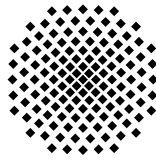


Geplanter Sonderforschungsbereich 732

Incremental Specification in Context

Universität Stuttgart



Finanzierungsantrag
2006/2 – 2010/1

Inhaltsverzeichnis

D5, Biased, Schütze	3
Schütze, Hinrich, Ph.D., Prof.	34

3.1 Allgemeine Angaben zum Teilprojekt D5

3.1.1 Titel: Biased Learning for Syntactic Disambiguation

Kurztitel: Biased

3.1.2 Fachgebiete und Arbeitsrichtung:

Computational Linguistics, Natural Language Parsing, Machine Learning, Knowledge Acquisition

3.1.3 Leiter:

Prof. Schütze, Ph.D., Hinrich, 1964
Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Azenbergstraße 12
D-70174 Stuttgart

Telefon: +49-(0)711-121-1400
Telefax: +49-(0)711-121-1366
E-Mail: schuetze-sfb732@ims.uni-stuttgart.de

Ist die Stelle des Leiters des Projektes befristet?

- nein ja, befristet bis zum _____
 eine weitere Beschäftigung ist vorgesehen bis zum _____

3.1.4 In dem Teilprojekt sind vorgesehen:

- Untersuchungen am Menschen oder am menschlichen Material ja nein
Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei ja nein
- klinische Studien im Bereich der somatischen Gentherapie ja nein
- Tierversuche ja nein
- gentechnologische Untersuchungen ja nein
- Untersuchungen an humanen embryonalen Stammzellen ja nein
Die gesetzliche Genehmigung liegt vor ja nein

3.1.5 Beantragte Förderung des Teilprojektes im Rahmen des Sonderforschungsbereichs (Ergänzungsausstattung)

Haushaltsjahr	Personalmittel	Sachmittel	Investitionsmittel	Gesamt
2006/2	53,1	5	22	80,1
2007	106,2	–	–	106,2
2008	106,2	–	–	106,2
2009	106,2	–	–	106,2
2010/1	53,1	–	–	53,1

(Beträge in Tausend EUR)

3.2 Zusammenfassung

Short summary. D5 investigates syntactic disambiguation as a particular type of specification. An ambiguous sentence is processed by a treebank-trained parser. A subset of the readings are extracted from the analysis. The most likely parse is then identified based on contextual knowledge acquired by way of biased learning from (monolingual and multilingual) unannotated corpora. We use the framework of Exemplar Theory for relating the ambiguous sentence to similar contexts in the unannotated corpora. Two similarity measures are considered: a language model as a baseline and a measure based on grammatical dependencies.

Extended summary. D5 investigates syntactic disambiguation as a particular type of specification. From the ambiguous output of a treebank-trained parser (delivered in the form of the n best parses), we identify the most likely parse based on contextual knowledge. All lexical, syntactic, semantic, pragmatic and world knowledge that a listener can use to interpret an utterance is viewed as context. One of the aims of the project is to model this linguistic and extralinguistic knowledge in the form of a database compiled from large unannotated text corpora.

As an example for how D5 will bring contextual information extracted from corpora to bear on the problem of disambiguation consider the phrase *an opening under the house that led to a fume-filled coal mine*. In an unannotated corpus, one can find similar contexts that mention an opening leading to a coal mine, but none that talk about a house leading to a coal mine. This suggests that the relative clause attaches to *opening*, not to *house*.

The project uses machine learning methods that are informed and guided by linguistic knowledge. We call these methods *biased learning*. Biased learning will be used for merging syntactic knowledge from the treebank parser with contextual knowledge from the database using the framework of Exemplar Theory. In this framework two similarity measures will be defined which, given an ambiguous parse, determine the most similar fragments in the exemplar database compiled from the corpus. The disambiguation decision will be based on these similar fragments. The first similarity measure is based on language models, the second on dependency structures. For the acquisition of the exemplar database both monolingual and multilingual parallel corpora will be exploited.

This project intends to contribute to the goals of the SFB by determining the effect of context on syntactic disambiguation; by investigating the different effects of contextual information acquired by means of shallow vs. deep analysis; by investigating the effect of

contextual information on disambiguation quantitatively; by investigating the effect of the domain “language” on syntactic disambiguation (English vs. German); by investigating the learnability of contextual information from monolingual and multilingual corpora; and by showing that complex statistical models based on Exemplar Theory can provide meaningful linguistic explanations.

Deutsche Zusammenfassung. D5 untersucht syntaktische Disambiguierung als einen Spezialfall von Spezifikation. Aus der ambigen Ausgabe eines auf einer Baumbank trainierten Parsers (den n besten Analysen) wird die beste Analyse basierend auf kontextuellem Wissen identifiziert. Als Kontext wird alles lexikalische, syntaktische, semantische, pragmatische und Weltwissen betrachtet, das vom Hörer zur Interpretation einer Äußerung herangezogen werden kann. Eines der Ziele des Projekts ist die Modellierung dieses linguistischen und extralinguistischen Wissens in Form einer Datenbank, die aus großen Textkorpora extrahiert wird.

Wie in D5 solche aus Textkorpora extrahierte kontextuelle Information zur Disambiguierung benutzt werden wird, kann anhand der Phrase *an opening under the house that led to a fume-filled coal mine* verdeutlicht werden. In einem unannotierten Korpus kann man ähnliche Kontexte finden, die eine Öffnung, die zu einer Kohlegrube führt, beschreiben, aber keine, die ein Haus, das zu einer Kohlegrube führt, erwähnen. Daraus kann man schließen, dass der Relativsatz sich auf *opening* bezieht, und nicht auf *house*.

Das Projekt benutzt Methoden des maschinellen Lernens, die durch linguistisches Wissen gelenkt werden. Wir nennen diese Methoden „wissensgesteuertes“ (biased) Lernen. Wissensgesteuertes Lernen wird benutzt, um syntaktisches Wissen aus einer Baumbank mit kontextuellem Wissen aus einer Datenbank zu verschmelzen. Dabei wird der theoretische Rahmen der Exemplartheorie verwendet. In diesem Rahmen wird ein Ähnlichkeitsmaß definiert, das für ein ambiges Sprachfragment die ähnlichsten Fragmente in der Exemplardatenbank, die aus dem Korpus gewonnen wurden, ausfindig gemacht. Die Disambiguierungsentscheidung basiert auf diesen ähnlichen Fragmenten. Für die Akquirierung der Exemplardatenbank werden sowohl monolinguale als auch multilinguale parallele Korpora benutzt. Auf die Korpora werden Werkzeuge wie Dependenzparser und Alignierungsalgorithmen angewendet.

Dieses Projekt beabsichtigt, folgendermaßen zu den Zielen des SFB beizutragen.

- Bestimmung des Einflusses von Kontext auf syntaktische Disambiguierung (die wir als Spezialfall von Spezifikation ansehen). Kontext ist als das kontextuelle Wissen definiert, das ein Hörer beim Versuch, eine Äußerung zu verstehen, anwendet.
- Untersuchung des Unterschieds zwischen zwei Arten von Kontext: kontextuelle Information extrahiert mittels einer flachen Analyse (language models) vs. kontextuelle Information extrahiert mittels eines Dependenzparsers.
- Untersuchung des Unterschieds zwischen linguistischem und extralinguistischem Kontext in der Disambiguierung
- Quantitative Analyse des Einflusses kontextueller Information auf Disambiguierung

- Untersuchung des Einflusses der Domäne *Sprache* auf syntaktische Disambiguierung im Kontext: Wie unterscheiden sich syntaktische Disambiguierung im Englischen und im Deutschen?
- Untersuchung der Lernbarkeit disambiguierungsrelevanter kontextueller Information aus monolingualen und multilingualen Korpora.
- Die statistische Sprachverarbeitung leidet unter zu stark simplifizierenden Modellen wie Hidden-Markov-Ketten und *bag of word* Modellen. Diese Modelle sind nur dann erfolgreich, wenn die Einfachheit des Modells durch große Trainingskorpora kompensiert wird. In manchen Gebieten (z.B. Spracherkennung) sind Trainingskorpora relativ billig, so dass eine solche Kompensierung möglich ist. Trainingskorpora für syntaktische Disambiguierung sind jedoch teuer, so dass die einzige Verbesserungsmöglichkeit in einem besseren Modell besteht. Wir hoffen, dass das exemplartheoretische Modell, das hier vorgeschlagen wird, den gegenwärtigen einfachen Modellen überlegen sein wird, und damit ein Fortschritt beim Stand der Forschung in der natürlichen Sprachverarbeitung sein wird.

3.3 Ausgangssituation des Teilprojekts

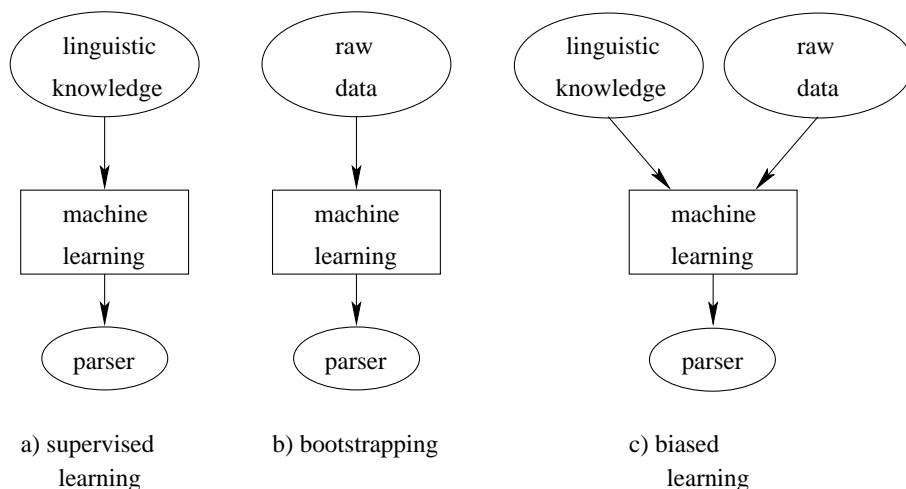


Figure D5.1: Supervised learning, bootstrapping and biased learning.

3.3.1 Stand der Forschung

We start with the observation that there are three approaches to training a statistical parser. The most common approach is supervised learning as depicted in Figure D5.1 (a). In supervised learning, the machine learning algorithm estimates the parameters of a model from linguistically annotated data, typically a treebank. This data source is labeled “linguistic knowledge” in the figure because we can view supervised learning as transferring a theory about the syntactic structure of sentences to the parser by (1) annotating the treebank according to the syntactic theory and (2) training the parser on the treebank.

A different approach to creating a statistical parser is bootstrapping (Figure D5.1, b). Here there is no linguistic knowledge involved in training the parser. Instead, the machine learning algorithm detects patterns in large amounts of raw data. For example, distributional clustering will identify the class of nouns as a category based on local co-occurrence patterns of words (Schütze, 1995). In a second step, syntactic phrases can then emerge as local patterns of the categories that are induced in the first step. Bootstrapping is appealing because it doesn't require expensive lexical resources. Many languages still don't have such resources, so that supervised learning is not an option for them.

In this proposal, we choose the third approach, biased learning, shown in (Figure D5.1, c). It combines the advantages of supervised learning and bootstrapping without being afflicted with their problems. The advantage of bootstrapping is that a virtually unlimited amount of data is available. In general, the more data a statistical parser is trained on, the better its performance will be. In contrast, the availability of training data is limited in supervised training since linguistic annotation of text is expensive. The disadvantage of bootstrapping is that insights from linguistic theory are not available to the machine learning algorithm. As a result, bootstrapped parsers often commit very basic errors that could be easily fixed if fundamental laws of language were taken into account. For example, they will construe "Monday" in "He arrived Monday" as a direct object of "arrive" because it is a noun right after a verb, a position typical of direct objects.

Supervised learning. Supervised training of statistical parsers on treebanks has been an active area of research in natural language processing for some time. After the creation of the Penn Treebank (Marcus et al., 1993), researchers at Stanford (Magerman, 1994), the University of Pennsylvania (Collins, 1997), Brown University (Charniak, 1997) and many other institutions published increasingly sophisticated parsing models with steadily increasing performance numbers. However, performance improvements have recently leveled off (Bikel, 2004; Klein & Manning, 2003). We believe that the information available in the treebank is being fully exploited by current parsers. Further improvements will either require larger treebanks or make use of non-treebank information. This is an important motivation for project D5: Supervised learning in statistical parsing has reached an impasse. Our goal is to show that this impasse can be broken by combining current supervised approaches with unsupervised learning from unannotated data.

A more recent approach to supervised parsing is discriminative reranking (Charniak & Johnson, 2005; Collins & Koo, 2005; Riezler et al., 2002). Reranking has the potential of incorporating more linguistic knowledge into the training process because complex linguistic features can be constructed, based on the entire parse tree, and then exploited by a discriminative classifier such as logistic regression. Such features are particularly important for German because they can capture complex non-local phenomena of German syntax such as the extraposition of relative clauses and the alternation of the finite verb in second vs. final position. Such non-local phenomena are hard to integrate into generative frameworks, but can easily be captured through features since they are extensively studied in the linguistic and computational linguistic literature (e.g., Uszkoreit et al. (1998); Frank et al. (2003)). Indeed, discriminative reranking has been successfully applied to German by Rohrer & Forst (2006). However, it is doubtful that current treebanks are large enough to support training on such relatively rare features. So despite their potential for more linguistic sophistication, discriminative approaches also need a strong unsupervised component to further improve parsing accuracy.

Supervised learning can also be applied to correcting attachment decisions of an existing parse. Siddharthan (2002a,b) trains a classifier on a training set to predict the correct attachment of a relative clause with multiple attachment points. Yeh & Vilain (1998) use a similar approach for both relative clause and prepositional phrase attachment.

Bootstrapping. Bootstrapping of syntactic models also has a long tradition. Most researchers worked on monolingual corpora (Finch & Chater, 1992; Schütze, 1993a, 1995; Klein & Manning, 2004; Solan et al., 2005), but there is also an increasing number of papers on inducing linguistic models from multilingual corpora (Dumais et al., 1997; Schütze, 1993b; Wu, 1997; Kuhn, 2004). It is important to note that the boundary between supervision and bootstrapping is somewhat blurry for aligned multilingual corpora. Approaches that do not use any linguistic knowledge for statistical machine translation (e.g., Ney’s group at RWTH Aachen, Zens & Ney 2004) are still supervised since the training material consists of correct pairs of input and output sentences. However, they can only be regarded as supervised for the task of machine translation. For the task of parsing, there is no supervision since there is no “output” in the form of correct parses present in the multilingual corpora.

The presentation in Figure D5.1 is somewhat simplistic because there are in fact two different ways of using prior knowledge in machine learning. In addition to the prior knowledge that is encoded in the training set (e.g., the detailed syntactic parses of sentences in the Penn treebank), one can also bias learning by specifying a model structure that favors linguistically plausible parses. For example, the model in Klein & Manning (2004) favors left- or right-branching tree structures as opposed to trees with deep center embedding. Similarly, bilingual grammar induction often posits a small set of possible syntactic correspondences between two languages (e.g. head-initial vs. head-final (Wu, 1997; Kuhn, 2004)). However, the amount of linguistic knowledge encoded in the statistical model (as opposed to the training data in supervised learning) is negligible for the purposes of learning parsers with good coverage of the syntactic phenomena of a natural language.

A note on terminology: We will use the terms *bootstrapping* and *unsupervised learning* interchangeably in this document. As pointed out above, even unsupervised learning cannot proceed without a minimal model that captures some knowledge about the domain. We have used the term *bootstrapping* in this section to stress that the learning procedures discussed under that rubric use no or almost no linguistic knowledge and are in that sense “non-linguistic.”

Biased learning. Given the obvious benefits of biased learning, it is surprising how little prior work has taken this approach. The only prior approaches that combine supervised and unsupervised learning and show an increase in performance for parsing seem to be Charniak (1997) and Johnson & Riezler (2000). However, the performance increase achieved is small and lacks significance in this prior work. Our aim is to show that biased learning can improve parsing performance significantly.

There is a large body of literature on PP-attachment (e.g. Volk, 2001; Calvo et al., 2005) that shares the overall goals of this proposal: using information from unannotated corpora for syntactic disambiguation. Volk (2001) counts the number of occurrences of word n-grams on the web to select the correct attachment of PPs. We believe that a more complex

analysis of corpora (i.e., extraction of dependencies) will provide better information for syntactic disambiguation than a simple frequency heuristic.

The early work by Hindle & Rooth (1991, 1993) is similar to our approach in that dependency statistics are extracted from an unannotated corpus and used for syntactic disambiguation. However, the system, while pioneering, was limited to one particular type of ambiguity, PP attachment. Our interest is in developing a framework that can disambiguate syntactic ambiguities in general, as opposed to solving a particular syntactic ambiguity problem.

A method of combined supervised and unsupervised learning different from the biased learning approach proposed here is *colearning* (Blum & Mitchell, 1998). *Colearning* relies on a clustering component to form groups of similar examples. These groups (learned by way of unsupervised learning) are then combined with supervised models. This approach has been successfully applied to many problems with relatively simple similarity spaces (e.g., text categorization). But it is unlikely that *colearning* would be applicable to the much more complex similarity spaces typical of linguistic data.

Another combination of supervised and unsupervised learning is *multiview learning* (Muslea et al., 2002). *Multiview learning* attempts to identify orthogonal aspects of the representation of unannotated “exemplars.” If two classifiers that are each trained on different aspects agree on an unannotated exemplar, then the resulting label is assumed to be correct. *Multiview learning* can improve classification accuracy considerably on small training sets (10–50 examples), but it has little or no benefit for training sets of 50 or larger. Since training sets tend to be larger in syntactic disambiguation, *multiview learning* would not be expected to be a promising method for the type of NLP problem investigated here.

Exemplar Theory. Exemplar Theory (Lacerda, 1995; Kirchner, 1999; Pierrehumbert, 2001) is a theory of speech perception and production. All percepts of speech events are stored in memory as exemplars in a perceptual space. This space can be represented as a multi-dimensional cognitive map with individual dimensions corresponding to phonetic and phonological properties of the exemplars. Percepts of nearly identical instances are located close to each other, whereas percepts of less similar instances are located in different regions. Thus, perceived realizations of speech events form clouds of exemplars on the map (Pierrehumbert, 2001). These exemplar clouds represent the phonetic and phonological categories of a given language. Within each category the distribution of exemplars indicates the range of variation of the category’s parameters. The cognitive system assumed by Exemplar Theory can thus be described as a mapping from locations in the perceptual space to labels of the language’s category system. Frequency of occurrence, or frequency of experience, are crucial factors in Exemplar Theory since the location of the center of a category and the boundaries between categories are influenced by the frequency (and salience) of exemplars.

Most statistical models used for language are either motivated by engineering considerations (e.g., Hidden Markov models) or are simple frequency models (e.g., Bybee’s model of French liaison (2001)). Except for some versions of Optimality Theory, Exemplar Theory is perhaps the only statistical theory of an aspect of language that is at the same time mathematically sophisticated and cognitively explanatory.

In this project, we want to learn from the success of Exemplar Theory in phonetics and phonology and apply it to the problem of syntactic disambiguation. For this pur-

pose, we conceptualize text corpora as collections of examples of context (or exemplars) that were experienced by an imaginative listener for particular words, phrases and sentence fragments. This proposal describes the first phase of this project, which is concerned with formulating and testing the basic idea of formalizing syntactic disambiguation as an exemplar-theoretic process. Future extensions of the project would then deal with transferring the specific predictions Exemplar Theory makes from phonetics and phonology to syntactic disambiguation.

An example would be to show the effect of prototypicality effects in disambiguation. A less prototypical adjective like “fun” in “a fun game” would be predicted to behave differently from a prototypical adjective like “beautiful”. The acceptability of *marginal* adjectives like “fun” and “key” in typical adjective positions (e.g., “this problem is key”) should improve after a subject has stored exemplars that reinforce their adjectival nature. The precise form of these predictions and the experiments that would verify them will depend on the outcome of this project.

Nearest neighbor methods. Exemplar Theory bears some resemblance to nearest neighbor methods in that both approaches determine category membership based on a local neighborhood of similar examples. In contrast to kNN and classical Exemplar Theory, the database of exemplars that is searched for nearest neighbors in our version of Exemplar Theory consists of *unlabeled* exemplars (based on data from large unannotated corpora). This “unlabeled” neighborhood is used to predict the correct reading of the ambiguous query.

Several research groups have worked on nearest-neighbor approaches to parsing (Bod et al., 2003; Argamon et al., 1998). Again, the approach in D5 differs in that the contextual information implicit in *unannotated* text is used. The basic philosophy of the approach is that grammar and context interact in natural language understanding. In our testbed, the model for the grammar is the treebank parser, the model for the context is the unannotated corpus.

Kübler & Hinrichs (2001) and Hinrichs (2005) also develop a nearest neighbor method, which they call *holistic memory-based learning* (MBL). This holistic approach is similar to our approach in that it tries to find similar fragments that are large (as opposed to restricting itself to the local environment of the syntactic ambiguity). Hinrichs (2005) creates tools for automatically annotating text corpora using holistic MBL. Treebank-parsing is a possible application of these annotated corpora, but the focus is on corpus annotation. As with other nearest neighbor methods, exemplars have to be labeled in MBL. We pay special attention to the problems of adapting nearest neighbor methods for the type of noisy and partially non-disambiguated data sets that partial parsers produce.

Language models. Language models are a well studied area of research in NLP (Manning & Schütze, 1999). However, adapting standard language models to the syntactic disambiguation framework proposed here poses a number of difficulties. In particular, if the “contextual plausibility” of different readings of a sentence are to be compared based on different parses, then a method for parse-based segmentation has to be developed. If the language model were to be applied to a sentence as a whole, then different parses would receive the same language model score and would not be differentiable for the purpose of disambiguation. This problem has not been addressed before as far as we know.

Acquisition from monolingual corpora There is a large literature on acquisition from unannotated corpora, but most of it has not been conducted in the context of treebank-trained parsers. For example, Carroll & Rooth (1998) acquire subcategorization frames from a corpus parsed with a PCFG. Schulte im Walde (2003) also uses a PCFG to parse corpora for the acquisition of information about verbs. The resulting feature representation of a verb provides a fine-grained characterization of the verb’s behavior. Verbs are then semantically clustered based on the assumption that the lexical meaning of a verb is related to its syntactic behavior (the argument structure of the verb, the prepositions of prepositional complements, lexical selectional preferences etc.). The focus of this line of research has been to create lexical resources as opposed to enhancing a treebank-trained parser. Treebank parsers consistently outperform parsers trained without the benefit of a treebank.

Parallel corpora. Hwa et al. (2005) bootstrap parsers for Spanish and Chinese from the information inherent in parallel corpora. They create noisy annotations of the Spanish and Chinese versions of the corpus by transferring a syntactic analysis of the English version to the aligned “second” languages. The Spanish and Chinese parsers are then trained on these noisy annotations. Hwa et al.’s goal is the creation of treebanks for languages without such a lexical resource. In contrast, our goal is to mutually disambiguate languages for the purpose of improved syntactical parsing.

Languages other than English. For languages other than English, there are treebanks for Czech (Hajič, 1998), Chinese (Xue et al., 2002, 2005), Dutch (van der Beek et al., 2002), German (Skut et al., 1998), French (Abeillé et al., 2003), and Spanish (Moreno et al., 2000). A number of researchers have created parsers based on these treebanks: Collins et al. (1999) for Czech, Bikel & Chiang (2000) and Chiang & Bikel (2002) for Chinese, Malouf & van Noord (2004) for Dutch, Dubey & Keller (2003) and Schiehlen (2003b) for German, Arun (2004) for French, and Moreno et al. (2000) for Spanish. However, to our knowledge the state of the art is the same for these languages as for English as far as the combination of supervised and unsupervised training is concerned. There is as little research on what we call biased learning for these languages as for English.

3.3.2 Eigene Vorarbeiten

The IMS offers an excellent infrastructure for corpus-based work. The HGC corpus (“Huge German Corpus”, about 240,000,000 tokens, mainly newspaper texts) was compiled here and several other large corpora including the .de corpus (Markowetz et al., 2005) and the Reuters corpus (Lewis et al., 2004) are available. The IMS has developed a large number of tools for corpus-based work that will be used in D5: taggers (Schmid, 1994, 1995), chunkers (Schiehlen, 2003b,a), statistical parsers (Schmid, 2004a; Schiehlen, 2004) and other tools.

The IMS also has a long tradition of extracting lexical knowledge from unannotated corpora (Kermes & Heid, 2003; Schulte im Walde, 2003; Evert, 2004). In D5, the intended application of the acquired knowledge is syntactic disambiguation (as opposed to building lexical resources) and the analysis mode is partial analysis (as opposed to tagging or PCFGs), but this prior work will be important in designing and implementing the acquisition procedures proposed here.

An exploratory study on exemplar-theoretic biased diambiguation has been submitted for publication by Atterer & Schütze (2005). The search for fragments in an unannotated corpus that are similar to an ambiguous fragment is formalized in a lattice framework. The search first retrieves the most specific fragment available (in the most extreme case this would be a fragment identical to the ambiguous fragment) and then proceeds to increasingly less specific similar fragments. Experiments are conducted for relative clause attachment. The two noun phrases that are possible attachment sites for a relative clause are made less specific by stripping them of modifiers and generalizing their head nouns (e.g. “American Bell telephone Co.” → “company”). The relative clause is made less specific by removing the object of the main verb. The elements produced by various combinations of these “despecification operations” form a lattice that has most specific (original) context as its supremum and the empty context as its infimum. For each tuple in the lattice the corpus is searched and a mutual information score is calculated that evaluates how many similar constructions occur in the corpus. The attachment site with the highest score for any of the tuples is chosen. Compared to a baseline of always attaching low, a considerable improvement in disambiguation performance is demonstrated on the Penn Treebank for *which*-clauses and *that*-clauses.

The most important formal element of Exemplar Theory is the similarity measure between exemplars. When designing this similarity measure we can draw on knowledge and experience gained from previous work on similarity based on surface syntactic features (Schütze, 1995), on similarity based on cooccurrence features (Schütze, 1998), and on measuring the similarity of data objects that consist of several disparate data sources (Schuetze et al., 2005a,b).

Hull et al. (1996) investigate methods for combining assessments provided by multiple text classifiers. A machine learning based framework is proposed for computing weights for the combination method. This framework can serve as a starting point in this investigation since we most likely will integrate different types of information sources in the similarity measure.

3.3.3 Liste der publizierten einschlägigen Vorarbeiten

Referierte Veröffentlichungen – in wissenschaftlichen Zeitschriften

Schütze, H., 1998. Automatic word sense discrimination. *Computational Linguistics* **24** (1), 97–124.

Referierte Veröffentlichungen – auf wesentlichen Fachkongressen

Hull, D. A., Pedersen, J. O., Schütze, H., 1996. *Method combination for document filtering*. In: SIGIR '96. 279–287.

Schütze, H., 1995. *Distributional part-of-speech tagging*. In: EACL 7. 141–148.

Schütze, H., 1993a. *Distributed syntactic representations with an application to part-of-speech tagging*. In: Proc. of the IEEE International Conference on Neural Networks. 1504–1509.

Schütze, H., 1993b. *Translation by confusion*. In: B. Dorr (ed.), *Working Notes of the AAAI Spring Symposium on Building Lexicons for Machine Translation*. AAAI Press, Menlo Park CA, 82–85.

Referierte Veröffentlichungen – in Monographien

Manning, C. D., Schütze, H., 1999. *Foundations of statistical natural language processing*. MIT Press, Boston, MA.

Zur Publikation eingereichte Arbeiten

Atterer, M., Schütze, H., 2005. *A lattice-based framework for unsupervised syntactic disambiguation with an application to relative clause attachment*, submitted.

Patente

Schuetze, H., Chen, F. R., Pirolli, P. L., Pitkow, J. E., Chi, E. H., Li, J., 2005a. *System and method for identifying similarities among objects in a collection*. United States Patent 6,941,321.

Schuetze, H., Chen, F. R., Pirolli, P. L., Pitkow, J. E., Chi, E. H., Li, J., Gargi, U., 2005b. *System and method for quantitatively representing data objects in vector space*. United States Patent 6,922,699.

3.4 Planung des Teilprojekts (Ziele, Methoden, Arbeitsprogramm)

3.4.1 Fragestellung

In D5, we view context as *contextual knowledge* – linguistic knowledge (lexicon, syntax etc.) and extralinguistic knowledge (encyclopedic or world knowledge) about the current situation that the listener uses to interpret an utterance. Our basic hypothesis is (1) that situations that were previously experienced and are similar to the current one are an important part of contextual knowledge; and (2) that contextual knowledge can be modeled with fragments and sentences extracted from unannotated corpora.

Specifically, the approach in D5 is to use *acquisition parsers* to extract contextual knowledge from unannotated *acquisition corpora*, store this information in a *context database* (C-DB) and then incorporate it into a *target parser*, which is trained on a *treebank corpus*. The overarching goal is to develop a method that can take advantage of information mined from unannotated corpora to improve parsing performance of parsers trained on a treebank. Our approach is thus a combination of supervised and unsupervised machine learning, an instance of biased learning as defined in Figure D5.1 (c).

The main difficulty in executing the program of biased learning is the formalization of how to merge syntactic knowledge extracted from the treebank with contextual knowledge extracted from the acquisition corpora. In D5 the merging framework will be Exemplar Theory.

Figure D5.2 shows how contextual knowledge is brought to bear on syntactic disambiguation using an exemplar-theoretic process. First, an ambiguous linguistic expression is analyzed. Example: A parser computes the dependency parse of the ambiguous expression: “an opening under the house that led to a fume-filled coal mine.” The analysis result is used as a query to search the contextual knowledge extracted from the acquisition corpora

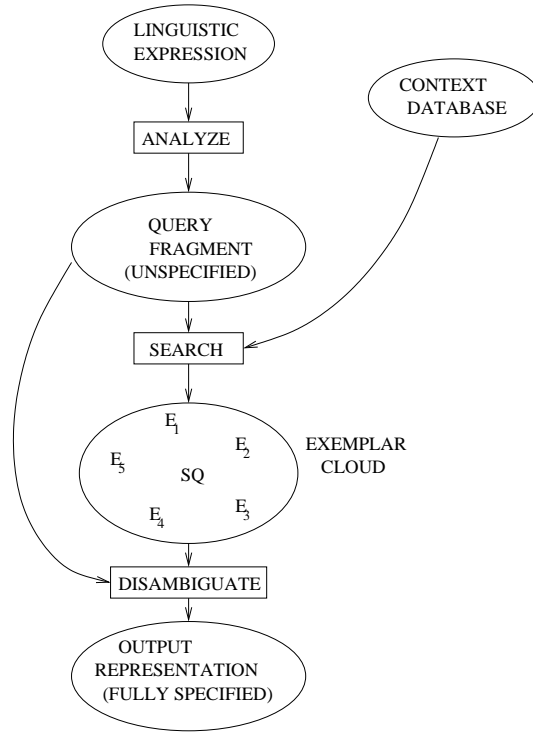


Figure D5.2: Exemplar-theoretic disambiguation.

for similar contexts (or exemplars). A disambiguation decision is then made based on the composition of the resulting exemplar cloud: the number and quality of exemplars found for the two possible relative clause attachments. Openings are more often subjects of “lead to” than houses, indicating that the relative clause should be attached to opening in the example.

The key formal requirements for an exemplar-theoretic approach to syntactic disambiguation then are the similarity measure that determines which fragments are deemed similar to the query and the inference mechanism that takes the nearest neighbors as input and returns a disambiguation decision. We first discuss the two similarity measures that are considered in this proposal (language models and dependency models) and then the inference mechanism.

Language models. The baseline similarity measure in D5 is based on language models. Language models capture local coherence patterns of words. For example, “Die Mehrheit der Abgeordneten stimmten dafür.” is better than “Die Mehrheit sind ohne Partei.” which in turn is better than “Die Mehrheit enthielten sich.” Our hypothesis is that local coherence explains the difference between these sentences. The phrase “Mehrheit enthielten” is an unconcealed violation of agreement, whereas “Die Mehrheit sind” and “Abgeordneten stimmten” have some local coherence due to the frequency of sentences like “Die Mehrheit sind Frauen.” (singular predicate nominal followed by plural verb) and “Die Abgeordneten stimmten dagegen.” It is clear that local coherence is not the dominant factor that determines grammaticality. But we believe that investigating its importance is an interesting research task in the context of this proposal.

Language models cannot be directly used for syntactic disambiguation. A standard language model assigns the same probability to all parses of a sentence since this probability depends on the string of words and not on the parse structure. Our hypothesis is that local cohesion is more important *within* syntactic phrases and less important at transition points *between* syntactic phrases. If the language model is applied selectively, then parses with locally cohesive phrases can be given preference.

Dependency models. Language models are the baseline for defining similarity between an ambiguous sentence fragment and sentence fragments in the context database. In the second model class, dependency models, similarity is determined with respect to syntactic dependencies as computed by a partial parser. An important part of the work planned for this project consists of designing a dependency-based similarity model that accurately predicts correct syntactic attachment. Starting points for our work will be the Jaccard measure on dependencies (see below) and the similarity measure proposed by Atterer & Schütze (2005).

Refinement of dependency models by means of multilingual acquisition. The alignment between the languages of a multitext provides additional information for syntactic disambiguation. For example, the English sentence “Peter saw the man with the TELESCOPE.” has a prepositional phrase attachment ambiguity that the corresponding German sentence does not exhibit: “Mit dem FERNROHR hat Peter den Mann gesehen.” We can exploit this lack of ambiguity in one language to learn how to disambiguate sentences in a second language. This approach is similar to the word sense disambiguation algorithm described by Dagan & Itai (1994), but here we exploit a similar mechanism for *syntactic* instead of lexical disambiguation. Note that the likelihood of being able to disambiguate a given syntactic ambiguity increases with the number of aligned languages available. Kuhn (2004) uses the same basic intuition for the induction of a multilingual grammar.

There will be many cases where multiple languages cannot be used for disambiguation because the same ambiguity occurs in all languages considered; or because the relevant sentences are expressed so differently that no inference about the underlying unambiguous structure is possible. We can either include these cases, but with a higher risk of error, to maintain high recall. Or we can discard them and only include parses that are correct with high probability to achieve higher precision than in monolingual acquisition. We thus have a classical precision-recall tradeoff. Multilingual acquisition can deliver higher precision at the cost of diminished recall. At the cost of a further reduced recall, we can increase the number of languages considered. We will limit ourselves to three here: English, German, and Spanish.

If the other languages are only used to filter out incorrect parses, then no fundamental difference between monolingual and multilingual acquisition may be apparent (except for higher precision). There is however one case where multilingual acquisition is superior to monolingual acquisition. This is true for systematic ambiguities in one language. For example, feminine nouns in German have identical morphological forms for nominative and accusative. As a result there is systematic ambiguity for German sentences that contain a feminine object and a feminine subject with the same grammatical number. It is difficult

to resolve this ambiguity monolingually, but languages like French and English with strict word order constraints will in most cases disambiguate this type of ambiguity.

Inference mechanism. The exemplar-theoretic inference mechanism (the box labeled “disambiguate” in Figure D5.2) consists of mapping the exemplar cloud to a set of features which are then used in discriminative reranking. (In the Work Packages, these are the three subtasks concerned with feature design.)

This approach is different from “classical” Exemplar Theory, in which the exemplar data base consists of labeled examples. It differs in the same way from k-nearest-neighbor classifiers that require labeled examples and assign the majority class. Since our approach is unsupervised, we don’t have any labels. The cognitive model would be that the “label” in this case is the situation that a particular sentence was experienced in.

Computing inference indirectly by way of features is simplistic, but it allows us to focus on the similarity measure in the 4-year project applied for in this proposal. The first problem that needs to be solved is to define a usable exemplar neighborhood. A more complex inference mechanism can then be developed in a possible follow-up project.

3.4.2 Ziele

Many of our goals in this project follow directly from the research program of the SFB (including its long-term goals) as stated in the Vorspann:

- Determine the effect of context on syntactic disambiguation (which we view as a particular kind of specification), where context is understood as the contextual knowledge that a listener applies when trying to understand an utterance;
- Investigate the differences of two types of context in disambiguation: contextual information obtained by shallow analysis (language models) vs. contextual information obtained by dependency parsing;
- Investigate the differences between linguistic and extralinguistic context in disambiguation;
- Investigate the effect of the domain “language” on syntactic disambiguation in context: How does syntactic disambiguation differ between English and German?
- Investigate the learnability of disambiguating contextual information from monolingual and multilingual corpora
- Statistical NLP suffers from overly simplistic models such as Hidden Markov chains and bag of word models. These models can only succeed if the simplicity of the model is compensated for by large training sets. In some areas (e.g., speech recognition), training sets are relatively cheap, so that such a compensation is possible. However, training sets for syntactic disambiguation are expensive, so the only way to make progress on this problem is a better model. We hope that the exemplar-theoretic model proposed here will be superior to current simplistic models and thereby advance the state of the art in NLP.

Within the context of the SFB we pursue the following goals that are specific to D5.

- Supervised learning in statistical parsing has reached an impasse: Performance improvements have leveled off in the last years. Our goal is to show that this impasse can be broken by combining current supervised approaches with unsupervised learning from unannotated data.
- The state of the art is that unsupervised learning does not improve treebank parsing or only marginally. Our goal is to show that substantial improvements are possible.
- Most research on treebank-based parsing has been done on a small number of static collections without systematic investigations of the relationship between treebank size and parsing performance. This question becomes even more important when a second corpus, the acquisition corpus, is used in disambiguation. The project will attempt to quantify the effects of treebank size and acquisition corpus size on parsing performance.
- We would like to develop Exemplar Theory into a theory that bridges the gap between statistical natural language processing and theoretical linguistics. Exemplar Theory already has had considerable success in this respect in phonetics and phonology. In this proposal, our aim is to show that it is explanatory and at the same time a good basis for applications.
- We also want to develop statistical methods for supporting linguistic analysis of specification in context (in collaboration with C2 and B5).

3.4.3 Methoden und Arbeitsprogramm

There are three aspects of the syntactic disambiguation model we want to develop: *formal*, specifically the inference mechanism and the precise mathematical formalization of the similarity measure used in Exemplar Theory; *algorithmic*, aiming at efficient algorithms that implement our models; and *applied*, where the framework is used on treebanks to empirically ground our approach on real data and evaluate its performance.

The methods we employ for empirical experimentation are common in machine learning and natural language processing. We conduct experiments under a blind testing regime whereby the data set is split into a training set and a test set. The test set is set aside for testing purposes only. We apply n-fold cross-validation in order to obtain more robust averages of the results. Cross-validation is often not applied in NLP because statistical parsing is time consuming, but we will make an effort to work with a stricter regime in our experiments.

The evaluation metrics for parsing are standard in the parsing literature (Black et al., 1991; Carroll et al., 2002). They measure the degree of match between gold standard trees and trees produced by the parser to be tested, where matches can be defined with respect to constituents, dependencies or similar syntactic units.

The proposed project consists of two phases. In the first phase, *acquisition parsers* extract contextual information from *acquisition corpora* and store it in the *context database* (C-DB). The acquisition corpora are monolingual in Work Package 1 and multilingual in Work Package 2. In the second phase, Work Package 3, *target parsers* are enriched with

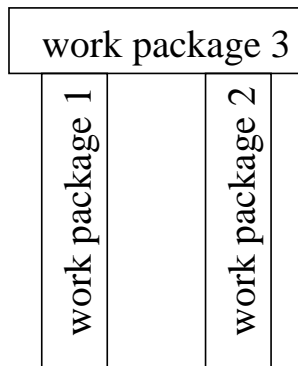


Figure D5.3: Structure of work program.

contextual information from the context database and results are evaluated on *treebank corpora*. Work Package 3 thus builds on the results of the first two work packages as shown in Figure D5.3.

In the description of the work packages that follows, headers of individual tasks contain the names of collaborating projects in parentheses.

WP1: Acquisition from monolingual corpora

WP1 will compute dependency parses of the English Reuters RCV1 corpus (using the Minipar parser, Lin 1998) and of the German .de (Markowetz et al., 2005) and HGC corpora (using Michael Schiehlen’s parser from D3). We will identify additional acquisition corpora if we find that an increase of the size of the acquisition corpora (or choosing acquisition corpora closer in genre to the treebank corpora) has a large impact on disambiguation performance.

1.1 Estimate language model. Here we will estimate a standard language model, probably either the Kneser-Ney model (Kneser & Ney, 1993) or Katz’s backing-off (Katz, 1987), with a maximum n-gram size of 3 or 4, depending on the available computational resources. The language model for German will be shared with D2.

1.2 Design and implement dependency model. The baseline model for comparing the dependency structure of an ambiguous fragment with fragments in the corpus is the Jaccard coefficient: the number of dependencies in the intersection of the dependencies of the fragments divided by the sum of the dependency counts of the two fragments. A more sophisticated, but inefficient similarity measure was described by Atterer & Schütze (2005) (see 3.3.2). The goal in this part of Work Package 1 is to devise a similarity measure that is at the same time efficient and more robust than the Jaccard coefficient.

The dependency model will be made available to D3 for clustering news stories and identifying sentences with similar content.

1.3 Investigate granularity. When looking for exemplars that resemble a “query fragment” (the fragment that is to be disambiguated) a critical question is what granularity of context to choose for fragments and whether this granularity should be variable or fixed. If too broad a context is chosen, then we will fail to find the specific information in the database that is needed for accurate disambiguation. Making the context too small could eliminate contextual information that is needed for disambiguation. The method in Atterer & Schütze (2005) starts with a large context and reduces it incrementally until a match in the database of exemplars (the C-DB) is found. This is an attractive tradeoff between context specificity and inclusiveness, but the search implementing this procedure is expensive. The goal of this part of Work Package 1 is to design a method for identifying the optimal granularity for a particular disambiguation task. The maximum fragment size that will be considered in this proposal is the sentence.

1.4 Design and implement C-DB. The goal of this part of WP1 is to design and implement the context database (C-DB) for storing the contextual information extracted from corpora. The basic representation of the parses will be a simple XML file, but we will need two additional representations for the items in Work Package 3 that compute features: aggregate statistics and a data structure that supports nearest-neighbor search. For computing aggregate statistics, a relational database is most appropriate. We will investigate both relational databases and information retrieval data structures (inverted indices) for supporting nearest-neighbor search, building on the work done by Atterer & Schütze (2005).

1.5 Extract data for C-DB. This task parses the English and German corpora and feeds the extracted information into the context database. The English partial parser used will be Minipar and the German partial parser will be provided by D3. As a preprocessing step the .de corpus will have to be “cleansed:” all material that does not consist of complete German sentences will be discarded.

WP2: Acquisition from multilingual corpora

Work Package 2 is concerned with acquisition from multilingual corpora that contain essentially identical content in several languages (“multitexts”). We choose the Europarl corpus since it contains 11 European languages. This will enable us to add more languages in possible follow-up projects. The main mode of acquisition is to parse selected languages partially, process the results and use them in WP3 for disambiguation.

2.1 Alignment. As a prerequisite for acquisition, we need to align the corpus with high quality. We will perform the alignment with state-of-the-art software, probably GIZA++ (Och & Ney, 2000).

2.2 Crosslinguistic referential context (C2). This task is concerned with exploiting the dependency parses of the multitext for providing C2 with a statistical characterization of referential contexts of certain nominal phrases in Spanish. The characterization will be in terms of the German phrases that are aligned with the Spanish phrases in Europarl. For example, one will be able to compare direct objects marked with “a” in Spanish that correspond to noun phrases in the *Vorfeld* in German with those that correspond to NPs in the *Mittelfeld*.

2.3 Design and implement dependency disambiguation. The main task in this work package will be to develop the framework for mutual syntactic disambiguation. In the simplest case, one can compute the intersection of the two sets of dependencies that occur in the two languages for a given sentence. For this purpose, dependency relationships produced by the two parsers must have a minimum level of compatibility. For example, both parsers should identify subject-verb dependencies using the same label. In this case, the mapping is simple (at least for European languages). However, there are many non-trivial cases. For example, cleft sentences in English have no direct equivalent in German. Part of the work package will therefore be to develop a methodology for mapping a set of dependencies in one language to the corresponding set in a second language.

2.4 Extract data for C-DB. The parsers for English and German identified in Work Package 1 will again be used, this time for parsing the English and German parts of Europarl. In addition, a Spanish partial parser (probably Conexor, based on Voutilainen et al.’s work (Voutilainen & Järvinen, 1996)) will be deployed to parse the Spanish part of the corpus. The extracted information is then fed into the context database.

WP3: Exploitation of acquired information

This work package is concerned with exploiting the contextual information acquired in the first two work packages as a resource for disambiguating syntactic parses.

We use Bitpar (Schmid, 2004b) to compute a set of k preferred parses and rerank them (Collins & Koo, 2005) based on contextual knowledge from the C-DB. Training and evaluation will be performed on the Penn treebank for English and Tiger or Negra for German.

3.1 Train Bitpar (D4). The proposed disambiguation procedure is discriminative: It needs a set of possible parses as input and identifies the most likely reading among these candidates. We use Bitpar to compute the initial set of parses. In this task, Bitpar will be trained on the English and German treebanks. In addition, any preprocessing of Bitpar parses that is required for applying discriminative disambiguation will also be performed here.

3.2 Design and implement discriminative disambiguation algorithm. In this part of the work package, the discriminative reranking algorithm will be designed and implemented following Collins & Koo (2005).

3.3 Signatures for French verbs (B5). This task will provide clustered verb data as an additional source for creating a cognitively adequate ontology of French verbs. Our hypothesis is that these clusters can serve as an approximate representation of prototypes in the exemplar-theoretic sense (except that in this case they would be semantic instead of phonetic/phonological prototypes). Each occurrence of a verb in Europarl will be represented as the signature of the verb’s translations in 5–10 of the other languages. For example, an occurrence of the verb “demander” in the sense of “demand” might be represented as “demand – fordern – domandare – exigir”. These signatures will then be clustered. The expectation is that occurrences of each prototypical use of a French verb

will be grouped into a separate cluster. Clusters will be presented with typical contexts to researchers in B5 as an aid in composing ontological entries. We will discuss with B5 to what extent an exemplar-theoretic framework can help provide formal criteria for making decisions about the structure of the ontology and the form and content of the individual entries. This exploratory study is intended to lay the foundation for a closer collaboration in a follow-on project.

3.4 Design and test language model context features (D2). In this task, we will design features for discriminative disambiguation that are based on language model similarity. An example of a language model feature would be the average log probability of transitions between words within phrases. On the hypothesis that transitions within a phrase are more likely than between phrases, higher average log probability should indicate that a parse is more likely. Features will be designed and evaluated in collaboration with D2.

3.5 Design and test dependency context features (D2). In this task, we will design features for discriminative disambiguation that are based on dependency model similarity. The simplest such features are frequency-based. If the information in C-DB indicates that the lexical dependencies corresponding to the low attachment of a prepositional phrase are more frequent than those corresponding to high attachment, then the former will be given preference. Features will be designed and evaluated in collaboration with D2.

3.6 Design and test multilingual context features (D2). In this task, we will design features for discriminative disambiguation that are based on multilingual context features. These features are similar to monolingual dependency features. However, the basic information retrieved from C-DB will be more accurate due to the mutual syntactic disambiguation of the different languages in the multitext, but also sparser due to decreased recall. Features will be designed and evaluated in collaboration with D2.

3.7 Formalize unified Exemplar Theory (A2). In the final year of the 4-year period applied for here, after having collected extensive experience with the exemplar-theoretic approach to syntactic disambiguation in D5 and with the ISC model of Exemplar Theory in A2, we will work on a common Exemplar Theory framework that is adequate for phonetics/phonology, syntactic disambiguation and other areas of linguistics. The most significant difference between the models in A2 and D5 is that exemplars in A2 are labeled whereas those in D5 are not. How exactly labels are attached to exemplars in speech perception is not well understood in Exemplar Theory. On the other hand, if the idea of “situational labels” (that is, context information that does not correspond to a clear category label) could be developed in more detail in D5, it might open up new research directions for Exemplar Theory in speech production and perception. This task will explore these issues and, together with A2, attempt to develop a unified version of Exemplar Theory.

3.8 Analysis of contextual effects. One of the central questions of this SFB is how different types of context influence specification. This task consists of linguistic analysis of successful and unsuccessful instances of the disambiguation algorithm designed and tested

in WP3. The result will be an empirical study that shows for the example of syntactic disambiguation what effect different types of context have on specification, concentrating on the different effects of linguistic context vs. extralinguistic context and the differences in disambiguation between English and German. The goal is 1) to characterize classes of syntactic ambiguities as to whether their resolution is based on linguistic regularities (e.g., a preference for low attachment) or on extralinguistic knowledge (e.g., a telescope is an instrument of seeing usually not associated with hills); and 2) to determine whether different ambiguity classes and different disambiguation strategies are used in English and German.

3.9 Quantitative analysis. It is well-known that disambiguation performance improves as the size of the treebank corpus increases, but it is not clear how this effect interacts with contextual information acquired from acquisition corpora. For example, it is possible that acquired information does not help beyond a certain treebank corpus size. We will characterize the behavior of the system with learning curves for different sizes of treebanks and acquisition corpora and their interactions.

Jahr	2006		2007				2008				2009				2010	
Quartal	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2
1 Monolingual acquisition																
1.1 Estimate language model (LM)			■	■												
1.2 Design and implement (D/I) dependency model					■	■	■	■								
1.3 Investigate granularity									■	■						
1.4 D/I C-DB			■	■	■	■	■	■	■	■	■	■	■			
1.5 Extract data for C-DB	■	■								■			■			
2 Multilingual acquisition																
2.1 Alignment	■	■														
2.2 Crosslinguistic referential context (C2)															■	■
2.3 D/I dependency disambiguation					■	■	■	■								
2.4 Extract data for C-DB									■	■			■			
3 Exploitation																
3.1 Train Bitpar (D4)	■	■														
3.2 D/I discriminative disambiguation			■	■							■	■				
3.3 Signatures for French verbs (B5)	■	■														
3.4 Design and test (D/T) LM context features (D2)							■	■						■		
3.5 D/T dependency context features (D2)									■	■				■		
3.6 D/T multilingual context features (D2)											■	■		■		
3.7 Formalize unified Exemplar Theory (A2)													■	■	■	■
3.8 Analysis of contextual effects													■	■	■	■
3.9 Quantitative analysis												■	■		■	■

Outlook

This proposal is meant to be the beginning of a longer investigation of exemplar-theoretic approaches to disambiguation. We initially investigate only simple forms of the inference process that infers a syntactic disambiguation decision from the neighborhood of an ambiguous fragment in “exemplar space.” We restrict ourselves to such simple inference processes to be able to concentrate on the similarity measures. An appropriate similarity measure is a prerequisite for a usable exemplar neighborhood and therefore needs to be addressed before more complex inference methods. A possible follow-up project could then develop a more sophisticated exemplar-theoretic account of syntactic disambiguation, including a complex inference method.

In follow-on projects after this Förderperiode, we intend to investigate how a better similarity measure can be constructed based on lexical resources such as WordNet (Miller et al., 1990), Germanet (Lemnitzer & Kunze, 2002), Cyc (Lenat & Guha, 1989), ConceptNet (Liu & Singh, 2004) and multilingual dictionaries. We also intend to look at the other languages in Europarl. The exploratory investigation of the other languages in Europarl (in collaboration with B5) would serve as the foundation for extending the project to other languages in subsequent Förderperioden.

3.5 Stellung innerhalb des Sonderforschungsbereichs

3.5.1 Stellung zum Gesamtkonzept des SFB

Many of our goals in this project follow directly from the research program of the SFB (including its long-term goals) as stated in the Vorspann:

- Determine the effect of context on syntactic disambiguation
- Investigate the differences of two types of context in disambiguation: contextual information obtained by shallow analysis (language models) vs. contextual information obtained by dependency parsing;
- Investigate the differences between linguistic and extralinguistic context in disambiguation;
- Investigate the effect of the domain “language” on syntactic disambiguation in context: How does syntactic disambiguation differ between English and German?
- Investigate the learnability of disambiguating contextual information from monolingual and multilingual corpora
- Statistical NLP suffers from overly simplistic models such as Hidden Markov chains and bag of word models. These models can only succeed if the simplicity of the model is compensated for by large training sets. In some areas (e.g., speech recognition), training sets are relatively cheap, so that such a compensation is possible. However, training sets for syntactic disambiguation are expensive, so the only way to make progress on this problem is a better model. Our goal is to show that the exemplar-theoretic model proposed here is superior to current simplistic models and thereby advance the state of the art in NLP.

3.5.2 Interaktion mit anderen Teilprojekten

D5 will collaborate most closely with A2, D2 and D4.

- A2 and D5 will collaborate on a new version of Exemplar Theory that is adequate for both phonetics/phonology and syntactic disambiguation.
- D2 and D5 will exchange features as well as data for constructing new features. They will also collaborate on methods and heuristics for discovering optimal features for discriminative reranking.
- D2, D4, and D5 share a common theme of using knowledge acquired from unannotated corpora for improving the performance of treebank trained parsers. The three projects differ in focus: generation and information structure in D2; the generative modularization framework in D4; Exemplar Theory, language models and multilingual corpora in D5. But many intermediate research results will be relevant for all three projects. We plan to have monthly meetings to coordinate the three projects.

D5 will use parsers developed in D3 and D4:

- D3 will provide the partial parser for German, a key component of the infrastructure for the project
- D4 will provide the Bitpar parser, the base parser used for syntactic disambiguation.

D5 will provide output for three projects:

- Corpus-based signatures for French verbs will be one of the information sources used by researchers in B5 for constructing the ontology of French verbs. We will also discuss with B5 to what extent an exemplar-theoretic framework can help provide formal criteria for making decisions about the structure of the ontology and the form and content of the individual entries. This exploratory study is intended to lay the foundation for a closer collaboration in a follow-on project.
- A characterization of referential context in Spanish in terms of German translations will be one of the information sources in C2 for analyzing statistical effects in referential contexts.
- A basic similarity measure for sentences, together with a clustering procedure, will be delivered to D3 in order to facilitate the extraction of inference rules from sentence groups.

We will also collaborate with A3 on how to represent context (e.g., in C-DB) and on statistical methodology in general.

3.6 Abgrenzung gegenüber anderen geförderten Projekten des Teilprojektleiters

SFB 627 “Nexus” (Spatial World Models for Mobile Context-Aware Applications) is supporting the project NexSem at IMS on “semantic” approaches to mapping new mobile applications to the internal world model of Nexus. NexSem focuses on a particular application whereas the project proposed here addresses syntactic ambiguity resolution as a fundamental problem of natural language processing and linguistics.

3.7 Ergänzungsaussstattung für das Teilprojekt

	2006/2			2007			2008			2009			2010/1		
	Verg.-Gr.	Anz.	Betrag EUR	Verg.-Gr.	Anz.	Betrag EUR	Verg.-Gr.	Anz.	Betrag EUR	Verg.-Gr.	Anz.	Betrag EUR	Verg.-Gr.	Anz.	Betrag EUR
PM	BAT IIa	1	29.400	BAT IIa	1	58.800	BAT IIa	1	58.800	BAT IIa	1	58.800	BAT IIa	1	29.400
	BAT IIa/2	1	14.700	BAT IIa/2	1	29.400	BAT IIa/2	1	29.400	BAT IIa/2	1	29.400	BAT IIa/2	1	14.700
	stud. HK	1,5	9.000	stud. HK	1,5	18.000	stud. HK	1,5	18.000	stud. HK	1,5	18.000	stud. HK	1,5	9.000
	zus.:		53.100	zus.:		106.200	zus.:		106.200	zus.:		106.200	zus.:		53.100

	Kostenkategorie oder Kennziffer	Betrag EUR	Kostenkategorie oder Kennziffer	Betrag EUR	Kostenkategorie oder Kennziffer	Betrag EUR	Kostenkategorie oder Kennziffer	Betrag EUR	Kostenkategorie oder Kennziffer	Betrag EUR
	SM	Kleingeräte	5.000	Kleingeräte	-	Kleingeräte	-	Kleingeräte	-	Kleingeräte
Verbrauchsmat.		-	Verbrauchsmat.	-	Verbrauchsmat.	-	Verbrauchsmat.	-	Verbrauchsmat.	-
Reisen		-	Reisen	-	Reisen	-	Reisen	-	Reisen	-
Sonstiges		-	Sonstiges	-	Sonstiges	-	Sonstiges	-	Sonstiges	-
zus.:		5.000	zus.:	-	zus.:	-	zus.:	-	zus.:	-

	IM insges.	IM insges.	IM insges.	IM insges.	IM insges.
IM	22.000	-	-	-	-

3.7.1 Personal im Teilprojekt

Grundausstattung

	Name, akad. Grad, Dienststellung	engeres Fach des Mitarbeiters	Institut der Hochschule	Beantragte Förderperiode: Mitarbeit im Teilprojekt in Std./Woche (beratend: B)					Vergütungsgruppe
				B 2006/2	2007	2008	2009	2010/1	
3.7.1.1 wissenschaftl. Mitarbeiter (einschl. Hilfskräfte)	1. Hinrich Schütze, Prof. Ph.D., Universitätsprofessor	Computerlinguistik	IMS, Univ. Stuttgart	5	5	5	5	5	C4

Aufgabenbeschreibung von Mitarbeitern der Grundausstattung für die beantragte Förderperiode

zu 1. Prof. Hinrich Schütze, Ph.D.

The principal investigator; coordinates project D5, oversees collaboration with other SFB projects and is responsible for D5 overall.

Ergänzungsausstattung

	Name, akad. Grad, Dienststellung	engeres Fach des Mitarbeiters	Institut der Hochschule	Beantragte Förderperiode: Mitarbeit im Teilprojekt in Std./Woche (beratend: B)					Vergütungsgruppe
				B 2006/2	2007	2008	2009	2010/1	
3.7.1.2 wissenschaftl. Mitarbeiter (einschl. Hilfskräfte)	1. Atterer, Michaela, Dr. phil., wiss. Mitarbeiterin	Computerlinguistik	IMS, Univ. Stuttgart	41	41	41	41	41	BAT IIa
	2. N.N.	Informatik	IMS, Univ. Stuttgart	20,5	20,5	20,5	20,5	20,5	BAT IIa/2
	3. N.N.	Linguistik	IMS, Univ. Stuttgart	15	15	15	15	15	stud. HK
	4. N.N.	Computerlinguistik	IMS, Univ. Stuttgart	15	15	15	15	15	stud. HK

Aufgabenbeschreibung von Mitarbeitern der Ergänzungsausstattung für die beantragte Förderperiode

zu 1. Dr. Michaela Atterer

Full-time researcher, computational linguist; will be responsible for all aspects of Work Packages 1, 2 and 3, except for system design and implementation.

zu 2. N.N. (BAT IIa/2)

Half-time researcher, computer scientist; will be responsible for all aspects of implementation and system design, including design of data structure and implementation aspects of Exemplar Theory, parsing, and alignment. The ideal candidate would be a doctoral student holding a diploma in computer science.

zu 3. N.N. (stud. HK)

Student assistant 1, linguist; will support all linguistic analyses in the work packages, including analysis of contextual effects (3.8), quality control of the signatures for French verbs (3.3) and quality control of the methods developed for investigating cross-linguistic referential context (2.2).

zu 4. N.N. (stud. HK)

Student assistant 2, computational linguist; will support computational linguistics tasks, including preprocessing of corpora, testing of language and dependency models and of the features designed in Work Package 3.

3.7.2 Aufgliederung und Begründung der Sachmittel (nach Haushaltsjahren)

	2006/2	2007	2008	2009	2010/1
Für Sächliche Verwaltungsausgaben stehen als Grundausrüstung voraussichtlich zur Verfügung:	1.000	2.000	2.000	2.000	1.000
Für Sächliche Verwaltungsausgaben werden als Ergänzungsausstattung beantragt (entspricht den Gesamtsummen „Sächliche Verwaltungsausgaben“ in Übersicht 3.7):	5.000	–	–	–	–

(Alle Angaben in EUR)

Begründung zur Ergänzungsausstattung der Sachmittel

Kleingeräte

We are applying for funding for two notebooks for the two researchers in the project.

Reisen

Bezeichnung	2006/2	2007	2008	2009	2010/1
Travel money	2.400	4.800	4.800	4.800	2.400

Travel money is applied for centrally. The following sums are necessary for D5:

For each researcher one national and one international conference per year. Relevant conferences: ACL, EACL, NAACL, EMNLP, HLT, Coling.

Expenses:

international conference per participant EUR 1.500,- (conference fee EUR 500, hotel EUR 500, flight EUR 500);

national conference EUR 900,- (conference fee EUR 300, hotel EUR 400, travel EUR 200)

3.7.3 Investitionen (Geräte über 10.000,- EUR brutto und Fahrzeuge)

D5 will be working with large text corpora that will be annotated automatically via partial parsers. It is desirable to store the parsed version of each corpus before experimenting with relational databases and information retrieval data structures as described in Workpackage 1 of D5 in order to avoid time-consuming re-parsing in each development and testing cycle of the system. In the experiments that (Atterer & Schütze, 2005) is based on, we have found that we need about four times as much space as the uncompressed corpus to store all derivative data structures. The size of .de is 1 TB, so we would need about 4 TB for this corpus. The other corpora are much smaller, so that the total we are applying for in this document is 4 TB. This requirement can be satisfied by the RAID system described in section 2.2.4.

For fast processing of our data with a variety of tools (partial parsers, aligner etc.), and for database access to the large amount of information extracted from the corpora, a fast server with 32 GB RAM and several processors is needed. We are applying for this server together with D4 (see section 2.2.4).

Bibliography

- Abeillé, A., Clément, L., Toussenel, F., 2003. *Building a treebank for French*. In: A. Abeillé (ed.), *Treebanks*. Kluwer, Dordrecht.
- Argamon, S., Dagan, I., Krymolowski, Y., 1998. *A Memory-Based Approach to Learning Shallow Natural Language Patterns*. In: ACL 36/COLING 17. 67–73.
- Arun, A., 2004. *Statistical Parsing of the French Treebank*. Master's thesis, School of Informatics, University of Edinburgh.
- Atterer, M., Schütze, H., 2005. *A lattice-based framework for unsupervised syntactic disambiguation with an application to relative clause attachment*, submitted.
- van der Beek, L., Bouma, G., Daciuk, J., Gaustad, T., Malouf, R., van Noord, G., Prins, R., Villada, B., 2002. *Algorithms for Linguistic Processing*. NWO PIONIER Progress Report, Groningen, Ch. Chapter 5. The Alpino Dependency Treebank.
- Bikel, D., Chiang, D., 2000. *Two statistical parsing models applied to the Chinese Treebank*.
- Bikel, D. M., 2004. Intricacies of collins' parsing model. *Computational Linguistics* **30** (4), 479–511.
- Black, E., Abney, S. P., Flickinger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M. P., Roukos, S., Santorini, B., Strzalkowski, T., 1991. *A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars*. In: Proc. of the Speech and Natural Language Workshop. Pacific Grove, CA, 306–311.
- Blum, A., Mitchell, T., 1998. *Combining Labeled and Unlabeled Data with Co-training*. In: COLT.
- Bod, R., Scha, R., Sima'an, K., 2003. *Data-Oriented Parsing*. CSLI Publications.
- Bybee, J., 2001. *Frequency effects on French liaison*. In: J. Bybee, P. Hopper (eds.), *Frequency effects and Emergent Grammar*. John Benjamins, Amsterdam, 337–359.

- Calvo, H., Gelbukh, A., Kilgarriff, A., 2005. *Distributional Thesaurus vs. WordNet: A Comparison of Backoff Techniques for Unsupervised PP Attachment*. In: Proc. of the sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing).
- Carroll, G., Rooth, M., 1998. *Valence Induction with a Head-lexicalized PCFG*. In: Proc. of EMNLP. Granada, Spain.
- Carroll, J., Frank, A., Lin, D., Prescher, D., Uszkoreit, H. (eds.), 2002. *Beyond PARSEVAL - Towards Improved Evaluation Measures for Parsing Systems. Workshop at the 3rd International Conference on Language Resources and Evaluation (LREC-02)*. Las Palmas, Gran Canaria, Spain.
- Charniak, E., 1997. *Statistical Parsing with a Context-Free Grammar and Word Statistics*. In: AAAI/IAAI. 598–603.
- Charniak, E., Johnson, M., 2005. *Coarse-to-fine n-best parsing and MaxEnt discriminative reranking*. In: ACL 43.
- Chiang, D., Bikel, D. M., 2002. *Recovering latent information in treebanks*. In: Proceedings of the 19th international conference on Computational linguistics. Association for Computational Linguistics, Morristown, NJ, USA, 1–7.
- Collins, M., 1997. *Three generative, lexicalised models for statistical parsing*. In: Proceedings of the 35th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, USA, 16–23.
- Collins, M., Koo, T., March 2005. Discriminative reranking for natural language parsing. *Computational Linguistics* **31** (1), 25–70.
- Collins, M., Ramshaw, L., j. Hajič, Tillmann, C., 1999. *A statistical parser for Czech*. In: Proceedings of the 37th conference of the ACL.
- Dagan, I., Itai, A., 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics* **20** (4), 563–596.
- Dubey, A., Keller, F., 2003. *Probabilistic parsing for German using sister-head dependencies*. In: ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, USA, 96–103.
- Dumais, S., Letsche, T., Littman, M., Landauer, T., 1997. *Automatic cross-language retrieval using latent semantic indexing*.
- Evert, S., 2004. *The Statistical Analysis of Morphosyntactic Distributions*. In: Proc. of LREC. Lisbon, Portugal, 1539–1542.
- Finch, S., Chater, N., 1992. *Bootstrapping Syntactic Categories Using Statistical Methods*. In: W. Daelemans, D. Powers (eds.), *Background and Experiments in Machine Learning of Natural Language*. Institute for Language Technology and AI, Tilburg University, 229–235.
- Frank, A., Becker, M., Crysmann, B., Kiefer, B., Schäfer, U., 2003. *Integrated shallow and deep parsing: TopP meets HPSG*. In: ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, USA, 104–111.
- Hajič, J., 1998. *Building a Syntactically Annotated Corpus: The Prague Dependency Treebank*. In: E. Hajičová (ed.), *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*. Charles University Press, Prague Karolinum.
- Hindle, D., Rooth, M., 1991. *Structural ambiguity and lexical relations*. In: Proc. of ACL 29. Association of Computational Linguistics, Morristown NJ, 229–236.

- Hindle, D., Rooth, M., 1993. Structural ambiguity and lexical relations. *Computational Linguistics* **19** (1), 103–120.
- Hinrichs, E., 2005. *Ergebnisbericht 2002-2004*. <http://www.sfb441.uni-tuebingen.de/a1/Publikationen/a1-ergebnis02-04.pdf>.
- Hull, D. A., Pedersen, J. O., Schütze, H., 1996. *Method Combination for Document Filtering*. In: SIGIR '96. 279–287.
- Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., Kolak, O., 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Special Issue of the Journal of Natural Language Engineering on Parallel Texts, to appear* **11** (3).
- Johnson, M., Riezler, S., 2000. *Exploiting auxiliary distributions in stochastic unification-based grammars*. In: Proceedings of the first conference on North American chapter of the Association for Computational Linguistics. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 154–161.
- Katz, S. M., 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing* **35** (3).
- Kermes, H., Heid, U., 2003. *Using chunked corpora for the acquisition of collocations and idiomatic expressions*. In: Complex 2003.
- Kirchner, R., 1999. Preliminary thoughts on “phonologization” within an exemplar-based speech processing system. *Technical Report, UCLA Working Papers in Linguistics* **6**.
- Klein, D., Manning, C. D., 2003. *Accurate Unlexicalized Parsing*. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics.
- Klein, D., Manning, C. D., 2004. *Corpus-Based Induction of Syntactic Structure: Models of Dependency and Constituency*. In: ACL. 478–485.
- Kneser, R., Ney, H., 1993. *Forming Word Classes by Statistical Clustering for Statistical Language Modelling*. In: R. Köhler, B. B. Rieger (eds.), *Contributions to Quantitative Linguistics*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 221–226.
- Kübler, S., Hinrichs, E. W., 2001. *From chunks to function-argument structure: A similarity-based approach*. In: ACL. 338–345.
- Kuhn, J., 2004. *Experiments in parallel-text based grammar induction*. In: ACL. 470–477.
- Lacerda, F., 1995. *The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory*. In: Proceedings of the 13th International Congress of Phonetic Sciences (Stockholm). Vol. 2. 140–147.
- Lemnitzer, L., Kunze, C., 2002. *GermaNet - representation, visualization, application*. In: Proc. LREC 2002. Third International Conference on Language Resources and Evaluation. 1485–1491.
- Lenat, D. B., Guha, R. V., 1989. *Building Large Knowledge-Based Systems*. Addison-Wesley, Reading MA.
- Lewis, D. D., Yang, Y., Rose, T. G., Li, F., 2004. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* **5**, 361–397.
- Lin, D., 1998. *Dependency-based evaluation of MINIPAR*. In: Workshop on the Evaluation of Parsing Systems. Granada, Spain.
- Liu, H., Singh, P., 2004. *Commonsense Reasoning in and over Natural Language*. In: International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES'2004). Springer.

- Magerman, D. M., 1994. *Natural language parsing as statistical pattern recognition*. Ph.D. dissertation, Stanford University.
- Malouf, R., van Noord, G., 2004. *Wide Coverage Parsing with Stochastic Attribute Value Grammars*. In: IJCNLP-04 Workshop Beyond Shallow Analyses - Formalisms and statistical modeling for deep analyses.
- Manning, C. D., Schütze, H., 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Boston, MA.
- Marcus, M. P., Santorini, B., Marcinkiewicz, M. A., 1993. Building a large natural language corpus of English: the penn treebank. *Computational Linguistics* **19**, 313–330.
- Markowetz, A., Chen, Y.-Y., Suel, T., Long, X., Seeger, B., June 2005. *Design and Implementation of a Geographic Search Engine*. In: 8th Int. Workshop on the Web and Databases (WebDB).
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J., 1990. Introduction to WordNet: An on-line lexical database. *Journal of Lexicography* **3** (4), 235–244.
- Moreno, A., Grishman, R., Lopez, S., Sanchez, F., Satoshi, S., 2000. *A Treebank of Spanish and its Application to Parsing*. In: LREC 2000.
- Muslea, I., Minton, S., Knoblock, C. A., 2002. *Active + Semi-supervised Learning = Robust Multi-View Learning*. In: ICML.
- Och, F. J., Ney, H., October 2000. *Improved Statistical Alignment Models*. In: ACL 2000. Hongkong, China, 440–447.
- Pierrehumbert, J., 2001. *Exemplar dynamics: Word frequency, lenition and contrast*. In: J. Bybee, P. Hopper (eds.), *Frequency and the Emergence of Linguistic Structure*. Benjamins, Amsterdam, 137–157.
- Riezler, S., King, T. H., Kaplan, R. M., Crouch, R. S., III, J. T. M., Johnson, M., 2002. *Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques*. In: ACL. 271–278.
- Rohrer, C., Forst, M., 2006. *Improving coverage and parsing quality of a large-scale LFG for German*. In: LREC. Submitted.
- Schiehlen, M., 2003a. *A Cascaded Finite-State Parser for German*. In: EACL.
- Schiehlen, M., 2003b. *Combining Deep and Shallow Approaches in Parsing German*. In: ACL.
- Schiehlen, M., 2004. *Annotation Strategies for Probabilistic Parsing in German*. In: Coling.
- Schmid, H., 1994. *Probabilistic Part-of-Speech Tagging Using Decision Trees*. In: Proceedings of the International Conference on New Methods in Language Processing. Manchester, UK, 44–49.
- Schmid, H., 1995. *Improvements in Part-of-Speech Tagging with an Application to German*. In: Proceedings of the ACL SIGDAT-Workshop. 47–50.
- Schmid, H., 2004a. *Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors*. In: Proceedings of the 20th International Conference on Computational Linguistics. Vol. 1. Geneva, Switzerland, 162–168.
- Schmid, H., 2004b. *Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors*. In: Coling.
- Schuetze, H., Chen, F. R., Pirolli, P. L., Pitkow, J. E., Chi, E. H., Li, J., 2005a. *System and method for identifying similarities among objects in a collection*. United States Patent 6,941,321.

- Schuetze, H., Chen, F. R., Pirolli, P. L., Pitkow, J. E., Chi, E. H., Li, J., Gargi, U., 2005b. *System and method for quantitatively representing data objects in vector space*. United States Patent 6,922,699.
- Schulte im Walde, S., 2003. *Experiments on the Automatic Induction of German Semantic Verb Classes*. Ph.D. dissertation, IMS, University of Stuttgart.
- Schütze, H., 1993a. *Distributed Syntactic Representations with an Application to Part-of-Speech Tagging*. In: Proc. of the IEEE International Conference on Neural Networks. 1504–1509.
- Schütze, H., 1993b. *Translation by Confusion*. In: B. Dorr (ed.), *Working Notes of the AAAI Spring Symposium on Building Lexicons for Machine Translation*. AAAI Press, Menlo Park CA, 82–85.
- Schütze, H., 1995. *Distributional Part-of-Speech Tagging*. In: EACL 7. 141–148.
- Schütze, H., 1998. Automatic word sense discrimination. *Computational Linguistics* **24** (1), 97–124.
- Siddharthan, A., 2002a. *Resolving Attachment and Clause Boundary Ambiguities for Simplifying Relative Clause Constructs*. In: Proceedings of the Student Research Workshop, 40th Meeting of the Association for Computational Linguistics (ACL 2002).
- Siddharthan, A., 2002b. *Resolving Relative Clause Attachment Ambiguities using Machine Learning Techniques and WordNet Hierarchies*. In: Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002).
- Skut, W., Brants, T., Krenn, B., Uszkoreit, H., 1998. *A linguistically interpreted corpus of German newspaper text*. In: B. Krenn, T. Brants, W. Skut, H. Uszkoreit (eds.), *Proceedings of the 10th European Summer School in Logic, Language and Information (ESSLLI'98), Workshop on Recent Advances in Corpus Annotation*.
- Solan, Z., Horn, D., Ruppin, E., Edelman, S., August 2005. Unsupervised learning of natural languages. *PNAS* **102** (33), 11629–11634.
- Uszkoreit, H., Brants, T., Duchier, D., Krenn, B., Konieczny, L., Oepen, S., Skut, W., 1998. Studien zur performanzorientierten Linguistik. Aspekte der Relativsatzextraposition im Deutschen. *Kognitionswissenschaft* **7** (3).
- Volk, M., 2001. *Exploiting the WWW as a corpus to resolve PP attachment ambiguities*. In: Proceedings of Corpus Linguistics 2001.
- Voutilainen, A., Järvinen, T., 1996. *Using the English Constraint Grammar Parser to Analyse a Software Manual Corpus*. In: H.-D. Koch, R. F. Sutcliffe, A. McElligott (eds.), *Industrial Parsing of Software Manuals*, pp. 57-87. *Language and Computers: Studies In Practical Linguistics, No. 17*. Rodopi, Amsterdam, 57–87.
- Wu, D., 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Comput. Linguist.* **23** (3), 377–403.
- Xue, N., Chiou, F.-D., Palmer, M., 2002. *Building a large-scale annotated Chinese corpus*. In: Proceedings of the 19th international conference on Computational linguistics. Association for Computational Linguistics, Morristown, NJ, USA, 1–8.
- Xue, N., Xia, F., Chiou, F.-D., Palmer, M., 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering* **11** (2).
- Yeh, A. S., Vilain, M. B., 1998. *Some properties of preposition and subordinate conjunction attachments*. In: Proceedings of the 17th international conference on Computational linguistics. Association for Computational Linguistics, Morristown, NJ, USA, 1436–1442.
- Zens, R., Ney, H., May 2 - May 7 2004. *Improvements in Phrase-Based Statistical Machine Translation*. In: D. M. Susan Dumais, S. Roukos (eds.), *HLT-NAACL 2004: Main Proceedings*. Association for Computational Linguistics, Boston, Massachusetts, USA, 257–264.

Prof. Hinrich Schütze, Ph.D.

Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart

Azenbergstr. 12, 70174 Stuttgart

Wissenschaftlicher Werdegang

1964 Geboren.

1989 Diplom Informatik, Universität Stuttgart.

1995 Promotion, Stanford University.

2000 – 2004 Consulting Assistant Professor, Stanford University.

seit 2004 Ordentlicher Professor für Theoretische Computerlinguistik am Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Ausgewählte Liste der wichtigsten 10 Publikationen der vergangenen 5 Jahre

- Atterer, M., Schütze, H., 2005. *A lattice-based framework for unsupervised syntactic disambiguation with an application to relative clause attachment*, submitted.
- Blessing, A., Klatt, S., Nicklas, D., Volz, S., Schütze, H., 2006. *Language-derived information and context models*. In: Proceedings of Comorea.
- Chang, J. T., Schütze, H., 2005. *Abbreviations in biomedical text*. In: S. Ananiadou, J. McNaught (eds.), *Text Mining for Biology and Biomedicine*. Artech House Books, to appear.
- Chang, J. T., Schütze, H., Altman, R. B., 2004. GAPSCORE: Finding gene and protein names one word at a time. *Bioinformatics* **20** (2), 216–225.
- Manning, C. D., Raghavan, P., Schütze, H., 2007. *Introduction to information retrieval*. Cambridge University Press, to appear. Web publication at URL: www.informationretrieval.org.
- Manning, C. D., Schütze, H., 1999. *Foundations of statistical natural language processing*. MIT Press, Boston, MA.
- Pitkow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., Breuel, T., 2002. Personalized search. *Communications of the ACM* **45** (9), 50–55, URL: <http://doi.acm.org/10.1145/567498.567526>.
- Raychaudhuri, S., Schütze, H., Altman, R. B., 2003. Inclusion of textual documentation in the analysis of multidimensional data sets: Application to gene expression data. *Mach. Learn.* **52** (1–2), 119–145.
- Schütze, H., 2000. *Disambiguation and connectionism*. In: Y. Ravin, C. Leacock (eds.), *Polysemy and Ambiguity: Theoretical and Applied Approaches*. Oxford University Press, Oxford.
- Schütze, H., Su, K.-Y. (eds.), 2000. *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Association for Computational Linguistics, New Brunswick NJ.