Michał; Ashley, Kevin; Atkinson, Katie; Bex, Floris; Borges, Filipe et al. (2012): A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. In *Artificial Intelligence and Law* (20), pp. 215-319.

Benz, Dominik; Körner, Christian; Hotho, Andreas; Stumme, Gerd; Strohmaier, Markus (2011): One tag to bind them all: Measuring term abstractness in social metadata. In : The Semantic Web, pp. 360–374.

BMJ (2014): Juris. Gesetze im Internet. Available online at http://www.gesetze-im-internet.de/, updated on 7/22/2014, checked on 7/22/2014.

Brown, R. (1958): Words and things: The Free Press.

Gangemi, Aldo; Guarino, Nicola; Masolo, Claudio; Oltramari, Alessandro; Schneider, Luc (2002): Sweetening Ontologies with DOLCE. In *Knowledge Engineering and Knowledge Management*, pp. 166-181. DOI: 10.1007/3-540-45810-7_18.

George, A. Miller (1995): WordNet: a lexical database for English. In *Commun. ACM* 38 (11). DOI: 10.1145/219717.219748.

Hamp, Birgit; Feldweg, Helmut (1997): GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Hutchison, David; Kanade, Takeo; Kittler, Josef; Kleinberg, Jon M.; Mattern, Friedemann; Mitchell, John C. et al. (Eds.) (2005): Law and the Semantic Web. Berlin, Heidelberg: Springer Berlin Heidelberg.

Köhler, Reinhard (2005): Quantitative Linguistik. Berlin [u.a.]: De Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

Larenz, Karl; Canaris, Claus-Wilhelm (1995): Methodenlehre der Rechtswissenschafft. Berlin [u.a.]: Springer.

McNamara, Danielle S.; Graesser, Arthur C.; McCarthy, Philip M.; Cai, Zhiqiang (2014): Automated evaluation of text and discourse with Coh-Metrix: Cambridge University Press.

Mellinkoff, D. (2004): The language of the law: Resource Publications.

Raabe, Oliver; Wacker, Richard; Oberle, Daniel; Baumann, Christian; Funk, Christian (2012): Recht ex machina. Berlin, Heidelberg: Springer Berlin Heidelberg.

Sartor, Giovanni; Casanovas, Pompeu; Biasiotti, Mariangela; Fernandez-B., Meritxellx (2013): Approaches to Legal Ontologies: Theories, Domains, Methodologies: Springer Publishing Company, Incorporated.

Schendera, Christian F. G. (2004): Die Verständlichkeit von Rechtstexten. In Kent D. Lerch (Ed.): Die Sprache des Rechts: Recht verstehen. Berlin, New York: De Gruyter, pp. 321–373.

Tiscornia, Daniela (Ed.) (2006): The LOIS project: Lexical ontologies for legal information sharing.

Toutanova, Kristina; Klein, Dan; Manning, Christopher D.; Singer, Yoram: Feature-rich part-of-speech tagging with a cyclic dependency network. In : Proceedings of the HLT-NAACL Conference 2003, pp. 173–180.