Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Detecting Ambiguity in Statutory Texts

Tom Zeller

Bachelor thesis

Studiengang:	Informatik
Prüfer:	Prof. Dr. Uwe Reyle
Betreuer:	Prof. Dr. Uwe Reyle

Beginn der Arbeit:	09.03.2018
Ende der Arbeit:	09.09.2018

Contents

1 Introduction

Ambiguity is ever-present in natural language production. A human typically has no difficulties in selecting the right interpretation for an ambiguous expression by using lexical and pragmatic knowledge. While the inclusion of broad semantic knowledge poses a challenge for general disambiguation systems and parsers, its utilization might be a feasible approach for disambiguation in a restricted context. A domain that is very sensitive to ambiguity is the legal domain, especially in the wording of statutory text. Some parsing systems deal with ambiguous input by specifying all possible interpretations without explicitly choosing a solution or by returning multiple parses along with their respective probability. This work serves two purposes: An application is created which allows the input of statutory texts or single text excerpts and which detects included structural ambiguities in the form of prepositional phrase attachments and coordination ambiguities, and semantic ambiguity in the form of scopal ambiguity. Furthermore, the found ambiguities are filtered by including subcategorizational information and by utilizing domain-specific semantic knowledge which is encoded in the form of a legal domain ontology and selectional preferences for common legal expressions. The filtering capability and the effect of including the semantic knowledge are evaluated on the *DUBLIN3 Regulation.*

2 Ambiguity and Vagueness

2.1 The Concept of Linguistic Ambiguity

Ambiguity can be described as the property of a *sign* to be related to more than one meaning. A *sign* might be any physical or abstract object that can be interpreted, like a picture, but most often, the term is used to refer to some kind of linguistic expression. Consider the sentence

The man fed her cat food.

This sentence can be interpreted as a man feeding a woman's cat some food or alternatively, as a man feeding a woman food made for cats.

Ambiguity is a phenomenon which can be found in many domains, literature, visual art, social interaction, but this work will be focusing on linguistic ambiguity in the restricted domain of legal texts. With differing backgrounds, the perceived role of ambiguity varies: In the literary domain, which is a mode of communication special to what Bauer et al. (2010) calls *normal communication*, expressions conveying multiple meanings are often intended, to invoke contrasting associations and to move liberty of interpretation to the reader. In other domains, the range of possible interpretations shall be as restricted as possible. Obvious examples would be the field of requirements engineering, a discipline which aims to provide exact specifications of the properties of a product, or maths, where universally equal interpretation of signs forms a necessity to be able to operate on common ground. In general, the tolerance for ambiguity is low in all domains that use a formal language or subsets of natural language, as this is the case for programming languages or for *Controlled Natural Language* (*CNL*). As Wasow et al. (2005) emphasize, in these formal or restricted languages, a sign denotes some kind of information, and to communicate this information to another human or to a machine, the sign must be correctly assigned to its corresponding denotation. (Wasow et al. (2005), 271) A legal text is expressed in unrestricted, natural language, but the same conditions on its definiteness as for formal or controlled language apply. Requirements that are special to the domain of legal texts will be discussed in chapter *3*. If ambiguity is regarded from a purely linguistic perspective, a common distinction is made between *lexical*, *syntactic* and *semantic* ambiguity, whose characteristics will be explained in the following.

2.2 Lexical Ambiguity

This type of ambiguity can be characterized as a lexem denoting two or more different meanings. An often used example is *bank*, which can refer either to the financial institution or to a riverside. This property is called *homonymy*, which is attributed to words having the same lexem, but differing **and** unrelated meanings. Related to this is the concept of *polysemy*, which is the property of a word to refer to a palette of related meanings. Contrary to *homonymy*, which is conicidential, the contained meanings of a *polyseme* often stem from a common etymological origin. The above mentioned example of *bank* is a polyseme too, given its reading as *the building where a bank (financial institution) is located.* So *bank* is polysemous in its meanings *financial institution* and *location of said institution* and homonymous in its meanings *financial institution* and *riverside.* It is up to debate, if polysemy is distinct from lexical ambiguity, or if it is a subset of it (Wilson, 2001).

2.3 Syntactic Ambiguity

Syntactic ambiguity (often called *structural ambiguity*) occurs on the sentence level. A sentence is syntactically ambiguous if two or more possible sentence structures can be assigned to it. The classic example for a structurally ambiguous sentence is

I saw the man with the telescope.

whose ambiguity is present in the german translation as well:

Ich sah den Mann mit dem Fernglas.

The two constituent structures that can be inferred from the sentence are illustrated below.

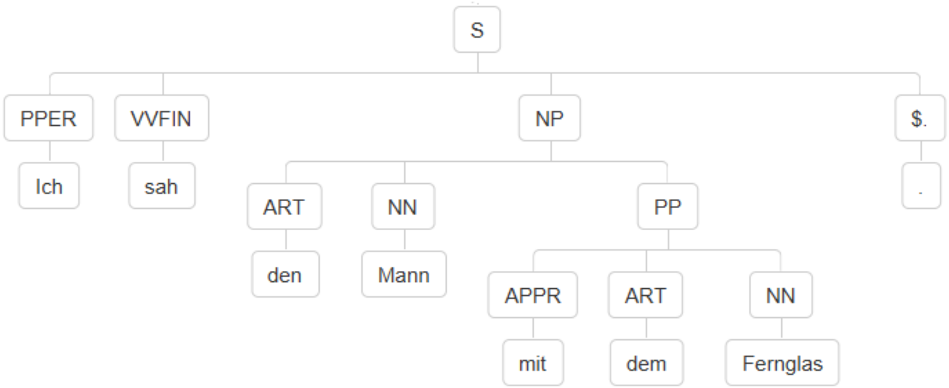


Figure 1: Attachment to *Mann*

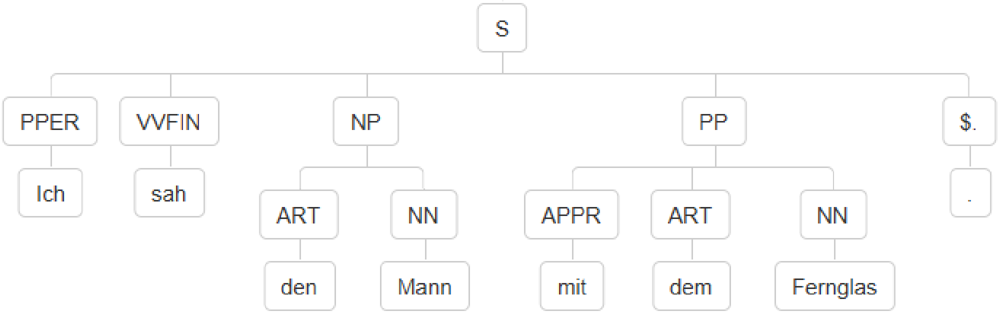


Figure 2: Attachment to *sah*

Figure 1 and 2 illustrate the most common form of syntactic ambiguity which concerns the attachment of prepositional phrases and which will be introduced in detail in the next subchapter. By definition, syntactic ambiguity cannot occur on single words like lexical ambiguity, it is however not restricted to whole sentences. An example for local syntactic ambiguity is illustrated by the english sentence

The old man the boat.

This sentence is globally unambiguous, which means there is only one sentence structure we can assign to it. However, when processing the sentence, a human typically assumes *man* to be part of a nominal phrase *the old man*, although it is actually a finite verb. In psycholinguistics, sentences that induce this effect of interpreting them incorrectly first, and reach the correct interpretation with more context (having advanced further in the sentence) are called *garden path sentences.*

This type of structural ambiguity is very apparent though, and it can be safely assumed, that human proof readers of any legal document will notice them. Ambiguities resulting from multiple possible attachments as illustrated by the former two examples are much more widespread, and of greater importance for this work.

2.3.1 PP Attachment

One of the most prevalent forms of structural ambiguity concerns the attachment of prepositional phrases. Niles and Pease (2003) report that from a set of 710 prepositional phrases, 501 phrases (70,7%) could not be identifed by a human to clearly either belong to a noun or a verb. However, humans often have little difficulty in assigning the correct structure to a sentence. If we reconsider the example from above,

Ich sehe den Mann mit dem Fernglas.

pragmatic knowledge tells us that *mit dem Fernglas* is more likely to be attached to *sehen* than to *Mann.* This kind of knowledge needed for disambiguation poses a problem for Natural Language Processing and will be further discussed in chapter 4. As Mehl et al. (1998) point out, in German, the problem of attachment gets even more complicated through its comparably free word order. A PP may not only be attached to its direct NP predecessor, as the following example shows:

die Unterhaltung der Teilnehmer über mögliche Anbindungen

It may be attached to a noun following it as well:

Er steht mit ihm in Verbindung.

As this kind of ambiguity is so prevalent, its detection and resolution in legal domain texts will be a major focus of the system.

2.3.2 Other Structural Ambiguities

Relative Clause Attachment This kind of ambiguity occurs when it is unclear, to what part of the preceding main clause a relative clause is referring to. An example would be:

Ich sah den Sohn meines Nachbarn, der mein Fernglas hat.

The relative clause *der mein Fernglas hat* can be attached to either *Sohn* or *Nachbarn*. In German, this kind of ambiguity occurs more sparsely than prepositional phrase attachment, due to the fact that the relative clause's grammatical properties are determined by the associated noun. If the neighbour was female, the relative clause attachment would be unambiguous:

Ich sah den Sohn meiner Nachbarin, der mein Fernglas hat.

Coordination Coordination ambiguity is a kind of structural ambiguity that generally arises if a coordinated structure (e.g. CNP or CVP) is used in combination with a modifier. A frequent coordination and modifier combination consists of a coordinated noun phrase along with an adjective or adjective phrase. Expressions like *reife Bananen und Äpfel* can be interpeted as *reife* referring to both, *Bananen* and *Äpfel*, or only to the former. Another possible manifestation is exemplified by the combination of verb phrases and NP arguments, as the following example illustrates:

Ich las und schrieb einige Bücher.

Einige Bücher can be argument to both verbs or only to the latter. Two further possible sources of coordination ambiguity are:

Ich verstand und reagierte sofort.
(Coordinated verb phrase and adverbial modifier)

Er ist nicht geeignet und interessiert.
(Coordinated adjective phrase and adverbial modifier)

2.4 Semantic Ambiguity

A sentence may have an unambiguous syntactic structure, but still offer multiple interpretations. *Mark und Lisa sind verheiratet* can either be interpreted as the two persons being married to each other or as both being married, but to different people. The two most important subclasses of semantic ambiguity, *scopal* and *anaphoric* ambiguity will be introduced in the following.

2.4.1 Scopal Ambiguity

Scope in linguistic terms denotes the area of effect of an expression within the boundary of a sentence. The german negation *nicht* (*not*) is an example of a scoped expression, which demands to determine the part of the sentence which has to be negated. Word classes that exhibit scope are quantifiers like *alle (all)*, *jede- (every)*, *manche- (some)* that are prepended to a NP, negations like *nicht (not)*, or modal adverbs like *vielleicht (maybe)* or modal verbs (*müssen (must)*, *sollen (shall)*).

Pafel (2006) identifies the following sentence structure to be generally scope ambiguous in German (Pafel (2006), 52):

$$[_{\mathrm{DP}_{\mathrm{acc}}} \text{ ein- }] \text{ - } \mathrm{V}_{\mathrm{fin}} \text{ - } [_{\mathrm{DP}_{\mathrm{nom}}} \text{ jed-}] \text{ (- } \mathrm{V}_{\mathrm{nonfin}})$$

An exemplary sentence with this structure would be:

Einen Kandidaten hat jede Partei nominiert.

In this case, the determiner *einen* is an existential quantifier, whereas the determiner *jede* is universal. *Alle Klagen sind nicht gerechtfertigt* is another example, the negation *nicht* can have scope over *alle* or vice versa, resulting in two different interpetations of this sentence. In first-order logic, the first interpretation, where *nicht* has scope over *alle* would be expressed as

$$\exists \text{ x (Klage(x)} \land \neg \text{ gerechtfertigt(x))}$$

while the interpretation of *alle* having scope over *nicht* can be expressed as

$$\forall \text{ x (Klage(x)} \Rightarrow \neg \text{ gerechtfertigt(x))}$$

A relative clause can exhibit scope too, the sentence *die Filme, die jeder gesehen hat* can be interpreted as

die Filme, für die gilt, dass jeder sie gesehen hat

or as

für jede einzelne Person, die Menge an Filmen, die er oder sie gesehen hat.

Proposed strategies to resolve scopal ambiguities include statistical approaches (Andrew and MacCartney, 2004) or resolution based on semantic argument structure (Gambäck and Bos, 1998).

2.4.2 Anaphoric Ambiguity

This class of semantic ambiguity occurs when the antecedent of a referring expression, which can be a personal pronoun or a demonstrative pronoun, is unclear. In the sentence *Mark hatte seinen Freund Tim eingeladen, aber er musste das Treffen leider absagen*, the pronoun *er* may refer to either *Mark* or *Tim*. As for scopal and PP attachment ambiguites, the resolution of anaphoric ambiguity demands pragmatic knowledge and world knowledge.

2.5 Vagueness

Tuggy (1993) defines vagueness as follows:
"*Two or more meanings associated with a given phonological form are united as non-distinguished subcases of a single, more general meaning.*"
Often a predicate is considered vague if bordercases exist, for which it is unclear, whether the predicate applies or not (Wilson (2001), 861). The term *groß* can be considered vague, as it does not define what size boundaries are contained within it. The border between polysemy and vagueness is fuzzy, in general, a vague term is not clearly defined at its 'boundaries' whereas a polsemyic term may refer to two or more well-defined terms, that are related to each other. It might be argued that vagueness applies to most terms in a language, because to not be vague, a term would have to include its exact boundaries of meaning. The question if vagueness poses a problem for legal text is under debate, some researchers argue, that it is not the vagueness in natural language terms which challenges interpretation, but rather the context sensitivity of natural language, which is the basis for assigning concrete meaning to vague terms. (Kompka in Keil and Poscher (2016), 205-224). It shall be emphasized here that the detection of vagueness is out of scope for this work.

3 Requirements on Texts in the Legal Domain

In 1985, a trial was held at the Supreme Court. The defendant had illegally aquired food stamps. The corresponding statute reads: *... whoever knowingly uses, transfers, acquires, alters, or possesses coupons or authorization cards in any manner not authorized by [the statute] or the regulations.* (Liparota, 1985)

This case became famous for the defendant's choice of strategy. The structural ambiguity which is inherent in this sentence made him argue that the court had to prove that he was aware of the illegality of his actions, interpreting the sentence as

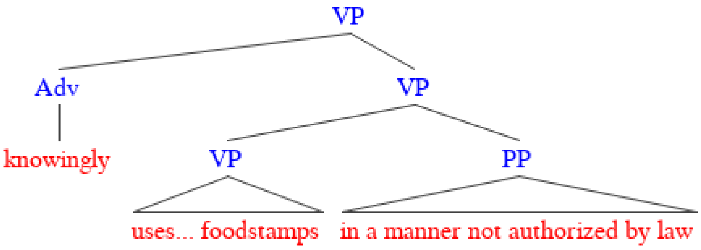


Figure 3: The defendant's interpretation

Whereas the prosecution argued that the statute had to be interpreted as

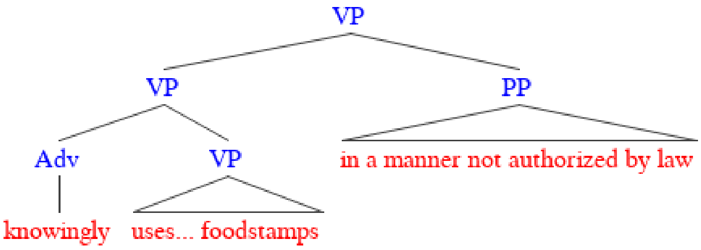


Figure 4: The prosecution's interpretation

Although this example stems from the US justice system, it is no less relevant for the german jurisdiction, as it demonstrates, how the simple structural interpretation of a sentence can be of life-changing importance in an extreme case.

3.1 Interpretation of Statutory Text

The aim of interpreting statutory texts is the mappping of abstract terms to a concrete meaning. For Larenz (2013) the difficulty in this task grounds in the use of natural language which naturally introduces vagueness as opposed to the more formally language used for example in scientifc contexts (Larenz (2013), 312).
To guide the interpetational process, legal methodology has developed a set of criteria, originating in the work of Friedrich Carl von Savigny. These criteria, which Savigny called *canones*, are given below.

Word Sense / Grammatical Interpretation
This category addresses the interpretation using common language meaning, which is grounded in the assumption that the creator of a legal utterance aimed to communicate with his audience in a generally understandable language, and that the probability for an expression to be meant in its common sense is the highest.

Systematic Interpretation
This category emphasizes the importance of the connections between an expression that has to be evaluated and its legal context.

Historical Interpretation
If the grammatical and systematic interpretations fail to eliminate all but one possible interpretation, the historical context has to be taken into consideration. This context is composed of the text creator's normative values

and what is assumed to have been his intention with the creation of the given text. To gain insight into these normative values, available historical evidence like protocols, reports and personal writings have to be examined.

Teleological Interpretation
This category emphasizes the relevance of the objective of a law, thus it is related to the historical interpretation, but differs by abstracting away from the creator, since he or her might not have been aware of the entirety of meaning of his work (Larenz (2013), 333).

As Larenz (2013) emphasizes, these interpretational categories are not mutually exclusive, nor is there a clear hierarchy between them, instead, the jurisdiction should include all of them in their reasoning about a law's meaning. As Reyle pointed out, the grammatical interpretation is of fundamental importance (Reyle (2016), 23), as it sets a frame for the possible interpretations which have to be narrowed down by reasoning in the other categories. In the german jurisdiction, the grammatical category is of special importance. This stems from Germany's special historical context and grounds in the reasoning that the literal text has a democratic legitimation, whereas all other categories are more prone to abusive interpretation.

3.2 Requirements on Statutory Text

As this work will make use of both, german law texts and european law texts in their german translations, two references concerning the textual quality could be determined. The first, the *Joint practical guide of the European Parliament, the Council and the Commission for persons involved in the drafting of European Union legislation* (EU-Kommission, 2003) is a collection of practical, non-binding advice for the creation of european law texts. Reyle (2016) determined two passages that concern the role of ambiguity. The first guidance note reads *The drafting of a legal act must be: clear, easy to understand and unambiguous.* The only clear reference to syntactic ambiguity is found in (EU-Kommission (2003), 5.2.3) which denotes: *The grammatical relationship between the different elements of the sentence must be clear.* The text also emphasizes the general principle of *legal certainity*, which states that "it should be possible to foresee how the law will be applied" (EU-Kommission (2003), 1.2). For the creation of german laws, the equivalent would be the *Handbuch der Rechtsförmlichkeit*, which contains "*general recommendations for the phrasing of legal texts*" (Bundesministerium der Justiz (2008), translation by the author), which states in chapter 1.4: "*Ebenso muss auf die Beziehung der Wörter zueinander und den Sinnzusammenhang geachtet werden.* (*Furthermore, the relationship among the words and the context of meaning should be taken into consideration.* Bundesministerium der Justiz (2008), 1.4, Translation by the author)". As stated in chapter 2.5, the role of vagueness in legal text is under debate, but, as Solan writes in (Keil and Poscher (2016), chapter 10), vagueness is inherent to natural language, and thus is inseparable from statutory language. As this work is focussing on structural and semantic ambiguity, the reader interested in the relation between vagueness and jurisdiction may consider (Keil and Poscher, 2016) for a survey of different positions.

4 Parsing and Ambiguity

Ambiguity in natural language input is a notorious problem for the computational processing of text. To reach human levels of disambiguation facilities, a parser would have to incorporate structural information, pragmatic knowledge, and lexical information (Manning and Schütze (1997), 18). This poses a problem for NLP systems, which have to operate with limited resources compared to humans. In the following, two approaches to handle ambiguity are discussed.

4.1 Statistical Parsing

One approach to handle ambiguous input in constituency parsing are *probabilistic context free grammars* (*PCFGs*). These grammars are constructed by assigning probabilities to each production rule in the grammar. These probabilities are derived from manually annotated treebanks. PCFG grammars can be used by dynamic programming algorithms such as the *CYK* algorithm, to find the highest scoring parse tree for a given sentence. The Earley parser (Earley, 1970) in particular is able to output a *parse forest*, which is a set of parse trees, weighted by their respective probabilities. Outputting multiple parses may either demand to restrict the number of parses to the top k (most likely) parses, or to find an efficient representation of their shared structure, as the number of potential parses for an english or german sentence of average length of 10 to 20 words can already be very high: Church (1982) noted that the number of possible interpretations for a given sentence like *Put the block in the box.*, which produces a single parse tree, grows exponentially (Church (1982), 141). One can increase the number of valid parses by appending another PP, which already introduces syntactic ambiguity in the form of two possible PP attachments:

Put the block in the box on the table.

If another PP is added, the number of interpretations grows to five.

Put the block in the box on the table in the kitchen.

The number of possible interpretations for n attached PPs equates to 2^{n-1}. This property of natural language text to produce a large number of syntactically valid parses is called *combinatorial explosion paradox* (Poesio (1994), 1).

4.2 Preserving Ambiguity through Underspecification

Underspecification in parsing aims to avoid uncertain decisions by including in its output any ambiguities it could not solve given its available information. Schiehlen (2003) developed a rule-based dependency parser called

FsPar, which aims to preserve ambiguous dependencies in its output, in the form of context variables. The parser operates by chunking the text into simple clauses, in a cascaded manner, while the clauses from each previous iteration are used as input for the next iteration. Schielen states the main motivation for this approach as keeping the speed of a deterministic parser, while keeping structural ambiguities for later disambiguation by 'expert modules'[1] (Schiehlen (2003), 166).

5 Subcategorization

Subcategorization frames are a way of formally describing the argument structure of predicates[2]. The verb *trinken* for example has the frames

$$\begin{array}{l}[\mathrm{NP}_{\mathrm{nom}}\ \_\ ]\\ [\mathrm{NP}_{\mathrm{nom}}\ \_\ \mathrm{NP}_{\mathrm{akk}}]\\ [\mathrm{NP}_{\mathrm{nom}}\ \_\ (\mathrm{PP\ aus}\ (\mathrm{NP}_{\mathrm{dat}}))]\end{array}$$

This means that *trinken* can be either used without a direct object, as in *ich trinke*, with a direct object as in *ich trinke das Wasser* or with a prepositional phrase starting with *aus* containing a NP. This example is obviously not exhaustive, as there are other possible PPs like *von* (NP_{dat}). Subcategorizational lexicons can be either manually constructed or automatically extracted from corpora. The latter approach typically involves some kind of measure of co-occurrence, which means, the strength of association between a verb and an argument or adjunct is represented by the relation

$$\frac{\#\ \text{of occurrences of verb - argument}}{\#\ \text{of occurrences of verb}}$$

[1] This work may be considered as such an expert module for the legal domain.

[2] Subcategorization is not restricted to verbs, for example a prepositional phrase starting with *in* would require a nominal phrase in dative, as in *im Himmel*.

The first automatically derived subcategorization lexicon for the german language was built by Schulte im Walde (2002), who used a lexicalized PCFG along with a hand-written grammar to infer subcategorization counts from a corpus of newspaper text. Roberts et al. (2014) presented a similar system, specifically designed for free word order languages like german, which was used to parse a subset of 3 million sentences from the *SdeWaC* corpus, which is in turn a subset of the *deWaC* corpus (consisting of 10^9 parsed web searches). Subcategorization frames may differ in their detail, they may distinguish between arguments and adjuncts, or specify a different number of frame types (Schulte im Walde (2009), 3).

6 Selectional Restrictions

Similar to the concept of subcategorization, *selectional restrictions* are restrictions that verbs impose on their arguments or adjuncts. The following example illustrates this behaviour:

Der Zirkusartist isst einen Salat.
(*The circus artist is eating a salad.*)

The ambitransitive verb *essen* is most commonly used with an argument that could be denoted as belonging to the semantic class of *edible* or *food.* Selectional restrictions can be seen complementary to the notion of subcategorization, where the latter imposes a syntactical restriction, selectional restrictions limit the range of possible arguments semantically. These constraints are not to be regarded as equal to the membership in a predefined class of words. Resnik (1993) describes a statistical approach that models selectional restrictions as probability distributions over a taxonomy of word classes. In this regard, the term *selectional restriction* is often referred to as *selectional preference.*

If we reconsider the example given above, the sentence

Der Zirkusartist isst Glas.
(The circus artist is eating glass.)

is sensical, but the argument to *essen* does not fit in the semantic class of *edible* or *food*. Similar to the acquisition of subcategorization frames, selectional preferences can be derived from large corpora by measuring the association strength between a predicate and a semantic class in terms of co-occurrence. This association can be expressed as a function

$$\sigma : (v, r, c) \mapsto a$$

which maps a verb v, a role r (subject, object, <preposition>) and a class c to a scalar a which denotes the strength of association. Approaches in inducing this association strength differ, used are statistical significance measures, *Hidden Markov Models* (*HMM*s), or bayesian reasoning. (Light and Greiff, 2002)

7 Ontologies as a Representation of Domain Knowledge

Knowledge about the world we are living in assists us in resolving ambiguity. In the first chapter, an example for lexical ambiguity was given in the form of the homonymous word *bank*. If we consider the sentence

I brought all my money to the bank.

it is evident to any recipient of this sentence, that *bank* does not denote *riverside* in this context, but rather the *financial institute*. World knowledge tells us that money is brought to a bank with a much higher possibility than it is thrown into a river.

NLP systems however typically don't have access to this knowledge. It has to be encoded in a form that is accessible and understandable to a computer. The domain of *ontology engineering* tries to solve this problem by representing knowledge in a formal, standardized way. Stemming from greek philosophy, the concept of an ontology has been used by various philosphers with sometimes differing meaning. The general notion of the discipline of ontology is the study of the nature and structure of reality. (Staab et al. (2010), 1) This subdiscipline of metaphysics is concerned with the question of what exists and what relations can be drawn between what is existing.

7.1 Ontology Engineering

Ontology Engineering is a subfield of *information engineering*, and is concerned with the practical development of ontologies. Ontologies have been put to use in information management systems, biomedical assistance systems (Gruber, 1995), for data processing in bioinformatics (Staab et al. (2010), 735) or for recommender systems (Staab et al. (2010), 779). While in the philosphical field of ontology, the question of the existance of categories plays an important role, in ontology engineering, this question is answered from a more pragmatic point of view: If something can be represented, it exists. (Gruber (1995), 1) Thomas Gruber gave a widely used definition of ontologies in the field of information engineering: "An ontology is an explicit specification of a conceptualization". (Gruber (1995), 1) This conceptualization can be defined as tuple $(D, \mathbf{R})$ with D being the *universe of discourse* and $\boldsymbol{R}$ being a set of relations on D. The universe of discourse is the set of "*objects, concepts and other entities that are assumed to exist in some area of interest*" (Gruber (1995), 1). The three fundamental components of formal ontologies are *classes*, *relations* and *individuals*.[3] A class corresponds to a category or collection of entities while individuals are concrete objects, persons, or abstract objects like words. The relations in an ontology often take the form of taxonomical relations like the *is-a* relation or a mereological relation like

[3] The terminology may vary, classes are sometimes called terms and individuals may be denoted as instances.

part-of, but are not restricted to these. Relations may be defined between individuals, classes, or between classes and individuals.

7.2 Top-Level and Domain Ontologies

A *domain ontology* aims to be a representation of the entities and relations of a specific field, which means its *universe of discourse* is restricted. The field of ontology engineering is usually concerned with the creation of domain ontologies. *Top-level* or *upper* ontologies aim to provide a basic foundation for the creation of domain ontologies, by defining general concepts that are common to all specific domains (Hoehndorf, 2010). Often employed concepts are *time*, *space*, *objects* or *processes.* Several implementations for top-level ontologies exist, like the *Basic Formal Ontology* (*BFO*), the *General Formal Ontology* (*GFO*) or the *Descriptive Ontology for Linguistic and Cognitive Engineering* (*DOLCE*).

8 WordNet as a Lexical Database

In 1985, Miller et. al. began the development of a lexical database for english, called *WordNet.* This database can be seen as an enhanced thesaurus, which contains nouns, verbs, adjectives, and adverbs, originally restricted to english (Miller et al. (1990), 2). Entries are grouped into collections called *synsets.* A *synset* is a group of words of the same word class, whose entries can be seen as synonyms of each other. It therefore expresses a lexical commonality. The homonymous word *Bank* can be assigned to at least two synsets, one corresponding to the financial institute, along with words like *Geldhaus*, or *Finanzinstitution* and one referring to the seating-accommodation along with *Sitzbank.* These different conceptual groups a word can belong to are identified as *senses* in WordNet terminology, the word *Bank* would include at least two senses for the two synsets illustrated above. WordNet defines five relations between synsets:

Synonymy* and *antonymy are the two *lexical* relations that WordNet includes, the former being of a central importance, because it is, as stated above, the relation responsible for grouping into synsets. *Synonymy* between two expressions can be defined as one expression being interchangeable with the other in a given textual context, without changing the truth value of the proposition (Miller et al. (1990), 241). The other lexical relation, *antonymy*, is used to relate adjectives and adverbs. In WordNet, antonymy does not connect synsets, which would be a semantic relation, rather it is restricted to single expressions.

Hypernymy* and *hyponymy connect synsets that can be seen as forming a superset-subset connection. It is therefore a semantic relation. The noun *Gebäude* is a hypernym of *Bürogebäude*. Hypernymy and hyponymy can thus be seen as building a *taxonomy* of synsets.

Meronymy is the last semantic relation and expresses a part-of relationship between two synsets. The synset containing *Raum* and *Zimmer* is a meronym of the synset containing *Gebäude* and *Bauwerk*.

8.1 WordNet as an Ontology

WordNet shares many features of an ontology: The *hypernymy* and *hyponymy* relations can be seen as *is-a / parent-of* subclassing resp. superclassing relations, and *meronymy* as *part-of* is a fundamental structuring relation in most ontologies as well. However, *WordNet* does not distinguish between *hyponyms* that would be subclasses in ontological terms and *hyponyms* that would be instantiations or *individuals* of a class.

9 A System for the Detection and Filtering of Ambiguities in Statutory Texts

9.1 Overview

This chapter describes the approach used in this thesis to detect and filter PP attachment, coordination and scopal ambiguities in german statutory texts. The results will be discussed in the next chapter. The detection step will apply to both, prepositional phrase attachments and possible scopal / coordination ambiguities, however only the former will be filtered.

The procedure of the working system can be divided into the following steps:

1. Sentence extraction
2. PCFG parsing of the extracted sentences
3. Pattern matching on generated parse trees for PP attachments
4. Pattern matching on parsed sentences for possible scopal ambiguities
5. Pattern matching on parsed sentences for possible coordination ambiguities
6. For each potential PP ambiguity:
   (a) Query subcategorizational lexical resource
   (b) Query selectional preferences
   (c) Query for ontological connections
7. Give a graphical report of the processing results

In the following, the steps and their underlying implementations are presented.

9.2 Parsing

The sentence extraction and parsing step are performed by the Stanford *CoreNLP* pipeline (Manning et al., 2014). This framework includes a PCFG parser trained on the *NEGRA* corpus, using the *STTS* tagset (Rafferty and Manning, 2008). The advantage of this framework lies in the inclusion of various preprocessing tools, such as a tokenizer and a module for sentence extraction. A drawback is the absence of a parser that allows for explicit underspecification of attachments, which means, these have to be extracted from the generated parse trees.

Obtaining ambiguous PP attachments The approach used to infer the possible ambiguities with the output of the parser can be described as follows: The PCFG parser assigns each parse a score indicating the probability it assigns to the given result. After a sentence is processed, the ordered list of the 50 most probable parse trees according to their score is retrieved. The count of 50 was chosen for practical reasons, for a smaller set of parse trees, the variation of possible attachments was found to be too low, and for higher numbers, the performance of the system drops. The obtained list is iterated, checking each tree for a set of predefined patterns of POS tags, that signal PP attachment. This will be illustrated by the parse for the sentence

Eine solche Formel sollte auf objektiven und für die Mitgliedstaaten und die Betroffenen gerechten Kriterien basieren.

The tree to which the parser attaches the highest score, contains the following subtree:

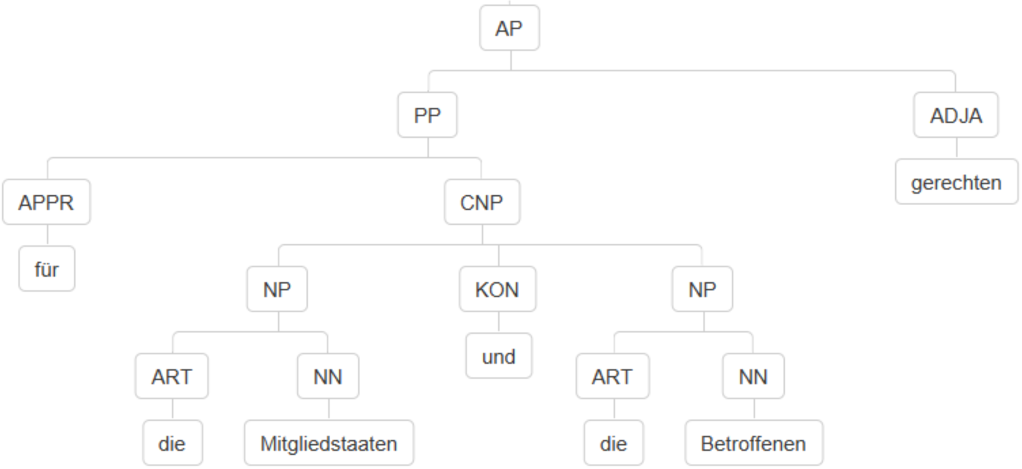


Figure 5: A subtree of the most probable parse

This subtree contains a prepositional phrase which is part of an adjective phrase (AP) whose head is the adjective *gerechten*. The pattern to detect this attachment includes the head of the common phrase (AP) and the sequence of its child constituents: *PARENT: AP, SEQUENCE: PP ADJA*

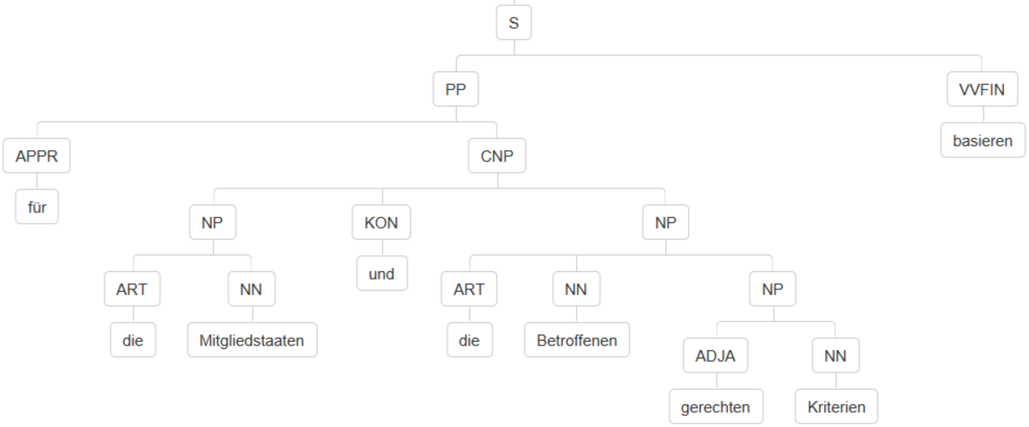


Figure 6: A subtree of the second-most probable parse

The second-most probable parse interprets the PP to be part of a clause

together with the finite verb *basieren*. The corresponding pattern would be
PARENT: S, SEQUENCE: PP VFIN

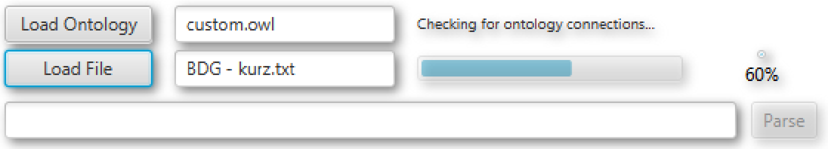
If two or more parse trees are found, that differ in the attachment of a given PP phrase, it is considered ambiguous. An alternative approach would be to utilize an underspecified dependency parser, like the *FsPar* (Schiehlen, 2003), which preserves the possible attachments as features in its output.

9.3 Subcategorizational Lexicon

For the disambiguation using subcategorizational information, a lexical resource was used. This resource was obtained by Roberts et al. through parsing the *SdeWaC* corpus (Roberts et al., 2014). Each entry in the generated file contains, amongst other contextual information, the lemmatized form of the verb, the extracted subcategorizational frame, the information whether the complement is a prepositional phrase along with the head of the complement phrase. This means that if a verb has multiple possible argument structures, it is likely that the resource contains multiple entries with the given verb lemma. This list was processed into the form of a map, that contains the lemma as key, and the found subcategorised prepositions as values, which is queried for every found ambiguous attachment. For the determination of possible attachments, it is of special importance to keep a the number of false negatives (incorrectly ruled out attachments) low. The usage of statistically built subcategorization frames is beneficial in this aspect, as uncommon attachments are included as well.[4]

[4] Roberts et al. (2014) set a threshold of at least 5 occurrences in the used corpus.

9.4 Creation of a Legal Domain Ontology



custom.owl

BDG - kurz.txt

Checking for ontology connections...

60%

Figure 7: As part of the disambiguation process, attachments are checked for connections in the loaded ontology

An important aspect in this work is the inclusion of domain knowledge in the disambiguation process. The approach of using ontological information for disambiguation tasks has already been used by Kang and Lee (2001) for word-sense dismbiguation, Batista et al. (2012) utilized semantic similarity between toponyms (location names) and ontological classes for disambiguation and Cimiano et al. (2005) have used ontological information for information extraction systems in the biomedical domain. As the creation of an ontology for the legal domain is not a trivial task in itself, the focus on the engineered ontology was laid more on the inclusion of 'practical' domain knowledge than on an exact and thorough representation of the legal domain.[5] As a base structure, the top level ontology *LKIF Core* was utilized, which aims to provide a grounding for knowledge representation in the legal domain (Hoekstra et al., 2007). The class names and relations were translated into German, and additional classes, relations and especially individuals (the original top-level domain contains none) were added.

[5] It can be safely assumed that the internal structure of the created ontology does not accurately represent the amount of or relations between ontological concepts a thorough representation of this domain would contain. This task would require a domain expert and an experienced ontology engineer.

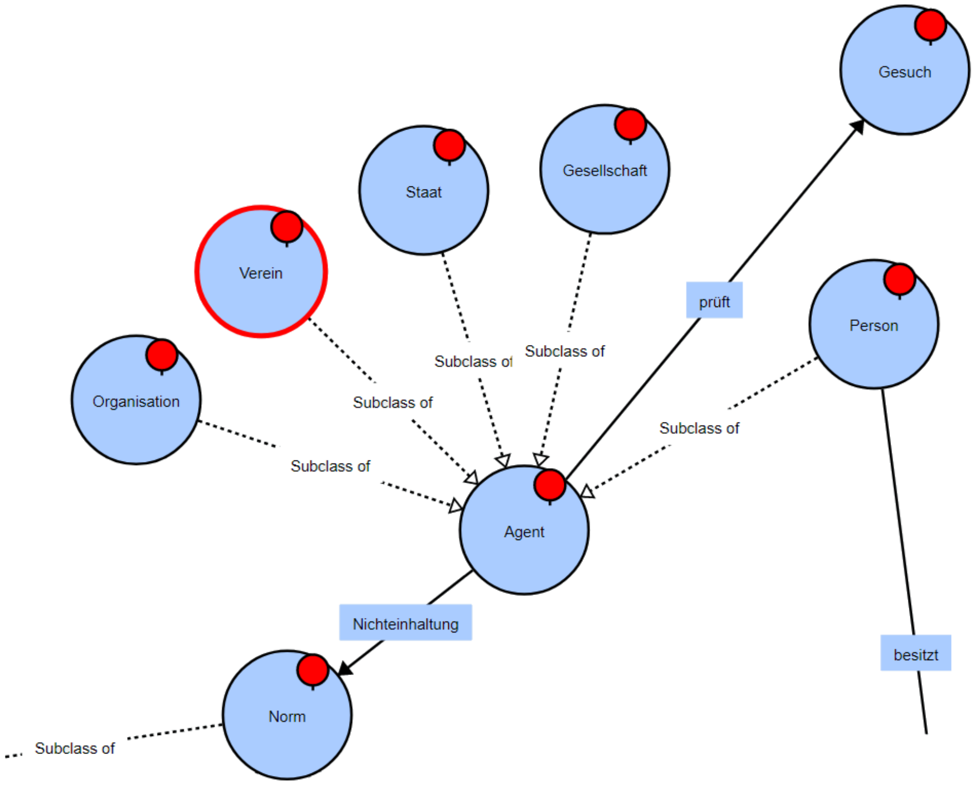


Figure 8: An excerpt from the created ontology, not all relations shown

The full ontology in OWL syntax is included in this work.

9.5 Disambiguation by Ontological Relatedness

The utilization of the ontology in the disambiguation process can be described as follows: Given a pair of a possible attachment target t and a prepositional phrase p, the ontology is queried for the following connections:

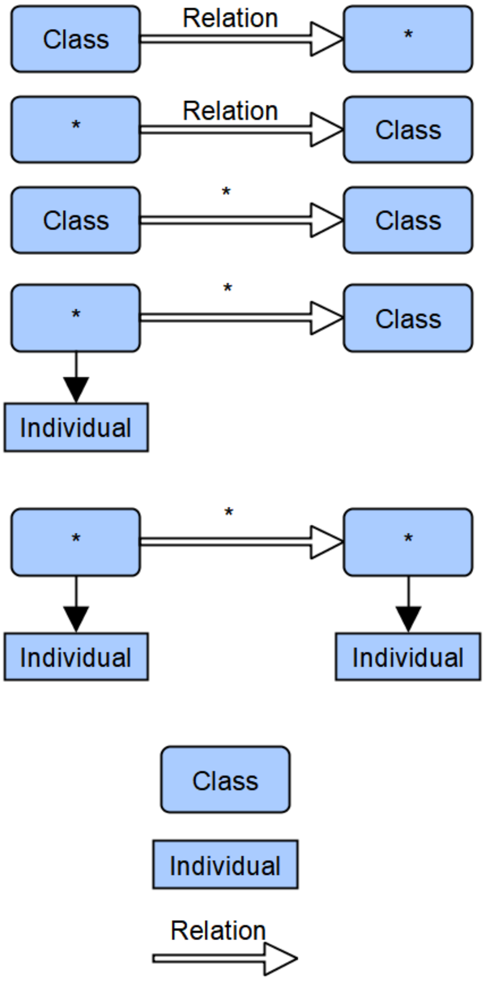


Figure 9: Connections searched for

Where * denotes a placeholder. The general approach can be described as trying to find a connection in the ontology between the given expressions t and p. It is expected that if such a connection has been found, the attachment is more likely. This connection can appear in the form of one expression corresponding to a class name that appears in a relation, either on the left side (in OWL terms: the *domain*) or on the right side (in OWL: *range*). Another possible connection would be two classes, connected through some unknown relation, or an individual which is of a class that is connected to some other class or individual. The last three cases denote no connection, instead the two terms t and p are combined and matched against a class name, indivdidual name, or relation name, either p appended by t or vice versa. The matching between an ontological entity and a given expression is a difficult task, as the ontology contains single terms for each of its entities, and the text excerpt may morphologically deviate from its base form. Another difficulty is posed by the inevitable encountering of synonyms or paraphrasings. The mapping thus has to take lexical and semantic similarity into account. This issue is addressed in two ways. Morphological differences are smoothed through stemming both, the ontological entity's name and the text excerpt, and by using *Levenshtein* distance to measure the similarity between the two terms. The difficulty in mapping proves to be a more challenging task. Sanfilippo (2006) pointed out that most ontologies do not contain much lexical information, and proposed a transformation of *WordNet* concepts to ontological classes. As this work is restricted to a specific domain, the approach will differ slightly. The ontology is created first, and, to introduce a notion of semantic similarity, classes and relations in the ontology are related to synsets from the *Open-de-WordNet*, which is an open-source WordNet initiative for the german language[6]. This approach has been put to use already by Niles and Pease (2003) who annotated *SUMO*, a top-level ontology with WordNet synsets. As the domain and thus the created ontology is restricted for this work, this assignment can be done manually, by querying the Wordnet for classes and relations from the ontology. If the term which is used for the ontology component is not found in the wordnet, the synonym (if existing) which appears closest in meaning is used. Most expressions result in multiple senses and thus synsets, which are sighted, and the sets which are most similar to a human judge are linked to the class or relation in the ontology. An exemplary linkage is illustrated in the following example.

The ontology contains a relation *Einhaltung*, that links an *Agent* to *Norm* or *juristisches Dokument*. The chosen synsets for *Einhaltung* contain entries

[6] https://github.com/hdaSprachtechnologie/odenet

such as *Befolgung*, *Recht und Gesetz*, *Gesetz*, *Disziplin* or *Anerkennung*. The corresponding verb form *einhalten* yields *akzeptieren*, *befolgen*, *im Sinne*, *gemäß*, and *je nach*. For a possible attachment of the prepositional phrase *von Artikel 3 des Grundgesetzes* to the noun *Befolgung*, a connection would be found in the form of the relation *Einhaltung* which has a *Norm* as domain. With the mapping of *Gesetz* to *Einhaltung*, a connection between the terms *Gesetz* (maps to the relation *Einhaltung*) and *Gesetz* (which maps to *Norm*) will be found too, which in this case might be sensical, but might prove as misleading for other mappings. Concerning the attachment decision between *Befolgung* and *von Artikel 3 des Grundgesetzes*, suppose that the set of possible attachments for the PP contains the term *Staatsbürger*. In this case, the ontological relatedness would lose its decisiveness, as there are now two connections:
Agent ('Staatsbürger') - Norm ('Artikel 3 des Grundgesetzes')
Befolgung ('Einhaltung') - Norm ('Artikel 3 des Grundgesetzes')
This is actually a case, where subcategorizational information could not help in the disambiguation process either, as both, *Staatsbürger* and *Einhaltung* allow a PP with *von*. A possible solution for this problem will be presented in the next subsection.

9.6 Disambiguation through Domain-Specific Selectional Preferences

The concept of using selectional preferences for PP attachment disambiguation has already been proposed by Calvo et al. (2004), who automatically aquire preferences in the form of WordNet synsets from a corpus. However, as the disambiguation task is not restricted to a domain, these synsets are very broad (like *place* or *substance*) (Calvo et al. (2004), 5). In this system, the third disambiguation module makes use of selectional preferences that are special to the legal domain. The noun *Verpflichtung* for example kann be used without a complement, or with a prepositional phrase starting with *aus*, *von*, *durch*, or *zu*:

[NP Verpflichtung [PP *aus/durch/zu* _]]

If we consider the clause

, dann entsteht durch seine Verpflichtungen aus der geltenden Verordnung die Notwendigkeit für eine eindeutige Zuweisung.

the *PP aus der geltenden Verordnung may* be attached to either *Verpflichtungen* or *entsteht*. Both, *entstehen* and *Verpflichtung* allow for a PP starting with *aus*. However, in legal texts, the noun *Verpflichtung* is likely to be linked to a member of the semantic class of *Norm* or *Abkommen*. These selectional preferences are encoded in rules of the following form:

Verpflichtung - aus <Norm> | *aus <Abkommen>*

These preferences are obtained by parsing a corpus of legal documents, and manually sighting all possible PP attachments for a given noun, verb or adjective. The legal texts employed in this task are taken from both, german translations of european laws and regulations and from the german body of laws. In particular, the texts used where: The *Bundesdatenschutzgesetz BDSG* (*Federal Data Protection Act*), the *Straßenverkehrsgetz* (*Road Transport Law*), the *Grundgesetz* (*German Basic Law*), the *Verordnung des europäischen Parlaments über eine gemeinsame Einfuhrregelung* (*Common Rules for Import*) and the *Vertag von Lissabon* (*Treaty of Lisbon*). These texts were chosen to represent a broad scope of legal texts, with specialized vocabulary (import, data protection) and texts that use a vocabulary that includes more basic legal terms like the *Basic Law* and the *Treaty of Lisbon*.

To further illustrate this approach, an excerpt of found attachments with selectional preferences is given below.

verletzen - in <Recht>
Bestätigung - durch <judikative Instanz>
erlassen - gemäß <Norm> | *gemäß <juristisches Dokument>*

The semantic classes in these preferences are not refering to WordNet synsets, instead they point to ontological classes. *Open-de-WordNet* does not contain a Synset *judikative Instanz*, but the ontological class *Judikative Instanz* has the subclasses *Gericht*, which in turn is instantiated by different individuals like *Amtsgericht* or *Landesgericht*. This allows for both, leveraging the internal taxonomy of the ontology which is more detailed in the given domain (*Amtgericht* and *Landesgericht* are both not present in the wordnet) and utilizing the synsets that may be linked to an individual or class.

9.7 A Score-based System for the Attachment Decision

The primary aim of the analysis of legal texts with the built system is the filtering of possible attachment ambiguities, that means, attachments like *Anschlag - über mögliche Gefahren* shall be ruled out, in the given case due to *Anschlag* not subcategorizing a PP starting with *über*. However, in many cases, nouns allow for the presence of many different prepositional phrases. The noun *Gesetz* allows for PPs starting with *zu-*, *über*, *von*, *betreffs*. The common auxiliary verb *haben* allows for the attachment of *zu* and *von* too, as in *ich habe den Tipp von dir*. As it is not unlikely to have both *haben* and *Gesetz* in the same sentence, and as intitial tests have shown, many ambiguous PP attachments could not be fully resolved. For this purpose, the lexical resource of selectional preferences and the search for ontological connections shall be used as decisive elements. If two or more attachments are possible for a given prepositional phrase, the attachments are ranked according to a simple score:

The attachment's score is incremented by 1 point, if the subcategorization frame of the target matches the PP's beginning:

Verweis: [_ (PP *auf* -)]

and by another point if the selectional preference of the PP is met as well:

Verweis: [_ (PP auf <Norm>)]

If a connection between *Verweis* and *Norm* is found in the legal ontology, the attachments score is incremented by one point.

The highest-scored attachment will be the one which is selected. If a selectional preference is met, often an according ontology connection will be found too, as in the above case between *Verweis* and *<Norm>*. The reason for the stronger weighting of the selectional preferences is the reasoning that there is evidence that these attachments are likely to be seen in legal texts (as they have been manually sighted) whereas a found ontology connection does not always indicate the right attachment, as the evaluation will show.

9.8 Detection of Potential Scopal Ambiguities and Coordination Ambiguities

The detection of potential scope ambiguities is performed by matching for predefined patterns that include two or more scope-bearing operators. This search is performed per clause, as the scope of of a quantifier is restricted to its minimal surrounding clause (Pafel (2006), 4). In a first version, all clauses that contained two or more terms from a predefined list of operators where included in the output. This proved to include too many false positives, as terms like *ein-* are included in many sentences. Subsequently, a list of common patterns was created, which is exemplified by the following examples:

(alle/jeder/manche/einige/...) ... (nicht/nie/gelegentlich/manchmal/ ...)
"Jeder macht gelegentlich einen Fehler."

(keine/nicht/niemals/nur/...) ... (können/müssen/dürfen/sollen)
"..., die nur zur Erfüllung der gesetzlichen Regelung angewandt werden dürfen."

(können/müssen/dürfen/sollen) ... (nur/niemals/nicht/...)
"Der Mitgliedstaat darf diese Regelung nicht verletzen."

ein ... (jeder/alle/die meisten/manche/einige/zwei/...)
"Ein gesetzlicher Vertreter spricht für jeden Minderjährigen."

The output of the PCFG parser is also used for the detection of possible coordination ambiguities. These are extracted by pattern matching on the flattened parse tree. Patterns in the form of

*(VAFIN *) (NP *) (CVP (VP .. (VVPP *)) (KON *) (VP .. (VVPP *))*
habe einige Briefe .. gelesen und .. geschrieben

are defined for the different types of coordination ambiguities illustrated in chapter *2.3.2*. The full list of defined scope and coordination patterns can be sighted in the accompanying source code.

10 Testing the System on the DUBLIN3 Regulation

10.1 Overview of the Evaluation Procedure

The system was tested on the *DUBLIN3* regulation, which is the legal text that Reyle examined in her work in (Reyle, 2016). This choice offers the advantage of having been manually examined already, which allows for comparison with the automated approach. The experiment was performed in the following steps:

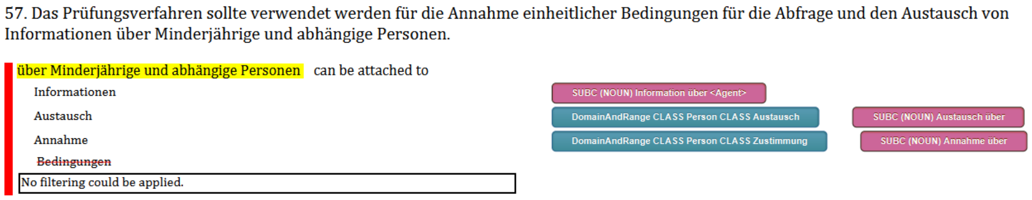
1. Manual preprocessing of the text

To smooth the process of sentence extraction and parsing, the source text was trimmed of headings, annotations of date, and all other fragments of text that did not form a complete sentence.

2. Loading the text in the application and processing it

3. Manual examination of the generated report

The report document lists all sentences that have been determined to either contain a PP attachment ambiguity, a potential scopal ambiguity or a coordination ambiguity, in the order in which they appear in the source text.



57. Das Prüfungsverfahren sollte verwendet werden für die Annahme einheitlicher Bedingungen für die Abfrage und den Austausch von Informationen über Minderjährige und abhängige Personen.

über Minderjährige und abhängige Personen can be attached to

Informationen

Austausch

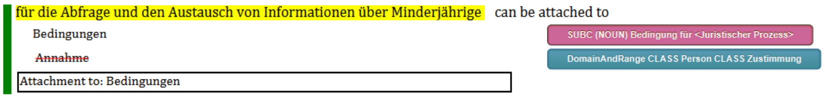
Annahme

~~Bedingungen~~

No filtering could be applied.

SUBC (NOUN) Information über <Agent>
DomainAndRange CLASS Person CLASS Austausch
SUBC (NOUN) Austausch über
DomainAndRange CLASS Person CLASS Zustimmung
SUBC (NOUN) Annahme über

Figure 10: The attachment could not be decided



für die Abfrage und den Austausch von Informationen über Minderjährige can be attached to

Bedingungen

~~Annahme~~

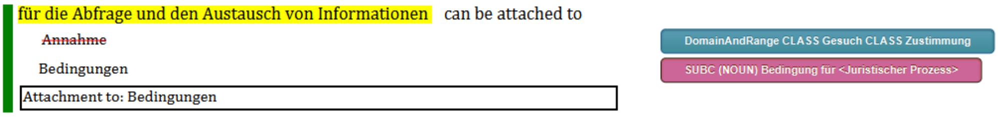
Attachment to: Bedingungen

SUBC (NOUN) Bedingung für <Juristischer Prozess>
DomainAndRange CLASS Person CLASS Zustimmung

Figure 11: Subcategorizational information excludes one possibility

Each ambiguous PP attachment is listed along with its possible attachment targets. If the subcategorization frame or the selectional preference of the attachment target matches the prepositional phrase, a visual indicator in form of a label is given, the same goes for any found ontological connection. Attachment targets that where found to not subcategorize for the given PP are ruled out. If a decision of attachment could be made according to the score, it is presented as well. If any clause in the sentence was found to contain a possible scopal ambiguity, it is marked and listed at the bottom of each entry, along with any detected coordination ambiguity.

The PCFG parser sometimes includes incorrect bounded prepositional phrases, which, if they are ambiguous in their attachment, are included in the output. This is illustrated in figure 11.



für die Abfrage und den Austausch von Informationen can be attached to

~~Annahme~~

Bedingungen

Attachment to: Bedingungen

DomainAndRange CLASS Gesuch CLASS Zustimmung
SUBC (NOUN) Bedingung für <Juristischer Prozess>

Figure 12: Incorrectly bounded PP

In such a case, only the correctly determined phrase is valued in the experimental result. The generated report is measured by the following metrics:

1. Share of possible attachments that could be ruled out due to subcategorizational information
2. Share of attachments where only one possibility remained
3. Accuracy of attachment assignments based on the scoring system
4. Accuracy of selectional preferences as single predictor
5. Accuracy of ontological relations as single predictor
6. Accuracy and recall of detected scopal ambiguities
7. Accuracy and recall of detected coordination ambiguities

10.2 Evaluation Results

The processing of the document yields a total of 459 sentences, of which 438 (95.4%) contain at least one prepositional phrase. Of these, a count of 389 sentences (88.81%) is considered ambiguous in the sense that at least one contained PP can be attached to more than one target. 28 sentences are marked to contain potential scopal ambiguities. The evaluated statistics from the manual sighting are given below.

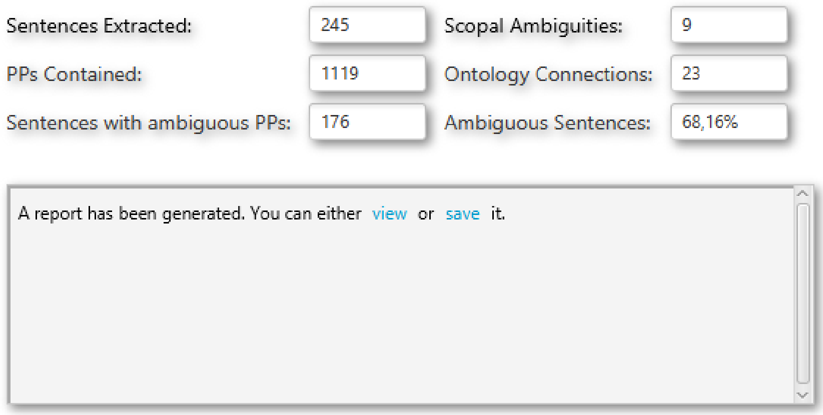
Metric	Value
Share of ruled out attachments	19.44%
Share of attachments where only one possibility remained	17.94%
Accuracy of attachment assignments based on the scoring system	73.84%
Accuracy of selectional preferences as single predictor	92.4%
Accuracy of ontological relations as single predictor	67.1%
Accuracy of detected scopal ambiguities	75.0%
Recall of detected scopal ambiguities	100%
Accuracy of detected coordination ambiguities	42.30%
Recall of detected coordination ambiguities	87.02%

Table 1: Evaluation results

10.3 Discussion of the Results

This work serves two main purposes:
1. The creation of a system that allows for the detection and reporting of ambiguities in legal texts.
2. The evaluation on how syntactic (in the form of subcategorization frames) and semantic (in the form of selectional preferences and a domain ontology) information can be used to help in the disambiguation process.

The first goal (detection and reporting) can, in the case of ambiguous prepositional phrases, not easily be measured numerically, as the count of reported possible attachments depends on the utilized parser.



Sentences Extracted: 245
Scopal Ambiguities: 9
PPs Contained: 1119
Ontology Connections: 23
Sentences with ambiguous PPs: 176
Ambiguous Sentences: 68,16%

A report has been generated. You can either view or save it.

Figure 13: A report is generated that can be viewed in the application or saved as .pdf file. Note: The displayed count of PPs is that high because all incorrectly bounded PPs are included as well.

The system identified all scopally ambiguous sentences that were found by manual sighting, it included however some false positives as the accuracy of 75.0% shows. A majority of the contained coordination ambiguities were found (87.02%), however, a majority (57.70%) was found to be unambiguous to a human reader. The majority of attachments could not be resolved by excluding candidates for which the subcategorization frame did not match, however, 19.4% of the possiblities could be excluded. Selectional preference rules proved to be very accurate in predicting the right attachment. Yet, they could be applied only to a subset of the ambiguous sentences, as not all attachments contain domain-specific vocabulary, e.g. *Informationen* and *über ihre Fähigkeiten*. The average count of possible attachment targets in the tested document was 2.86, which equates a 34.96% chance of choosing the correct attachment randomly. Ontological connections as sole predictor reached an accuracy[7] of 67.1%, which leaves space for improvement, but which supports the thesis that terms which are connected in the ontology

[7] As for selectional preferences, this accuracy was measured over sentences where only one connection was found.

have a higher probability of being attached.

Connections between two classes or instances of two classes have been found with a higher frequency than class - relation pairs. Relation and class names on their own did not match any expression, and only one expression (*Recht auf Asyl*) could be matched with the name of an individual (which is why it is excluded in figure 12).

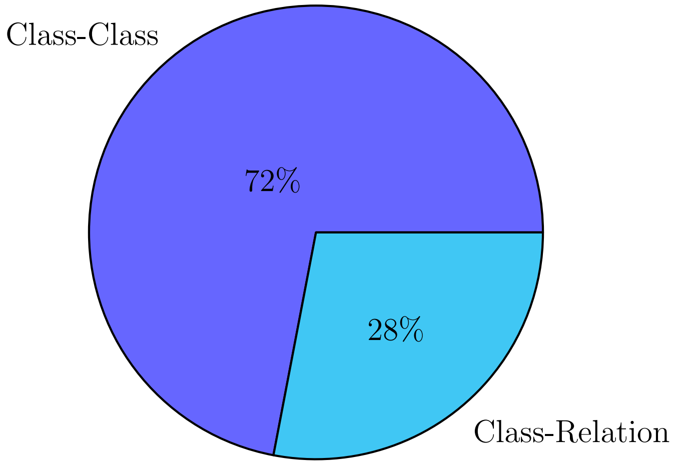


Figure 14: Ontological connections by type

It shall be emphasized here that the used approach might not be feasible for disambiguation in the parsing process, if performance is of importance. Especially the process of querying the ontology for connections proves to be very time-consuming (the total runtime for the tested document was about 7 minutes).

10.4 Proposed Improvements on the Semantic Disambiguation

One possible approach, which could improve the coverage through selectional preferences would be to automatically derive these preferences from a domain-specific corpus. The ontological creation and linking to wordnet classes would remain the same, but prepositional attachments, which strongly associate with a given ontology class would be determined statistically, as Calvo et al. (2004) did for *WordNet* classes from an untagged corpus. The ontological connections might profit of a weighting scheme, that takes the 'granularity' of the found connection into consideration: As most ontologies contain some kind of internal taxonomy, the distance between a matched class and the top level entity (in OWL: *thing*) can be measured, and a connection between two classes might be weighted according to their summed distances from the top level. If two terms correspond to highly specialized entities in the ontology that are linked together, an attachment might be more probable. In general, a matching individual name should be weighted higher than a matching class name, as individuals in the legal domain often have proper names (an example out of the test document, that could have been detected by a matching indivudal name is *Europäisches Unterstützungsbüro - für Asylfragen*). In this regard the performance could be improved by further increasing the granularity of the ontology, although it might be debatable if a term like the aforementioned *Europäisches Unterstützungsbüro für Asylfragen* falls under he scope of a legal domain ontology.

Another improvement which might be worth investigating is the inclusion of other sentence parts in the decision process: The ontology contains the connection *Person* - *wohnhaft-in* - *Staat.* Suppose the attachment in question was *Wohnsitz* and *im Ausland.* If the subject of the sentence mapped to *Person* too, as in
(Der Staatsangehörige <Person>) muss seinen (Wohnsitz <wohnhaft-in>) im (Ausland <Staat>) registrieren,
the connection might indicate a higher probability of attachment. y

11 Conclusion

Ambiguity is an integral feature of human language. However, in legal texts, it has to be minimized. The presented system allows for the detection of scopal, coordination and prepositional phrase ambiguities in statutory texts, although its basic approach can be generalized to many domains. Research has shown that in many cases, lexical and world knowledge are needed for the correct solving of ambiguous expressions. While for general ambiguity resolution, the inclusion of world knowledge poses a challenging task, it is a feasible approach for smaller domains. This work utilizes domain-specific knowledge in the form of selectional preferences and by querying a domain ontology, to reduce the number of possible readings. The detection of scopal ambiguities has proven to be precise in terms of recall, however, a human survey is still needed for the decision about whether a true ambiguity is present or not. Additionally, the applicability of ontological connections and selectional preferences as features in an attachment decision process has been investigated, and found to be promising. As the disambiguation in the described system is time-intensive, the possible integration into a parser is questionable. However, subcategorization frames, ontological relatedness and and selectional preferences might be used as features in an expert disambiguation module. An information extraction system for the legal domain would be an example of an application that might profit of a more costly disambiguation process, since, as it has been shown, exact representation of meaning is of special importance in the jurisdiction.

References

Jürgen Pafel. *Quantifier Scope in German.* John Benjamins Publishing Company, Germany, 2006.

Galen Andrew and Bill MacCartney. *Statistical resolution of scope ambiguity in natural language.* Citeseer, 2004.

Thomas Ede Zimmermann and Wolfgang Sternefeld *Introduction to Semantics: An Essential Guid to the Composition of Meaning.* Walter de Gruyter, 2013.

Björn Gambäck and Johan Bos *Semantic-head based resolution of scopal ambiguities.* COLING Proceedings of the 17th international conference on Computational linguistics, Vol. 1, 1998.

Sanford Schane *Ambiguity and Misunderstanding in the Law.* HeinOnline,T. Jefferson L. Rev. Vol. 25, 2002.

Christopher Kennedy *Ambiguity and Vagueness - An Overview.* Semantics: An international handbook of natural language meaning, Vol. 1, pages 507-535, 2011.

Matthias Bauer, Joachim Knape, Peter Koch *Dimensionen der Ambiguität.* Springer, Zeitschrift für Naturwissenschaft und Linguistik (40), pages 7-75, 2010.

Edward N. Zalta and others *Stanford Encyclopedia of Philosophy.* Stanford University, The Metaphysics Research Lab, 2003.

Robert Andrew Wilson and Frank C. Keil *The MIT Encyclopedia of Cognitive Sciences.* MIT Press, 2001.

M. K. Anjali and Anto P. Babu *Ambiguities in Natural Language Processing.* International Journal of Innovative Research in Computer and Communication Engineering, India, 2014.

Liparota v. United States. Supreme Court Database, ID: 1984-093, Case No. 471 U.S. 419.

Karl Larenz *Methodenlehre der Rechtswissenschaft.* Springer-Verlag, 2013.

Hannah Reyle *Rechtsunsicherheit durch Ambiguitäten.* Universität zu Köln, 2016.

Dick Grune and Ceriel Jacobs *Parsing Techniques - A Practical Guide.* Springer, Monographs in Computer Science, 2007.

Dan Jurafsky and James H. Martin *Speech and Language Processing - An introduction to natural language processing, computational linguistics, and speech recognition.* Pearson London, 2014.

Sandra Kübler, Ryan MacDonald, Joakim Nivre *Dependency Parsing.* Morgan and Claypool Publishers, Synthesis Lectures on Human Language Technologies 1, pages 1-127, 2009.

Andrew Carnie *Constituent Structure.* Oxford University Press, 2009.

Joakim Nivre *Algorithms for Deterministic Incremental Dependency Parsing.* MIT Press, Computational Linguistics 34, pages 513-553, 2008.

Joakim Nivre *An efficient Algorithm for Projective Dependency Parsing.* Proceedings of the 8th International Workshop on Parsing Technologies, 2003.

Massimo Poesio *Ambiguity, Underspecification And Discourse Interpretation.* Proceedings of the First International Workshop on Computational Semantics, pages 151-160, 1994.

Kenji Sagae and Alon Lavie *A Classifier-Based Parser with Linear Run-Time Complexity.* Association for Computational Linguistics, Proceedings of the Ninth International Workshop on Parsing Technology, pages 125-132, 2005.

Kenneth Church and Ramesh Patil *Coping with Syntactic Ambiguity or How to Put the Block in the Box on the Table.* MIT Press, Computational Linguistics 8, pages 139-149, 1982.

Jay Earley *An Efficient Context-Free Parsing Algorithm.* ACM, Communications of the ACM 13, pages 94-102, 1970.

David Tuggy *Ambiguity, Polysemy And Vagueness.* De Gruyter, Cognitive Linguistics 4, pages 273-290, 1993.

Geert Keil and Ralf Poscher *Vagueness and Law.* Oxford Unversity Press, 2016.

Steffen Staab and Rudi Studer *Handbook on Ontologies.* Springer Science and Business Media, 2010.

Thomas R. Gruber *Toward Principles for the Design of Ontologies Used for Knowledge Sharing.* Elsevier, International Journal of human-computer studies 43 p.907-928, 1995.

Cristina Romero Tris, Riao D., Real F. *Ontology-Based Retrospective and Prospective Diagnosis and Medical Knowledge Personalization.* in: Knowledge Representation for Health-Care, Springer, Lecture Notes in Computer Science 6512, 2011.

Robert Hoehndorf *What is an upper level ontology?.* Ontogenesis, 2010.

Philip Stuart Resnik *Selection and Information: A Class-Based Approach to Lexical Relationships.* IRCS Technical Reports Series, 1993.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, David McClosky *The Stanford CoreNLP Natural Language Processing Toolkit.* Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations p.55-60, 2014.

Anna N. Rafferty and Christopher D. Manning *Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines.* Association for Computational Linguistics, Proceedings of the Workshop on Parsing German, pages 40-46, 2008.

Will Roberts and Markus Egg and Valia Kordoni *Subcategorisation acquisition from raw text for a free word-order language.* Association for Computational Linguistics, Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 298-307, 2014.

Philipp Cimiano and Uwe Reyle and Jasmin Saric *Ontology-driven discourse analysis for information extraction.* Elsevier, Data and Knowledge Engineering Vol. 55, pages 59-83, 2005.

S. Kang and J. H. Lee *Ontology-Based Word Sense Disambiguation by Using Semi-Automatically Constructed Ontology.* In Proceedings of MT Summit 8, Spain, 2001.

David S. Batista, Joao Ferreira, Francisco M. Couto, Mario J. Silva *Toponym Disambiguation using Ontology-based Semantic Similarity.* Springer, International Conference on Computational Processing of the Portuguese Language, pages 179-185, 2012.

Hiram Clavo and Alexander Gelbukh *Acquiring Selectional Preferences from Untagged Text for Prepositional Phrase Attachment Disambiguation.* Springer, International Conference on Application of Natural Language to Information Systems, pages 207-216, 2004.

Rinke Hoekstra, Joost Breuker, Marcello Di Bello, Alexander Boer and others *The LKIF Core Ontology of Basic Legal Concepts.* LOAIT Journal, Vol. 321, pages 43-63, 2007.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, Katherine J. Miller *Introduction to WordNet: An On-line Lexical Database.*

Oxford University Press, International Journal of Lexicography Vol. 3, pages 235-244, 1990.

Sanfilippo, Antonio P and Tratz, Stephen C and Gregory, Michelle L and Chappell, Alan R and Whitney, Paul D and Posse, Christian and Paulson, Patrick R and Baddeley, Bob and Hohimer, Ryan E and White, Amanda M *Ontological Annotation with WordNet.* Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2006.

Ian Niles and Adam Pease *Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology.* Ike, pages 412-416, 2012.

Mehl, Stephan and Langer, Hagen and Volk, Martin *Statistische Verfahren zur Zuordnung von Präpositionalphrasen.* Proceedings of the 4th Conference on Natural Language Processing, Vol. 98, 1998.

Langer, Hagen and Mehl, Stephan and Volk, Martin *Hybride NLP-Systeme und das Problem der PP-Anbindung.* Berichtsband des Workshops" Hybride konnektionistische, statistische und symbolische Ansätze zur Verarbeitung natürlicher Sprache", 1997.

Christopher D. Manning and Hinrich Schütze *Foundations of Statistical Natural Language Processing.* MIT Press, 1999.

Europäische Kommission *Gemeinsamer Leitfaden des Europäischen Parlaments, des Rates und der Kommission für Personen, die in den Gemeinschaftsorganen an der Abfassung von Rechtstexten mitwirken.* Amt für amtliche Veröffentlichungen der europäischen Gemeinschaften, Luxemburg, 2003.

Bundesministerium der Justiz *Handbuch der Rechtsförmlichkeit.* Bundesanzeiger, Bundesministerium der Justiz, Jahrgang 60, 2008.

Sabine Schulte im Walde *A subcategorisation lexicon for German verbs induced from a lexicalised PCFG.* in LREC, 2002.

Sabine Schulte im Walde *The induction of verb frames and verb classes from corpora.* In Corpus Linguistics. An International Handbook, Chapter 61, De Gruyter, Berlin, 2009.

Marc Light and Warren Greiff *Statistical models for the induction and use of selectional preferences.* Weiley Online Library, in Journal of Cognitive Science, Vol. 26, pages 269-281, 2002.

Michael Schiehlen *A cascaded finite-state parser for German.* Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, Vol. 2, pages 163-166, 2003.

Thomas Wasow, Amy Perfors and David Beaver *The Puzzle of Ambiguity.* CSLI Publications, Stanford, Morphology and the web of grammar: Essays in memory of Steven G. Lapointe, Vol. 2, pages 265-282, 2005.

Statement of Authorship

I hereby declare that the work presented in this thesis is entirely my own. I did not use any other sources and references that the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

Date and Signature: