A5 Automatic classification of concept types

1 General information

1.1 Applicants

Prof. Dr. KILBURY, James

Prof. Dr. LÖBNER, Sebastian

1.2 Topic

Automatic classification of concept types

1.3 Scientific discipline and field of work

Semantics, statistical language processing, computational linguistics, corpus linguistics

1.4 Scheduled total duration

Six years

1.5 Application period

Three years

1.6 Summary

The goal of the project is the development of an automatic classification of types of nouns in text corpora by statistical methods. The classification exploits the fact that nouns of the relevant types (sortal, individual, relational and functional) differ in the grammatical contexts in which they occur. This is due to their semantic properties of inherent uniqueness (individual and functional concepts) and inherent relationality (relational and functional concepts). While relational nouns are much more frequently used with a possessor specification, inherently unique nouns occur significantly more often with definite determination and in the singular. These uses can be automatically recognized by computational methods requiring a certain degree of parsing, but no semantic analysis. Distributional differences between the four types allow the classification of a given noun, if it occurs with sufficient frequency. Refinement of the contextual criteria, a growing lexicon with relevant information, and advances in the statistical procedure itself will allow for a reduction of the number of occurrences needed for classification. Part of the research to be pursued in the project is the development of appropriate tagging software, a combination of existing tools with supplementary programming. This work is supported by a theoretical analysis of the contexts and constructions in which different types of nouns occur (a) if they are used in accordance with their underlying type and (b) if they undergo certain kinds of type shifts. The analyses of type shifts are important since these occur quite frequently and tend to blur the distributional characteristics of the nouns being classified. The object language is German, but later the project goals will be extended to French and English.

2 State of the art, preliminary work

2.0 Problem to be adressed

2.0.1 The type distinction and its relation to determination

The four types of nouns to be distinguished are sortal nouns (SC)[1], individual nouns (IC), relational nouns proper (RC), and functional nouns (FC). The four types differ along two independent semantic parameters: relationality and inherent uniqueness. These characteristics are immediately relevant for the way in which the respective nouns are used:

- IC and FC nouns are inherently unique; therefore, in their unmarked uses they are combined with definite determination and in the singular.
- SC and RC nouns do not semantically determine the number of their potential referents in a given context: they may apply to zero, one, or multiple instances; therefore, they are open for singular and plural use. The unmarked use is indefinite.
- RC and FC nouns are inherently relational. They therefore require the specification of their possessor argument for determining their referents in a given context. While the

[1] When dealing with these types of nouns, no distinction needs be drawn between nouns (as lexicalized concepts) and concept: a noun is a functional noun if it is the lexicalization of a functional concept. As the term concept can be used for words as well as their meanings, there is no harm when denoting in this context, e.g., functional nouns as functional concepts.

specification of a possessor is not syntactically obligatory, RC and FC nouns do occur significantly more frequently with possessor constructions of various sorts.

Table 1 displays the four types and the types of determination they occur with.

	not inherently unique [–D]	inherently unique [+D]
1-place [–P]	**SORTAL – SC** *table book adjective water* ↼definite determiner ✓**indefinite** ✓dem. ✓quant. ✓plural ↼possessive	**INDIVIDUAL – IC** *moon weather time Maria* ✓singular **definite determiner** ↼indefinite ↼demonstr. ↼quant. ↼plural ↼possessive
2-place [+P]	**RELATIONAL – RC** *sister blood leg attribute* ↼definite determiner ✓**indefinite** ✓dem. ✓quant. ✓plural ✓**possessive**, possessor must be specified	**FUNCTIONAL – FC** *father head age subject* (gramm.) ✓singular **definite determiner** ↼indef. ↼demonstr. ↼quant. ↼plural ✓**possessive**, possessor must be specified

Table 1: correlation of concept types and determination

The non-relational types SC and IC logically are 1-place predicators, while RC and FC are 2-place. There are also higher-place relational and functional nouns; but these are comparatively rare. The present study is therefore restricted to the most frequent four basic types. In the table, ✓ marks a type of determination which accords the type of the noun, while ↼ marks a type of determination at variance with the noun type. This does not mean that these combinations are ungrammatical, rather that they are accompanied or made possible by a type shift of the noun that gives it an appropriate type (See 2.0.4 below). Among the determinations mentioned, plural in general, quantification, indefiniteness, and contrastive demonstrative all require at least the possibility of an open number of potential referents in the given context. Therefore, these types of determination require a sortal or relational head noun ([–D]). Only definite singular use is in accordance with the inherent uniqueness of SC and FC nouns ([+D]). For the use with the other determinations mentioned, a type shift is required which suspends the property of inherent uniqueness. Conversely, [–D] nouns require a particular context, such as the prior introduction of the unique referent for definite singular use. Löbner (1985) argues that the use of the definite article inevitably renders a contextually construed IC reading of the noun, and hence a type shift, if the noun is not [+D]. As for the [±P] distinction; RC and FC nouns can only be used for reference, if the possessor argument is specified in the given context. They are frequently used with possessive constructions of various kinds. If they lack a possessor specification, they need a special context that provides it. Conversely, SC and IC nouns require a special context for possessive use.

Table 2 displays figures representing the frequencies of singular definite use and possessive use, respectively, for four representative nouns in a 110.000 word German text corpus on semantics (Löbner 2003).

SORTAL NOUN		**INDIVIDUAL NOUN**	
Nomen („noun“), 166 tokens		*Semantik* („semantics“), 152 tokens	
definite, singular	**37 %**	definite, singular	**82 %**
possessive	**0%**	possessive	**4%**
RELATIONAL NOUN		**FUNCTIONAL NOUN**	
Teil („part“), 124 tokens		*Bedeutung* („meaning“), 721 tokens	
definite, singular	**36 %**	definite, singular	**57 %**
possessive	**73%**	possessive	**74%**

Table 2: frequencies for types of determination for four nouns of different type

The table illustrates the highly significant differences in frequency with respect to the two types of determination. A further look at the data, however, also reveals unexpected results, such as the fact that the count noun *Bedeutung* appears in definite bare singular use without any determiner in about 110 cases.

We checked the significance of these two parameters of use with a larger set of nouns occurring in the same corpus with some frequency and received the following distribution (Fig.1):

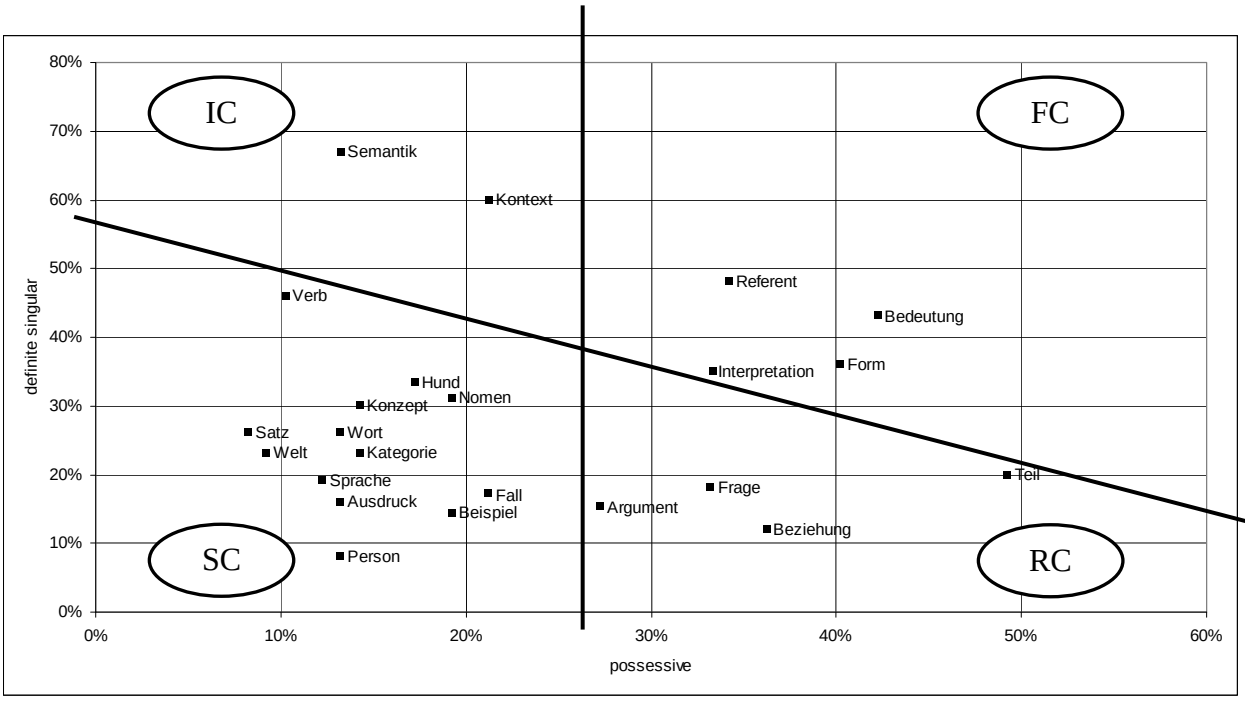

Figure 1: grammatical environments of the most frequent nouns in the corpus (automatic annotation)

On the one hand, the results show clusters for the different concept types: sortal nouns (such as *Verb, Hund, Nomen*, etc.) essentially occur with a lower percentage of definite singular and possessive use (lower left). Relational nouns (such as *Teil, Interpretation, Form*) have a

higher percentage of possessive use, whereas the concept types with unique referents are generally located in areas with higher percentages of definite singular use (*Referent* (FC), *Bedeutung* (FC), *Semantik* (IC)). On the other hand the distinction is not very sharp yet. Certain findings, such as the locations of *Welt* (IC), *Kontext* (FC), or *Beispiel* (RC), cannot be explained with respect to their respective underlying concept types. The analysis hence requires a deeper investigation of the factors responsible for the high quota of marked uses. An atypical contextual bias may be one source. For example, the noun *Welt* is in most cases used in the corpus as part of the A-N term *mögliche Welt* in the chapter about formal semantics. In this special use, the noun simply carries a SC reading responsible for the low rate of definite singular uses.

2.0.2 Relevant contexts

Since the differences between the types are primarily semantic and only secondarily grammatical, it cannot be expected that the types of nouns can be distinguished merely by determining the type of determination they come along with. The mere grammatical notion of context based on types of determination is meaningful, but it is not sufficient if a more exhaustive analysis is projected. For example, possessive use of an RC or FC is often, but not always, indicated by grammatical determination. Among the roughly 600 possessive uses of *Bedeutung* in the corpus, i.e. uses where a possessor argument was specified in the context, roughly 20% have a left possessor specification (mostly a possessive pronoun) and less than 50% a right possessive construction; in almost a quarter of the cases the possession was specified in a verbal construction (e.g. *POSSESSOR hat die Bedeutung XY*) or prepositional construction (e.g., *POSSESSOR mit der Bedeutung XY*). While these types of possessor specification can be captured with more elaborate parsing efforts, there is little chance of automatically recognizing those cases where the possessor is specified in the extrasentential context.

2.0.3 Polysemy

Polysemy is a problem faced by all semantic classification tasks. Polysemous nouns – and most nouns are polysemous – tend to occur in larger corpora in more than one meaning variant. For example, the English word *child* is consistently used in two meanings that differ in type: *child* in its SC sense „non-adult person" and in its (primary) RC sense „immediate descendant". Given that both meanings are of frequent use, one would expect that counts such as those underlying Fig. 1 would not yield a clear classification.

2.0.4 Type shifts

A particular challenge to the task of automatic noun-type classification is the massive occurrence of common type shifts. By this we mean re-interpretations of noun meanings in a particular context that requires or lends itself to a reading of different type, e.g. an SC reading for an FC noun. Examples of type shifts of IC and FC nouns may illustrate the point.

(1) a. *Semantics is concerned with meaning.* [IC]
b. *This is not a semantics but a syntax.* [IC → SC]
c. *The semantics of the definite article is elusive.* [IC → FC]
d. *There may be more than one semantics of the passive.* [IC → RC]

The noun *semantics* constitutes an IC noun. It is a plurale tantum and used without an article for definite singular reference. (Its German equivalent *Semantik* would be used with a definite article in the singular.) Given that what exactly semantics is like depends on one's point of view, semantics also exhibits certain sortal characteristics; these are involved in (1b), which would be interpreted as referring to a certain instance of a theory claimed to constitute semantics. In this context, both IC terms, *semantics* and *syntax*, are used as SC terms meaning „semantics-like theory", or „syntax-like theory", respectively. (1c) represents a type shift common to terms for scientific disciplines; the meaning is shifted from a denotation of the discipline to an expression for an account in that discipline of a particular object of the discipline: „semantics" shifts to „the semantic account of", i.e. an FC notion. (1d) adds another shift to this one, a shift which cancels the uniqueness condition, yielding the relational notion „(possible) semantic account of".

(2)	a.	*Mary is Peter's <u>mother</u>.*	[FC]
	b.	*In this family, the <u>mother</u> earns the living.*	[FC → IC]
	c.	*A <u>mother</u> should not smoke.*	[FC → SC]
	d.	*You cannot very well have more than one <u>mother</u>.*	[FC → RC]

(2a) illustrates the common FC use of the term *mother*. Given a context narrow enough for restricting the possible referents of the noun to one (there may be several children, but sharing the same mother), the relationality of the concept can be dispensed with, reducing it to Type IC (cf. 2b). In (2c) the FC „mother (of)" is turned into the SC „mother of a child" by existential binding of the possessor argument, a process analogous to the antipassive alternation of two-place verbs (to be observed, e.g., with the intransitive use of the two-place verb *eat* in the sense of „eat something" or „have a meal"). (2d) illustrates a context in which the property of unique reference is canceled (for the sake of the argument expressed). In general, existential contexts like the object position of *have* lead to a type shift that suspends inherent uniqueness.

Relational and sortal nouns also frequently undergo type shifts in certain contexts. In general, each of the four types can be shifted to each other type. The type-shift mechanisms are both common and manifold; like other shifts, partly better known, such as metaphor, metonymy, mass-count or count-mass shifts, or the broad class of verb alternations, they can be applied to larger classes of semantically similar nouns. Part of the work planned in the project is the semantic analysis of the major types of common type shifts.

As can be seen from Table 1 and Examples (1) and (2), the determinations of type-shifted nouns are in accordance with the resulting type. The question arises if all cases of marked (↼) determination involve type shifts. Intuitively, there are rather specific meaning shifts like the one in (1c) and (1d). They not only bring about a change in type but also some other, in this case metonymic, shift in meaning. They tend to be available only to narrower classes of concepts (e.g. terms for scientific disciplines), relying on certain more specific characteristics of the concepts themselves. Let us call these types of shifts „lexical type shifts", as they depend on the narrower lexical meaning of the noun. On the other hand, there are very common type shifts available for large classes of nouns which are only restricted by very general, unspecific semantic conditions. These shifts include the ones indicated in (3a), (3b), and (3c), among many others:

(3) a. *The orange was very sour.* [SC → IC]
b. *My orange was very sour.* [SC → FC → IC]
c. *My tongue is burning.* [FC → IC]

The noun *orange* is a typical sortal noun. With a definite article in anaphoric use, as might be the case in an utterance of (3a), it becomes a contextually defined IC for unique reference, e.g. the IC „orange I just ate" (see Löbner 1985 for the thesis that any noun with definite article is interpreted as IC or FC). Used with the possessive pronoun in (3b), the noun *orange* is turned into a functional concept „the orange of" (note that the construction is only used definitely). In a second step, the specification that the possessor is the speaker saturates the possessor argument and reduces the FC „the orange of" to the IC „the orange of the speaker". In (3c) the first of these two shifts is missing because the noun *tongue* as a unique body-part term is inherently FC. The shift here is the same as the second one in (3b): saturation of the possessor argument. The second kind of type shifts, illustrated in (3a-c), might be called „functional type shifts".

We assume that all cases of non-standard (↼) determination necessarily involve some kind of type shift. Some of the shifts are special lexical type shifts, others simply functional shifts.[2]

2.0.5 Underlying type vs. actual use

Given the high frequency of type shifts, the question may be raised if it makes sense to try to draw conclusions from actual use of nouns about their underlying meaning. It might even be questioned if there is anything like an underlying meaning of a noun with a fixed type. Our hypothetical answer to both questions is positive. We assume that the human lexicon is built, among other things, on the observation of the frequencies of types of uses. If a particular noun is used very frequently with a possessor specification or with a definite article and in the singular, the language learner will record this fact in their linguistic memory. We also assume (1) that the language learner comprehends the conceptual modifications associated with either lexical or functional type shifts and (2) that these shifts involve a certain amount of cognitive effort. The learner will then arrange their lexicon in such a fashion that the most frequent determination type(s) exhibited by a given noun do not require type shifts. Thus, for example, the learner will arrange their lexicon such that the concept associated with *Bedeutung* involves a possessor argument (as it is in 80% of all uses with a possessor specification) and that it is functional rather than relation, since it occurs in definite singular use most of the time.

We thus hypothesize that when polysemy is resolved for a given noun there will be a most frequent use of the noun, which is in accordance with the distributional characteristics of one of the four types (or some further type not yet included). We also assume that the type of a noun so determined is non-random but rather consistent with its lexical analysis (since the lexical analysis is basically founded on the same observations as the language learner). The project will subject these hypotheses to empirical scrutiny.

---

[2] It follows from this point of view that the most important function of the variety of types of noun determinations is to shift types between lexical meanings and the actual interpretations in context. This line of theorizing is pursued in Project A1.

Additionally, we assume that all type shifts, lexical or functional, are triggered by or at least require specific contexts. In many cases such contexts may be linguistically given in the surrounding text of a noun token. In principle it might be possible to automatically detect contextual cues indicative of, or frequently co-occurring with, certain type shifts. If at least some of the type shifts were detected, the underlying type of the noun could be reconstructed even if it occurs in a context indicative of a different type.

2.1 State of the art

Since the project seeks to provide a solution to a problem not yet studied, there is only a small amount of relevant literature available. The systematic identification and classification of definite descriptions in English texts was pioneered by Poesio & Vieira (1998), Vieira (1998), and Vieira & Poesio (2000). They apply heuristics to a subset of the Penn Treebank (Marcus et al. 1993) to identify definite NPs with functional nouns and are concerned with anaphora resolution. In Poesio & Vieira (1998) human annotators were assigned to classify definite descriptions. The aim of the study was the determination of inter-annotator agreement with respect to different classification schemata. They had to distinguish direct from associative anaphora, which turned out to be a difficult task. Associative anaphora NPs have a functional head (FC noun) and an anaphoric possessor argument which can be explicit or implicit (Löbner 1998). Poesio (2004) describes a comparative analysis of manually annotated corpora that were created by two groups of annotators with different annotation criteria: one group had a list of criteria based on anaphoric relations, while the other group had criteria based on functionality. It turned out that the inter-annotator agreement of the second group was significantly higher than that of the first. These results support the Löbner's (1985) theory that functionality is the defining basis for definiteness, while anaphora is a secondary phenomenon. The studies by Poesio and Vieira do not, however, represent an attempt to automatically determine functional nouns.

The classification of conceptual noun types is a disambiguation problem. The resolution of ambiguities in natural language is a challenge on all linguistic levels. Most of the solutions to these problems use probabilistic methods, because maximum likelihood estimation of a classifier allows clear-cut decisions compared to unweighted disjunctive alternatives resulting from rule-based methods. While Bayesian decision theory (cf. Gale et al. 1992) usually suffers from the hypothesis that the contextual features of a classifier are independent of each other, maximum entropy frameworks (Ratnaparkhi 1998) overcome this limitation. They allow a uniform treatment of contextual features from arbitrary levels of linguistic description while the features can overlap in arbitrarily complex ways. But supervised approaches for the acquisition of classifiers like maximum entropy models normally require large training corpora which have to be annotated manually at high expenses. Fortunately, they can be combined with bootstrapping or semi-supervised methods like Yarowski's (1995) to start with less annotated training data and learn unsupervised from arbitrarily large corpora in a secondary step. The resulting model can then be used to automatically create new training data for repeated supervised learning.

Research on meaning shifts has been carried out by Bierwisch (1983), Partee (1983, 1986), Dölling (1992, 1995), and Michaelis (2004), among others. Bierwisch (1983) observes that numerous nouns can have different interpretations without being polysemous in a nar-

rower sense. He argues that these nouns are based upon one primary meaning, from which other interpretations can be derived due to certain contextual factors. The shifts Bierwisch addresses are metaphor, metonymy, and conceptual differentiation. None of these is immediately related to the type distinctions we have chosen to investigate. Dölling (1992) focuses on the mass-count distinction and provides templates for the analysis of sort shifts but neither explains the general restrictions on sort shifts nor the restrictions of their use.

2.2 Preliminary work, progress report

The systematic distinction between sortal (SC), relational (RC), individual (IC), and functional (FC) concepts was essentially developed by Löbner (1979, 1985, 1998). In the first phase of the project, we started out with manual countings of determination types for selected nouns of higher frequency in the corpus. This led to the results depicted in Table 2 and Figure 1 and helped to assess the significance of the types of determination as the relevant contextual cues for the automatic classification procedure.

Apart from these case studies, we developed a framework for the automatic classification of conceptual noun types consisting of three main components:

- manual annotation of concept types in a training corpus,
- automatic annotation of contextual features,
- machine learning of a classifier for automatic classification.

The first two components provide a training corpus for supervised learning in the third component. To support manual annotation we modified a text editor to insert XML tags corresponding to concept types by keyboard shortcuts. For the annotators, different concept types were visualized by different colors. As a starting point for further investigation we classified the occurrences of about 100 nouns in a German text corpus according to their underlying concept types and their respective uses. The corpus we used is Löbner's text (2003), which was chosen because it represents a scientific text, in which the effects of polysemy are reduced by observing terminological conventions for the application of many nouns frequently recurring in the text.

The automatic annotation of contextual features is based on deep morphosyntactic analysis. We use the software Machinese Syntax for German (http://www.connexor.com) to generate morphologically annotated dependency trees for our corpora. It turned out that the manual identification of some relevant grammatical features (e.g. determination, place of determination, possessive) was indispensable to adjust the existing software for our purposes, since some determiners in the software were given wrong definiteness values or were not defined at all as 'definite'or 'indefinite'. In a step-by-step process we managed to significantly improve the automatic identification of the appropriate grammatical features from less than 80% initially to ca. 95% by now. Further improvement of this rate is still in progress.

In a postprocessing step we filter out the contextual features considered relevant for the classification of concept types. While some features are comparatively easy to determine, such as grammatical number or the co-occurrence with certain determiners, others are hard to approach, in particular possessive constructions of various syntactic nature. We have to identify a considerably large set of possibilities that realize possessive constructions and can only do this by mapping regular expressions to dependency trees with the regular expressions en-

gine of the programming language Perl. Finally, we obtain a set of tuples <class, context> which serves as a training sample for machine learning. Typical instances are:

<individual, {tok=semantik, num=sg, det=def}>

<relational, {tok=teilgebiet, num=sg, det=indef, poss=rgen}>

<functional, {tok=bedeutung, num=sg, det=def}>

<sortal, {tok=buch, num=sg, det=indef}>

We then compute a classifier p(class|context) as the probability for a class assignment (some concept type) in a given context. Since contexts are made up of multiple attribute value pairs with arbitrary interdependencies and overlaps, we have to choose a model that can cope with such richly structured data. Maximum entropy (ME) models (Ratnaparkhi 1998) meet these requirements and result in classifiers, which reflect the specific amount of evidence that individual contextual features as well as their combinations contribute for the selection of a class in a given context. We implemented the computation of ME models in Prolog and used the algorithm Generalized Iterative Scaling (GIS) to approximate feature weights. An important subtask for ME models consists in feature selection, in order to eliminate infrequent features before feature weights are distributed with GIS. In our implementation, the representation of contextual features is based on bit vectors, allowing for efficient computation of overlappings between two vectors in terms of bit vector operations.

Classifiers ascertained from a training sample derived from a training corpus are then used to automatically classify test samples derived from test corpora. Every context of the test sample is assigned the concept type having the highest probability regarding the classifier.

The results of the first phase were encouraging, since the development of an automatic classifier with acceptable results now appears to be possible. The insights won from semantic analysis gave the experimental work an encouraging direction. On the other hand the corpus-linguistic approach to the semantic matter of concept types yielded interesting data for semantic analysis, such as the frequency of type shifts or of unexpected applications such as the bare singular use of non-mass FCs.

3 Goals and work schedule

3.1 Goals

The results of our investigations support our expectation that the distribution of some grammatical features (such as number, determination, possessivity) varies with certain conceptual noun types. When analyzing the differences of the distribution two approaches can be considered: on the one hand a statistic investigation can be conducted to sharpen the observation and obtain clues for the crucial factors responsible for the differences in the distribution. An automatic annotation of the relevant grammatical features, as well as the automatic classification of conceptual noun types (not possible with existing software) could help to process large amounts of data. On the other hand a semantic investigation of the respective noun types could also shed light on the varying distribution and explain why certain noun types occur more often in certain grammatical environments than in others and to what extent certain

kinds of determination occur with the respective noun types or even mark type shifts.

In this project we will continue to analyze the grammar and semantics of conceptual types of nouns (sortal, relational, individual, and functional) and develop a method for their automatic classification. With respect to the methods we use the project consists of two parts. The objects of investigation of the semantic part are the compositional properties of the different concept types, their unmarked uses, and the type shifts they undergo. These properties determine, in which contexts the different types of nouns occur.

On the other hand, the existing software will be developed to improve the automatic tagging of the relevant contextual cues. In contrast to our previous work we want to extend the annotations in our training corpora to both, underlying (lexical) concept types as well as specific readings in a context, which may include type shifts. This extends the classification task to pairs of types in the Cartesian product of types. From one perspective, this is what we have already done with respect to the decomposition of our types *individual*, *sortal*, *functional*, and *relational* into the underlying types *relational* and *functional*. We can also distinguish between '*underlying relational*' and '*contextual relational*', etc. In addition, we want to abstract away from disjunctive types or pairs of types to arrive at subsuming types, which cover two or more disjunctive types in a type lattice. The induction of type lattices can be conducted by exploring the probability spaces we have at the moment: in addition to maximum likelihood estimation we also have to consider the probability space below maximum likelihood to identify overlaps. Type lattices will lead to partially underspecified types. In terms of disambiguation decisions this is not appealing, but we expect deeper insights by exploring problem cases for the classification task. Furthermore, the association of types in a type lattice with contextual features will be compared with Formal Concept Analysis (FCA, Ganter et al. 2005), where formal concepts are arranged in lattices according to their features. Petersen (2008) uses FCA to induce type signatures for typed feature structures (Carpenter 1992). We will apply her method to our domain and try to extract implicational type constraints for conceptual noun types like those in constraint-based grammars. This will bridge a gap between our probabilistic methods used for classifier acquisition and the rule-based methods used in grammar engineering. We also expect a valuable feedback to our own work when exploring the grammar of conceptual noun types.

The type lattice account is more adequate than a plain type assignment of maximally distinguished types, because the type of a noun indicated by its context is often underdetermined. For example, if the noun *Bedeutung* occurs in the NP *die Bedeutungen des Wortes*, it can be concluded that it is used with a possessor phrase and is either RC or FC; it cannot be FC because it occurs in the plural with a singular possessor. But if we add a plural possessor specification such as *die Bedeutungen dieser beiden Wörter*, it cannot be determined here whether *Bedeutung* is RC or FC, since the phrase allows for both readings (each word may have one meaning or one or both may have more than one.) Thus, a type assignment of just [+P] would be adequate.

Our classifier has so far been a maximum entropy model (Ratnaparkhi 1998), in which the supervised computation with *Generalized Iterative Scaling* presupposes fully annotated training corpora. It turned out to be both difficult and expensive to manually create sufficient training data for our task. Therefore, we want to extend our classifier computation to a semi-supervised method using Yarowsky's algorithm (Yarowsky 1995, Abney 2004). This boot-

strapping method starts with a supervised classifier acquisition in the so-called inner loop and uses it to automatically annotate a new (larger) corpus, which can subsequently be used to learn a new classifier in the inner loop. This procedure can be iterated to fine-tune a classifier with the enriched contextual material provided by new corpora. We want to combine Yarowsky's algorithm with maximum entropy estimation in the inner loop. To our knowledge, this combination has not yet been reported in the literature, so we can contribute with our experiences.

Last but not least, we will evaluate the quality of the results of our automatic classification in terms of precision and recall in comparison to a gold standard corpus.

The goals of this project in detail:

1. To identify the possible uses of the different concept types and their specific context features. We will approach this task with the following sub-goals for our analysis of German and English texts:
   - 1.1 Our investigations for German have indicated that it is possible to identify groups of determiners, which are typical for the use of a noun according to its underlying concept type (cf. 2.0). If a noun occurs unmarked, a determiner of the respective group is used. Our first sub-goal consists in identifying all determiners and similar expressions to indicate the respective groups.
   - 1.2 The next step will be to analyze major type shifts that interfere with the classification task.

2. Develop and implement a method for the automatic classification of concept types in texts based on morphosyntactic features. We will
   - 2.1 create training corpora, in which concept types of nouns are manually annotated based on guidelines according to goal 1
   - 2.2 develop tools for the automatic annotation of contextual features in our training corpora
   - 2.3 compute classifiers from our training corpora with maximum entropy models and Yarowsky's algorithm in a semi-supervised fashion
   - 2.4 use these classifiers for the automatic classification of unseen texts and evaluate the quality of the results in terms of precision and recall
   - 2.5 extract from classifiers rules for constraint-based grammars that describe the grammar of conceptual noun types

3. Our third goal for the later phase of the project is to develop analogous machinery for French, in order to support the corpus-linguistic research in project A4. Having a more elaborate article system than German, the task might be easier for modern French; but if the method is applied to historical texts, we have to confront historical changes in the grammar of determination, such as stages of the gradual development of the present use of the definite article.

3.2 Methods and work schedule

The first goal will be pursued by conventional methods of semantic analysis, supported, and challenged by the corpus-linguistic evidence provided by the empirical part of the project. As the types of determination are of primary importance, a substantial part of the analysis is addressed to understanding determination. Which types of nouns are eligible for a given type of determination (e.g. definite singular, quantificational, or demonstrative) without type shift? What kind of type shifts are involved if the type of determination is applied to other types of nouns? What is the semantic effect of the determination, i.e. which is the resulting type of NP? For example, Löbner (1985) asserts that the result of the definite article is a [+D] concept, independently of the type of noun it is applied to.

We will use the same methodology consecutively to analyze the conceptual noun types and their environments in French. Finally, a comparison of the data for the different corpora should provide more information about the occurrences of the conceptual noun types.

As to the second goal, we will develop, maintain, and improve our software for the automatic classification, which is hybrid in that it involves rule based and statistical systems. The pre-statistical part of the framework requires rule-based determination of the relevant contextual features. The most important are grammatical number, definiteness of the NP, and the existence of a possessor specification. Experiments have shown that the automatic annotation of the grammatical features provided by the existing software still needs to be adjusted for our purposes, which will be a constant task. The input texts are annotated linguistically with Machinese Syntax for German, a deep syntactic analyzer by the Finnish company Connexor (www.connexor.com). The output is in the form of dependency trees. In addition, each running text token is assigned a part-of-speech tag (POS), a lemma, a shallow syntax (phrase) tag, and a set of relevant morphological tags. Here is an example:

(4) Input: *Thetys ist ein Mond des Saturn.*

Output of Machinese Syntax (in text format[3] with tab separated fields):

1	Thetys	thetys	subj>2	@NH Heur N SG NOM
2	ist	sein	main>0	@MAIN V IND PRES SG P3
3	ein	ein	det>4	@PREMOD DET Indef MSC SG NOM
4	Mond	mond	comp>2	@NH N MSC SG NOM
5	des	der	det>6	@PREMOD DET def MSC SG GEN
6	Saturn	saturn	mod>4	@NH N MSC SG NOM

Of the features we are interested in only grammatical number of the nouns is delivered directly by the analyzer. Definiteness is a feature with scope over the whole NP and is realized in German texts in different ways (e.g. in (4) the proper name *Thetys* is definite without an article, while the proper name *Saturn* is combined with a definite article). The same applies to possessive constructions. Currently, we are examining the relevance of some thirty types of definite constructions and some twenty types of possessive constructions for our investigation.

---

[3] The output can be in xml format. In this case we use XQueries to obtain the necessary information.

(5) a. *Marias Auto ist rot.* [prenominal genitive]
b. *Maria ist eine Freundin von mir.* [prepositional genitive]
c. *Maria hat ein rotes Auto.* [possessive relation with *haben*]
d. *Maria hat ein neues Auto. Die Sitze sind rot.* [associative anaphora]

Information on definite and possessive constructions exists only implicitly in the output of the analyzer. Thus, the output of Machinese Syntax is used as input for another postprocessing module that will add the relevant information on definiteness and possession explicitly to the noun, e.g.:

4 Mond mond comp>2 @NH N MSC SG NOM **i re:von**

"i re:von" stands for indefinite article, possessor von-PP to the right (indefiniter Artikel, rechter Possessor:von-PP). This means that the individual noun *Mond* 'moon' is assigned a relational noun reading on the basis of extracted contextual indicators.

The output of the postprocessing will serve as an input for the automatic classification module.

Another postprocessing module that takes the output of Machinese Semantics as input will determine the frequency of the occurrence of a given noun in different contexts in the given corpus. This will help determine the lexicalized conceptual type of the noun, the kinds of the type shifts, and the kinds of contexts that trigger or support these shifts. We expected that *Mond* would only occasionally appear in a context such as (4), where it is used as a relational concept. However, other nouns might not show such clear-cut distributions and thus suggest that a permanent type shift is in progress, which may lead to another lexicalized type, e.g. the use of the (proper) name of a particular brand of a product as a common noun to denote the product. This technique will be especially suitable for the diachronic investigation of type shifts.

Yet another postprocessing module will examine whether there are specific collocations of a certain noun with modifiers or verbs.

Our postprocessing modules are and will be implemented with the programming language Perl, which has a built-in engine for regular expressions that makes it superior to other programming languages. Although regular expressions are not powerful enough to recognize tree structures, a flat analysis of trees seems sufficient for our purposes since the vast majority of trees occurring in the corpus will be of very limited complexity.

The maximum entropy framework is implemented in Prolog and will be embedded in a bootstrapping framework based on Yarowsky's algorithm, which we will also implement in Prolog as well as an evaluation suite to assess the quality of the automatic classification.

Work Schedule

1st year
2008/2 and 2009/1

- corpus annotation with underlying and contextual concept types
- further investigation of the grammatical environments of the concept types and analysis of the type shift mechanisms in German

- compilation of an advanced set of context features and experiments with our software for automatic annotation
- maintenance of our maximum entropy framework
- extension to bootstrapping with Yarowsky's algorithm
- development of an evaluation suite

2nd year
2009/2

- analysis of the type shift mechanisms in German corpora; compilation of a profiling set of context features for German
- construction of type lattices for concept types
- extraction of constraint-based rules for concept types

2010/1

- investigation of the grammatical environments of the concept types in French texts and analysis of their most frequent uses; compilation of a first set of context features
- preparation of manually annotated training corpora in French and adjustment of the software for automatic annotation

3rd year
2010/2 and 2011/1

- investigation of the grammatical environments of the concept types in French texts with focus on type shifts
- compilation of an advanced set of context features for French and experiments with these

Publications and talks by the research group

Bontcheva, K. (1999). On Generics in Bulgarian and English. Second Finnish Colloquium of South-East European Studies in Helsinki, November 1999, 22nd-23rd.

Horn, C. (in prep.). Determination und Begriffstyp. Dissertation, Heinrich-Heine-Universität Düsseldorf.

Horn, C. (2007). *Grammatische Kontextmerkmale von Begriffstypen*. Tag des wissenschaftlichen Nachwuchses. Heinrich-Heine-Universität Düsseldorf.

Horn, C. & Rumpf, C. (2007). Conceptual noun types: semantics, grammar and automatic classification. Vortrag CTF 07. Universität Düsseldorf.

Löbner, S. (1979). *Intensionale Verben und Funktionalbegriffe*. Tübingen: Narr.

Löbner, S. (1985). Definites. *Journal of Semantics, 4*, 279-326.

Löbner, S. (1998). Definite Associative Anaphora. Retrieved Mar 04, 2008, from HHUD Institute for Language and Information Web Site: http://web.phil-fak.uni-duesseldorf.de/ ~loebner/publ/DAA-03.pdf

Löbner, S. (2003). *Semantik. Eine Einführung*. Berlin: de Gruyter.

Löbner, S. (2005). Begriffstypen aus logischer, kognitiver und grammatischer Sicht. Handout im Rahmen des Philosophischen Kolloquiums „Sprache und Weltbezug" (14.12.2005), TU Dresden.

Rumpf, C. (in prep.). Automatic classification of concept types. Dissertation, Heinrich-Heine-Universität Düsseldorf.

Rumpf, C. (2006). Ein Computermodell zur Bestimmung von Begriffstypen. Vortrag Tag der Forschung, Heinrich-Heine-Universität Düsseldorf.

Relevant Publications

Abney, S. (2004). Understanding the Yarowsky Algorithm. *Computational Linguistics 30(3)*.

Barsalou, L. (1992). Frames, Concepts, and Conceptual Fields. In A. Lehrer & E. V. Kittay (eds.), *Frames, Fields, and Contrasts*. Hillsdale, N.J.: Erlbaum.

Bierwisch, M. (1983). Semantische und konzeptuelle Repräsentationen lexikalischer Einheiten. In R. Ruzicka & R. Motsch (eds.), *Untersuchungen zur Semantik* (pp. 61-101), Berlin: Akademie-Verlag.

Carpenter, B. (1992). *The Logic of Typed Feature Structures*. Cambridge: Cambridge University Press.

Dölling, J. (1992). Flexible Interpretation durch Sortenverschiebung. In I. Zimmermann & A. Strigin (eds.), *Fügungspotenzen* (pp. 23-62), Berlin: Akademie-Verlag (Studia grammatica XXXIV).

Dölling, J. (1995). Ontological Domains, Semantic Sorts, and Systematic Ambiguity. In *International Journal of Human-Computer Studie,s 43*, 785-807.

Gale, W. A.; Church, K W. & Yarowsky, D. (1992). A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities, 26*, 415-439.

Ganter, B. & Stumme, G. & Wille, R. (Eds.) (2005). *Formal Concept Analysis: Foundations and Applications. Lecture Notes in Artificial Intelligence*, no. 3626, Springer.

Michaelis, L. A. (2004). Type shifting in construction grammar: An integrated approach to aspectual coercion. *Cognitive Linguistics, 15*(1), 1-67.

Marcus, M. P.; Marcinkiewicz, M. A. & Santorini, B. (1993). Building a Large Annotated Corpus of English: *The Penn Treebank. Computational Linguistics ,19*(2), 313-330.

Petersen, W. (2007). Representation of concepts as frames. In J. Skilters & F. Toccafondi & G. Stemberger (eds.), Complex Cognition and Qualitative Science. *The Baltic International Yearbook of Cognition, Logic and Communication*, (Vol. 2, pp. 151-170). University of Latvia.

Partee, B. H. (1983). Generalized conjunction and type ambiguity. In R. Bäuerle; C. Schwarze, A. von Stechow (eds.), *Meaning, Use and Interpretation of Language*. Berlin: de Gruyter.

Partee, B. H. (1986). Noun phrase interpretation and type shifting principles. In J. Groenendijk, D. de Jongh & M. Stokhof (eds.), *Studies in Discourse Representation Theory and the Theory of Generalized Quantifiers*. Dordrecht: Foris.

Petersen, W. (2008). Type Signature Induction with FCAType. To appear in: Lecture Notes in Artificial Intelligence, Springer.

Poesio, M. (2004). An empirical investigation of definiteness. *Proceedings of the International Conference on Linguistic Evidence*. Tübingen.

Poesio, M. & Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics, 24*(2), 183-216.

Ratnaparkhi, A. (1998). Maximum Entropy Models for Natural Language Ambiguity Resolution. PhD thesis, University of Pennsylvania.

Tapanainen, P. (1999). Parsing in two frameworks: finite-state and functional dependency grammar. PhD thesis, University of Helsinki.

Vieira, R. & Poesio, M. (2000). An Empirically-Based System for Processing Definite Descriptions. *Computational Linguistics ,26*(4), 525-579.

Vieira, R. (1998). *Definite Description Resolution in Unrestricted Texts.* PhD thesis, University of Edinburgh: Centre for Cognitive Science.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics.*

3.3 Experiments involving humans or human materials

☐ **yes** ☒ **no**

3.4 Experiments with animals

☐ **yes** ☒ **no**

3.5 Experiments with recombinant DNA

☐ **yes** ☒ **no**