Report

DFG reference number and title

KI 374/5-1 – Automatische Klassifikation von Nomen nach Begriffstyp

Project heads

Prof. Dr. Kilbury, James
Prof. Dr. Löbner, Sebastian

Staff

Dr. Katina Bontcheva (TV-L 13/2, EA, 9-10.2005)
Christian Horn M.A. (TV-L 13/2, EA, since 11.2005)
Christof Rumpf M.A. (TV-L 14/4, GA)

Student research assistants:

Pawel Sirotkin	until Aug 07, 10 hours/week
Stephane Onder de Linden	since Oct 07, 8 hours/week
Jan Bolten	since Oct 07, 5 hours/week

1 State of knowledge and goals

The four types of nouns to be distinguished are sortal nouns (SC)[1], individual nouns (IC), relational nouns proper (RC), and functional nouns (FC) (cf. Löbner 1979, 1985). 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; they are therefore in their unmarked uses combined with definite determination and in the singular.
- SC and RC nouns do not determine semantically the number of their potential referents in a given context: they may apply to zero, one or multiple instances; they are therefore 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 to be drawn between between nouns (as lexicalized concepts) and concept: a noun is a functional noun iff 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, in this context, in denoting, 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 are logically 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 is in accordance with 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 they are accompanied, or made possible, by a type shift of the noun that renders it an appropriate type. 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. In Löbner (1985) it is argued 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 therefore frequently used with possessive constructions (of various kinds). If they lack a possessor specification, they are in need of a special context that provides it. Conversely, SC and IC nouns require special context for possessive use.

Löbner (1998) amends these considerations through the analysis of associative anaphora. These require a noun which can be interpreted as a FC.

Our goals are:
The systematic investigation of the grammatical contexts of different noun types in order to identify contextual features that occur systematically with specific concept types.
Based on selections of features which we believe to be relevant for the identification of conceptual noun types, we will create training corpora for the machine learning of classifiers for conceptual noun types, where contextual features are weighted according to the evidence they give for the identification of noun types.

We will use these classifiers for the automatic classification of noun types in unseen text corpora.

2 Results and their significance

2.1 Noun types and contexts

In the first phase of the project, we started out with manual counting of determination types for selected nouns of higher frequency in our text corpus. The corpus we used was the text of Löbner (2003), because it represents a scientific text where the effects of polysemy are reduced by observing terminological conventions for the use of many nouns that are frequent in the text. The corpus contains about 110.000 words in German. For our investigation, we first developed an environment for the manual annotation of types of nouns for intellectual annotators to create annotated corpora in a relational database. We classified about 100 nouns in a German text corpus according to their underlying concept types and their respective uses as a starting point for further investigation. The identification and annotation of the grammatical contexts was first conducted manually only, in order to check if our software would deliver correct results. Since the appropriate output of the software is an indispensable prerequisite for further analysis, we concentrated on an accurate annotation of the respective nouns in the corpus. We annotated the nouns according to their kind of definite/indefinite determination (considering determiners, quantifiers, prepositions, demonstratives, pronouns, etc.), their number and their possessive use (such as prenominal genitive, prepositional genitive, the possessive relation with verbs like *haben,* generic and anaphoric possessives etc.). Then we used the software Machinese Syntax (www.connexor.com, cf. Tapanainen 1999) to annotate morphosyntactic features in the same corpus. As expected, it turned out that the manual identification of some relevant grammatical features (e.g. determination, place of determination, possessive) was essential to adjust the existing software for our purposes since some determiners in the software were given wrong definiteness values or were not defined as ‘definite’ or ‘indefinite’ at all. In a step-by-step process and on the basis of our manual annotations we managed to significantly improve the automatic identification of the appropriate grammatical features from less than 80% in the beginning to about 95% now. Further improvement of this rate is still expected.

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

SORTAL NOUN

Nomen ("noun"), 166 tokens

definite, singular	**37 %**
possessive	**0%**

INDIVIDUAL NOUN

Semantik ("semantics"), 152 tokens

definite, singular	**82 %**
possessive	**4%**

RELATIONAL NOUN

Teil ("part"), 124 tokens

definite, singular	**36 %**
possessive	**73%**

FUNCTIONAL NOUN

Bedeutung ("meaning"), 721 tokens

definite, singular	**57 %**
possessive	**74%**

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

The table well 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 singular use without any article 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):

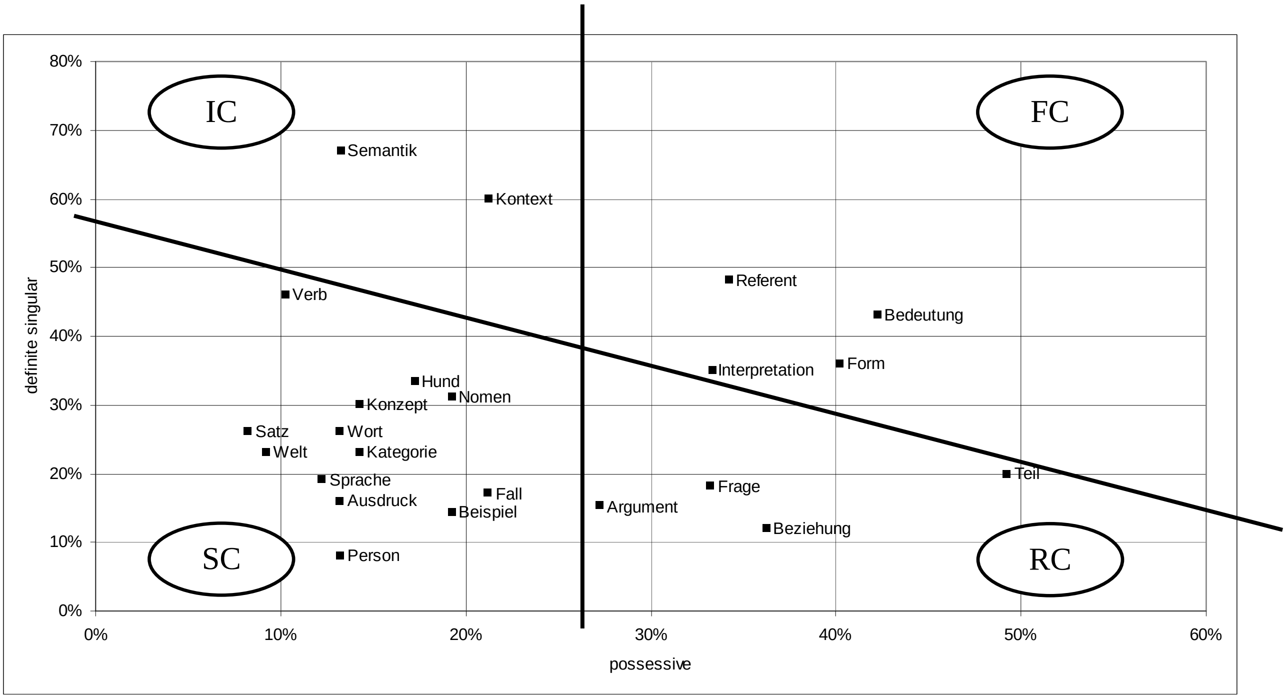

Figure 1: grammatical environments of the most frequent nouns in the corpus (automatic annotation)

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. 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.

2.2 Automatic classification

We developed a framework for the automatic classification of conceptual noun types consisting of three main components:

- manual annotation of concept types,
- 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 are visualized by different colors. The corpus we used so far is the book Löbner (2003).

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. In a post-processing step we filter out the contextual features which we declared relevant for the classification of concept types. Especially in the case of possessive constructions this is non-trivial. We have to identify a considerably large set of possibilities that realize possessive constructions and do this by mapping regular expressions to dependency trees with the regular expressions engine of the programming language Perl. Finally, we get 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 resemble 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* to filter infrequent features out before feature weights are distributed with GIS. In our implementation, the representation of contextual features is based on bit vectors, which allows the efficient computation of overlappings between two vectors in terms of bit vector operations.

During the development of the framework we created some additional analysis tools that support *ngram* views of the corpus data as well as unified views of the results of all analysis steps in a relational database for further analysis, error detection, and correction.

The classifier learned from a training sample derived from a training corpus is then used to automatically classify test samples, which are derived from test corpora. Every context of the test sample is assigned the concept type with the highest probability regarding the classifier.

Although we did not evaluate our system in terms of precision and recall (which will be one of our next steps), our expectations concerning the performance of our approach were met.

3 Relation of work schedule to outcome

We have achieved the main goals formulated in our application to a great extent: We examined the grammar and contextual features of conceptual noun types, constructed training corpora, and implemented a framework for the automatic classification of noun types.

At some points we experienced unexpected difficulties where we had to focus on the underlying problems: The necessity of a clear distinction between underlying and contextual readings of noun types for manually annotating a training corpus became apparent at the very start of our work. This led to systematic changes in corpus construction as well as in our perspective about which classifier should be learned for automatic classification. For the extraction of contextual features we experienced serious problems identifying all possessive constructions. We had to accept that a complete solution to this problem would be unrealistic, last but not least because of associative anaphora as possessive constructions. For a partial solution to this problem we decided to use full parsing instead of shallow processing as planned in our application.

Instead of using Yarowsky’s algorithm (Yarowsky 1995, Abney 2004) as described in our application, the computation of our classifiers has so far been based on maximum entropy models (Ratnaparkhi 1998). This is a prerequisite for making Yarowsky’s algorithm work, since it depends on some supervised learning method in its inner loop. We will combine it with maximum entropy models in the next project phase.

4 Cooperation

4.1 within the research unit

Albert Ortmann (project A1); Thomas Ede Zimmermann, Magdalena Schwager (project A3); Hans Geisler, Doris Gerland (project A4); James Kilbury, Gerhard Schurz, Wiebke Petersen (project B1)

4.2 with external partners

Prof. Dr. Gerhard Jäger, Lehrstuhl Theoretische Linguistik der Fakultät für Linguistik & Literaturwissenschaft, Universität Bielefeld
Prof. Dr. Anette Rosenbach, Institut für Anglistik und Amerikanistik, Fakultät für Kulturwissenschaften, Universität Paderborn
Maria Cieschinger, Institut für Kognitionswissenschaft, Universität Osnabrück

5 Publications and activities

Publications and talks

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. (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.

Further activities

Christian Horn was member of the organizing committee of our FFF Conference ‚Concept Types and Frames CTF07’ (august 10th-22nd) in Düsseldorf. Website of the conference: http://www.phil-fak.uni-duesseldorf.de/fff/ctf/.

References

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

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

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. ms. http://web.phil-fak.uni-duesseldorf.de/ ~loebner/publ/DAA-03.pdf

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.

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