Extracting Supertags from HPSG-based Tree Banks

Günter Neumann and Berthold Crysmann
LT-lab, DFKI &
Department of Computational Linguistics
Saarland University
Saarbrücken, Germany

This is a draft version, whose publication is under consideration for:
S. Bangalore and A. Joshi (eds): Complexity of Lexical Descriptions and its Relevance to Natural Language Processing: A Supertagging Approach, MIT press, in preparation.

Abstract

We describe a method for the automatic extraction of a Stochastic Lexicalized Tree Insertion Grammar from a linguistically rich HPSG Treebank. The extraction method is strongly guided by HPSG–based head and argument decomposition rules. The tree anchors correspond to lexical labels encoding fine–grained information. The approach has been tested with a German corpus achieving a labeled recall of 77.33% and labeled precision of 78.27%, which is competitive to recent results reported for German parsing using the Negra Treebank.

1 Introduction

In recent years several approaches have been proposed to improve the performance of Natural Language systems, which are based on the linguistic theory of Head–Driven Phrase Structure Grammars (HPSG, e.g., Kasper *et al.* [1995], van Noord [1997], Makino *et al.* [1998], Torisawa *et al.* [2000], Kiefer *et al.* [2002], Toutanova *et al.* [2002], and Toutanova and Manning [2002]). Common to all these approaches is that the coverage of the original general grammar is not affected. Thus if the original grammar is defined domain-independently it might also define a great many theoretically valid analyses covering a wide range of plausible linguistic constructions, including the rarest cases. However, in building real-world applications it has been shown to be a fruitful compromise to focus on frequency and plausibility of linguistic structures wrt. a certain domain. A number of attempts have been made to automatically adapt a general grammar to a corpus in order to achieve efficiency through domain-adaptation, e.g., using PATR-style grammars Briscoe and Carroll [1993]; Samuelsson [1994]; Rayner and Carter [1996] or lexicalized TAGs Srinivas [1997].

In Neumann and Flickinger [2002] and Neumann [2003] we applied the idea of a data–oriented approach for achieving domain-adaptation to HPSG.[1] We called this approach *HPSG–DOP*, because it had some strong corresponding relationships to the framework of Data–Oriented Parsing (DOP), cf. Bod *et al.* [2003]. The basic idea of HPSG–DOP is to parse all sentences of a representative training corpus using an HPSG grammar and parser in order to automatically acquire from the parsing results a *stochastic lexicalized tree grammar* (SLTG) such that each resulting parse tree is recursively decomposed into a set of subtrees. The decomposition operation is guided by the *head feature principle* of HPSG. Each extracted tree is automatically lexically anchored and each node label of the extracted tree compactly represents a set of relevant features by means of a simple symbol. A major drawback of this approach was that non–headed constructions were not factored out consequently due to the lack of structural refinements, e.g., recursive modifier constructions were restricted by the number of the largest embedding found in the corpus.

However in Hwa [1998], Neumann [1998], Xia [1999], Chen and Vijay-Shanker [2000], and Chiang [2000] a number of approaches for the automatic extraction of Tree Adjoining Grammars (TAGs) from treebanks are presented, which treat the factorization of modifier constructions more systematically. In particular, the approach developed by Chiang [2000] and further elaborated in Chiang [2004] is of interest, because his approach only requires a minimal amount of treebank preprocessing, and – more importantly – he interprets the head/argument rules exploited by Magerman [1995] and Collins [1997] as a heuristic for reconstructing full structural descriptions from partial ones rather than as a means for rearranging information in the training data, which eases the adaptation of his approach also to other natural languages and treebanks. For example, Bikel and Chiang [2000] apply this method to a Chinese treebank.

In this paper, we extend HPSG–DOP by combining it with Chiang's method and apply it on a linguistically rich HPSG treebank for German which is based on the recently developed Redwoods Treebank (cf. Oepen *et al.* [2002] and sec. 3). To our knowledge, our approach is the first time that a rich linguistic theory together with a stochastic TAG is applied to the German language. This is not a trivial task, as recently Dubey and Keller [2003] and Levy and Manning [2004] have shown that treebank parsing for German (using the rather shallow Negra Treebank) yields substantial lower performance compared to English Penn treebank parsing, probably due to the fact that differences in both languages and treebank annotation may be involved (cf. sec. 7). To give our new approach a name, we call it HPSG–Supertag following Srinivas [2003] who defines the elementary structures of a lexicalized TAG as *supertags*.

The rest of the paper is structured as follows. We begin by summarizing the formalism of the used tree structures. The HPSG treebank and details concerning the German HPSG are described in sec. 3, before in sec. 4 the method for the induction of the stochastic grammar from the HPSG treebank is described. This also includes a description of the

[1] A first initial approach for applying data–oriented methods to HPSG is described in Neumann [1994], where an approach for memory–based processing with HPSG based on Explanation–Based Learning is described.

HPSG–based head/argument rules used for the grammar reconstruction process. In sec. 5 we describe experiments using the standard PARSEVAL measurement. In sec. 7 important related work is discussed, before we conclude our chapter in sec. 8 by also giving an outline into future work.

2 Stochastic Lexicalized Tree Grammars

The set of lexically anchored trees extracted via the original HPSG–DOP method already characterizes a lexical tree–substitution grammar, i.e., a tree–adjoining grammar with no auxiliary trees, cf. Schabes [1990]. In Neumann [1998], and subsequently in Xia [1999], Chen and Vijay-Shanker [2000], and Chiang [2000] it is shown how tree adjoining grammars can be extracted from the Penn Treebank by performing a re–construction of the derivations using head–percolation rules. Here, we follow the approach developed in Chiang [2000], because his approach only requires a minimal amount of treebank preprocessing, which makes it easier to adapt it to other kind of treebanks.[2]

For efficiency reasons, a restricted form of lexicalized tree adjoining grammars is considered viz. lexicalized tree insertion grammars (LTIGs). LTIG has been introduced in Schabes and Waters [1995] as a TAG–formalism in which all auxiliary trees are either left or right auxiliary trees. No elementary wrapping auxiliary trees or elementary empty auxiliary trees are allowed. Furthermore, left (right) auxiliary trees cannot be adjoined to a node that is on the spine of an elementary right (left) auxiliary tree; and there is no adjunction allowed to the right (left) of the spine of an elementary left (right) auxiliary tree (cf. figure 1).

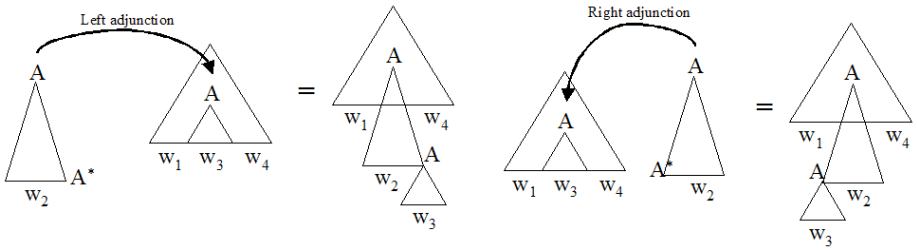

Figure 1: Left and right adjunction.

There is an additional tree composition operation called sister–adjunction used by Chiang [2000], which is based on the extended notion of TAG derivation introduced in Schabes and Shieber [1994]. In sister–adjunction, the root of a modifier tree is added as a new daughter to any other node, and multiple trees may be sister–adjoined at the same position. The main motivation for introducing this operation is its potential for deriving the flat structures found in the Penn Treebank (cf. figure 2). Note that in our case the

[2]And because his approach can be seen as a substantial improvement of the initial work we have layed out and described in Neumann [1998].

HPSG–derivations are deeply nested binary trees, so that sister–adjunction is actually not effective, however we leave it here for completeness.

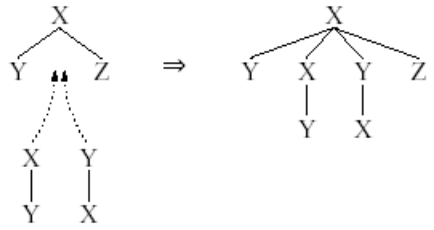

Figure 2: Sister adjunction.

The parameters of a probabilistic TAG which control the combination of trees by the substitution and adjunction are:

$$\sum_{\alpha} P_i(\alpha) = 1$$

$$\sum_{\alpha} P_s(\alpha \mid \eta) = 1$$

$$\sum_{\beta} P_a(\beta \mid \eta) + P_a(\text{NONE} \mid \eta) = 1$$

$$\sum_{\alpha} P_{sa}(\beta \mid \eta, i, X) + P_{sa}(\text{STOP} \mid \eta, i, X) = 1$$

where α ranges over initial trees, and β over auxiliary trees, and η over nodes. $P_i(\alpha)$ is the probability of beginning a derivation with α; $P_s(\alpha \mid \eta)$ is the probability of substituting α at η; $P_a(\beta \mid \eta)$ is the probability of adjoining β at η; $P_a(\text{NONE} \mid \eta)$ is the probability of nothing adjoining at η; $P_{sa}(\beta \mid \eta, i, X)$ is the probability of sister–adjoining, and $P_{sa}(\text{STOP} \mid \eta, i, X)$ is the probability of no further sister–adjunction. X is the root label of the previous tree to sister–adjoin at the site (η, i), or START if none. The probability of a derivation can then be expressed as the product of the probabilities of the individual operations of the derivation, cf. Chiang [2004] for more details.

LTIGs have context–free power and can be parsed in $O(n^3)$. Two parseres are available to us: a two–phase Early–style LTIG parser based on Schabes and Waters [1995] written in Lisp at our Lab, and a CKY–style bottom–up parser based on Schabes and Waters [1993] written in C by David Chiang. For the experiments reported in this paper in sec. 5, we are using David's parser, because currently, it is much faster than the Early–based Lisp parser, and can be handled much more flexible. The CKY–parser implements sister–adjunction, and uses a beam search, computing the score of an item $[\eta, i, j]$ by multiplying it by the prior probability $P(\eta)$. All items with score less than a given threshold compared to the best item in a cell are pruned (cf. Chiang [2000] and sec. 5 on details concerning the used parser settings).

3 HPSG TreeBank

The HPSG treebank (codename *Eiche*) we use in our study is based on a subset of the Verbmobil corpus which has been automatically annotated with a German HPSG grammar. The analyses provided by the grammar have then been manually disambiguated using the Redwoods treebanking technology, cf. Oepen *et al.* [2002].

The underlying HPSG grammar itself has originally been developed as a large-scale competence grammar of German by Stefan Müller and Walter Kasper in the context of the Speech-to-Speech machine translation project Verbmobil (see [Müller and Kasper, 2000]), and has subsequently been ported to the LKB (Copestake [2001]) and PET (Callmeier [2000]) processing platforms. In 2002, grammar development has been taken over by Berthold Crysmann. Since then, the grammar has undergone several major changes, most importantly the treatment of verb placement in clausal syntax [Crysmann, 2003].

3.1 Some basic properties of German syntax

The syntax of German features a variety of phenomena that makes syntactic analysis much harder than that of more configurational languages. Chief among these is the relative free word order in which syntactic arguments of a verb can appear within the clausal domain.

(1) a. weil der Lehrer dem Schüler das Buch schenkte
because the teacher.NOM the pupil.DAT the book.ACC donated
'because the teacher gave the book to the pupil as a present'

b. weil der Lehrer das Buch dem Schüler schenkte

c. weil dem Schüler der Lehrer das Buch schenkte

d. weil dem Schüler das Buch der Lehrer schenkte

e. weil das Buch der Lehrer dem Schüler schenkte

f. weil das Buch dem Schüler der Lehrer schenkte

Almost anywhere between the arguments modifiers can be interspersed quite freely.

(2) weil (gestern) der Lehrer (gestern) dem Schüler (gestern) das Buch (gestern) schenkte
because (yesterday) the teacher.NOM (yesterday) the pupil.DAT (yesterday) the book.ACC (yesterday) donated
'because yesterday the teacher gave the book to the pupil as a present'

This situation is further complicated by the combined effects of verb cluster formation and argument composition, which permit permutation even amongst the arguments of different verbs within the cluster.

(3) a. weil der Lehrer das Buch zu kaufen versprach
because the teacher.NOM the book.ACC to buy promised
'because the teacher promised him to buy the book.'

b. weil das Buch der Lehrer zu kaufen versprach
because the book.ACC the teacher.NOM to buy promised
'because the teacher promised him to buy the book.'

Furthermore, realisation of the verb cluster is often discontinuous, typically in matrix clauses.

(4) a. da versprach der Lehrer.NOM das Buch zu kaufen
there promised the teacher.NOM the book.ACC to buy
'There, the teacher promised him to buy the book.'

b. da versprach das Buch der Lehrer zu kaufen
there promised the book.ACC the teacher.NOM to buy
'There, the teacher promised him to buy the book.'

Assuming continuous constituents only, the argument structure is therefore only partially known in bottom-up parsing, until the other member of the discontinuous verb cluster is found.

In German matrix clauses, the finite verb typically surfaces in second position, the first position being occupied by some fronted, i.e. extracted, constituent. Thus, in contrast to English, presence of non-local dependencies is the norm, rather than the exception.

Taken together, permutation of arguments, modifier interspersal, discontinuous complex predicates and the almost categorial presence of non-local dependencies give rise to a considerable degree of variation in tree structure. As a consequence, we expect data-driven approaches to parsing to be more prone to the problem of data-sparseness. In the context of grammar induction from treebanks, it has already been observed, e.g., by Dubey and Keller [2003] that methods which are highly successful in a more configurational language, such as Collins PCFG parser for English (cf. Collins [1997]), give less optimal results when applied to German.

This problem is further enhanced by the fact that German is a highly inflectional language, with 4 distinct cases, 3 gender and 2 number distinctions, all of which enter into agreement relations. The same holds for the verbal domain, where up to 5 person/number combinations are clearly distinguished.

3.2 The grammar

In the spirit of HPSG as a highly lexicalised grammatical theory, most of the information about an items combinatorial potential is encoded in the lexical entries itself, in terms of typed feature structures. Syntactic composition is then performed by means of highly general rule schemata, again, implemented as typed feature structures, which specify the flow of information within syntactic structure. As a result, the DFKI German HPSG specifies only 87 phrase structure schemata, as compared to some 280+ leaf types for

the definition of parameterised[3] lexical entries, augmented by 56 lexical rules and 286 inflectional rules.

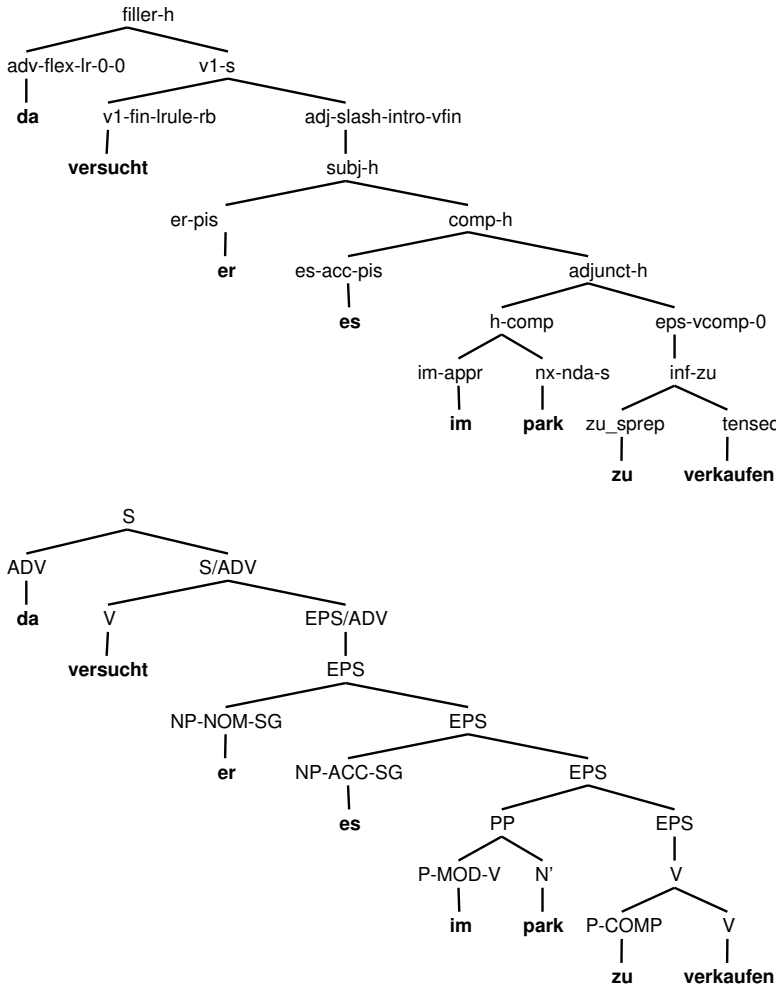

Figure 3: Examples of a derivation tree and its corresponding phrase tree representation. See text below for an explanation of the different symbols.

The rule schemata, which make up the phrase structure backbone of the HPSG grammar, correspond quite closely to principles of syntactic composition: by themselves they encode basic functional relations between daughter constituents, such as head-subject, head-complement, or head-adjunct, rather than intrinsic properties of the node itself. Thus, a rule like `h-comp` can be used to saturate a subcategorised complement of a preposition, a verb, or, a noun. Similarly, which constituents can function as the complement daughter of the `h-comp` rule is mainly determined by the information represented on the SUBCAT list

[3]Lexical entries may get further specialised beyond the information encoded in the lexical leaf type: typically, this includes subcategorisation for lexical case, selection of prepositional complements and verb particles, specification of auxiliary type (*have* vs. *be*), as well as sortal restrictions on complements.

of the lexical head. The rule schemata merely ensure that the subcategorisation constraints formulated by the head will actually be imposed on the complement daughter, and that the saturated valence requirement will be canceled off.

Since the underlying processing platforms (LKB/PET) do not currently support the segregation of immediate dominance and linear precedence, some rule schemata are further specialised according to the position of the head: alongside `h-adjunct`, `h-subj` and `h-comp` rules for verb-initial clauses and *pre*positional phrases, the grammar also defines their head-final counterparts (`adjunct-h`, `subj-h`, `comp-h`), required for verb-final clauses, adjectival phrases and *post*positional phrases. Within NPs some modifiers, e.g. adjectives are licensed by adjunct-h structures, whereas PPs are licensed in post-head position only. To summarise, the rules of the CF backbone provide crucial information about the position of the syntactic head, as well as the functional status of the non-head daughter.

Scrambling of complements is licensed in the German grammar by special lexical rules that permute the elements on a head's SUBCAT list. Modifier interspersal and scrambling across the subject are accounted for by permitting the application of h-subj, h-comp, and h-adjunct rules in any order.

Argument composition and scrambling of arguments from different verbs is captured by shuffling the SUBCAT lists of the upstairs and downstairs verb (e.g., `vcomp-h-0` ... `vcomp-h-4`). Discontinuous verb clusters are modelled by means of simulated verb movement (Müller and Kasper [2000] expanding an earlier idea proposed by Kiss and Wesche [1991]). Essentially, the subcategorisation requirements of the initial verb are percolated down the tree to be shuffled with those of the final verb.

Finally, extraction is implemented in a fairly standard way using slash feature percolation. Slash introduction is performed, at the gap site, by a unary rule. For subjects and complements, slash introduction saturates an argument requirement of the head by inserting its LOCAL value into the SLASH list. For adjuncts, the slash introduction also inserts a *local* object into SLASH, but since there is no valency to be saturated, it only semantically attaches the extracted modifier to the head. At the filler-site, SLASH specifications are retrieved, under unification: for semantic reasons, the grammar crucially distinguishes here between wh-fillers (`wh-h` rule) and non-wh-fillers (`filler-h` rule).

Besides these more basic constructions, the grammar also provides rule schemata for different types of coordinate structures, extraposition phenomena (Crysmann [in press]), dislocation, as well as some constructions more specific to German, such as auxiliary flip and partial VP fronting.

3.3 The treebank

The version of the HPSG formalism underlying the LKB and PET processing systems assumes continuous constituents only. Thus, the derivation tree of a sentence analysed by the grammar corresponds to a context free phrase structure tree. Given a grammar, the full HPSG analysis of a sentence can therefore always be reconstructed deterministically, once the derivation tree is stored together with the unique identifiers of the lexical entries on the terminal nodes. This fact is actually exploited by the Redwoods treebanking infrastructure

to provide a compact representation format. From the fully reconstructed feature structure representation of a parse, it is possible to extract additional derived structures: one such auxiliary structure that deserves particular mentioning is an isomorphic constituent tree decorated with more conventional node labels, such as S, NP, VP, PP, etc. These labels are obtained by testing the unifiability of a feature structure description against the AVM associated with the node, and assigning the label of the first matching description. Since these derived trees are isomorphic to the derivation history, the "functional" decorations provided by the rule backbone can be enriched straightforwardly with "categorial" information, providing for a very rich annotation.

As already mentioned before, the primary data used for the construction of the *Eiche* treebank are taken from the Verbmobil test corpora. In order to minimise duplication of annotation effort, only unique sentence strings have been incorporated into the treebank. Thus, redundancy in the data is limited to partial structures.

4 HPSG–Supertag Extraction from the Treebank

The main purpose of the grammar extraction process is twofold: 1) extract automatically all possible supertags, i.e., an LTIG, and 2) to obtain a maximum–likelihood estimation of the parameters of the extracted LTIG. The grammar extraction process actually re–constructs TAG derivations underlying the parse trees and is quite similar to the head–driven decomposition operation used in HPSG–DOP, but now adapted for the case of LTIG extraction.

4.1 The extraction method

Similar to Magerman [1995] and Chiang [2000], we use head–percolation and argument rules that classify for each node η exactly one child of η as the head and the others as either argument or modifier. However, as we will discuss below, our rules are based on HPSG and as such, are much more smaller in number and less heuristic in nature as those defined in Chiang [2000]. Using these rules, the derivations are re–constructed using the method described in Chiang [2000], and summarized here for your convenience:

- If η is an adjunct, excise the subtree rooted at η to form a modifier tree.
- If η is an argument, excise the subtree rooted at η to form an initial tree, leaving behind a substitution node.
- If η has a right corner θ which is an argument with the same label as η (and all intervening nodes are heads), excise the segment from η down to θ to form an auxiliary tree.

From the determined structures, supertags are generated in two steps: first the tree template (i.e., the elementary tree minus its anchor), then the anchor. From there, the

probabilities are decomposed accordingly and three back-off levels are computed, as described in Chiang [2000]. Furthermore, all words seen n or fewer times in training are treated as a single symbol UNKNOWN, in order to handle unknown words.

4.2 The rule definition

The following two tables contain the HPSG–based head and argument rules currently in use:

Parent:	Child:
SUBJ-H	last *
ADJUNCT-H	last *
COMP-H	last *
FILLER-H	last *
WH-H	last *
POS-ES	last *
DET-NBAR	last *
NP-NBAR	last *
VCOMP-H-0	last *
VCOMP-H-1	last *
VCOMP-H-2	last *
VCOMP-H-3	last *
VCOMP-H-4	last *
BINARY-COORD	last *
RECURSIVE-EV-COORD	last *
RECURSIVE-NOM-COORD	last *
*	first *

Table 1: Head rules for the HPSG Treebank. The symbol * stands for any label.

Parent:	Child:
SUBJ-H	first *
H-SUBJ	last *
COMP-H	first *
H-COMP	last *
H-COMP-EXTRAPOSED	last *
H-SUBJ-EXTRAPOSED	last *

Table 2: Arg rules for the HPSG Treebank. The symbol * stands for any label.

The list of rules is processed in the order specified and the first rule that fires is applied. A rule fires if the label of the current node matches with one of the parent node labels specified in the rule list. A head rule like "SUBJ-H last *" determines that the last child of a parent node with label `SUBJ-H` is the head, regardless of the child's label. The head rule "* first *" means that for a parent with an arbitrary node label its leftmost child is chosen as the head daughter. This rule plays the role of a default head rule. The argument rules work in the same way. For an explanation of the linguistic content of these rules, cf. sec. 3.

5 Experiments

We performed a ten–fold cross–validation over a corpus of 3528 sentences from the Verbmobil domain with an average sentence length of 7.2 words. We used "evalb"[4] to compute the standard PARSEVAL scores for our results, and focused on the Labeled Precision (LP) and Labeled Recall (LR) scores, as they are commonly used to rank parsing systems. During the experiments we used the same settings for the parser as used by Chiang [2000] for his Penn Treebank experimentations: a.) beam size set to 10^{-5}, b.) unknown word threshold set to 4.

The anchors of the extracted supertags consist of the preterminals of the derivation trees and are lexical labels. These are much more fine–grained than Penn Treebank preterminal tags, covering information about POS, morpho-syntactic, valence and other information. The input to the parser is then a sequence of pairs (LEX-LABEL WORDFORM) (i.e., we also ignore upper and lower case). For example for the sentence *Wann hätten Sie denn dann noch Zeit?* ('When would you have still time, then?'), the input to the parser is (see appendix A for the corresponding derivation and phrase tree represenations):

> (WANN-ADV WANN)(HABEN-T HAETTEN)(SIE_SH-PIS SIE)(DENN_SCADV-ADV DENN)(DANN_SCADV-VVPP DANN)(NOCH_PADV-ADV NOCH)(ZEIT-N ZEIT)

Since, we do not have any HPSG–based lexical tagger available, we used the (LEX-LABEL WORDFORM) sequence of each sentence extracted from the parse trees of the test corpus (i.e., from the 10% blind data used in each iteration of the cross–validation step). Note that the UNKNOWN symbol only relates to corresponding words in the training set (it maps words seen fewer than N times to this symobl), i.e., stems that only occur in the test set, but not in the training set, are not covered by the grammar. Hence, the parser will deliver no result for sentences which contain "out–of–vocabulary" stems.

We trained and tested our method on the full encoding of the symbols, which among others encode values for gender, number, person, case, tense and mood. Furthermore, the symbols also encode the valency of verbs. The number of different node labels is 2069;

[4] http://nlp.cs.nyu.edu/evalb/

appendix B shows the top fifty most frequent symbols in the corpus (together with the frequency count).

It seems clear that using lexical labels as anchors will effect at least the coverage and recall. In order to test this, we also run an experiment, where we used only the Part–of–Speech (POS) of the lexical labels, which are retrieval from the yield of the corresponding phrase tree (but note, that the labels of all non–terminal nodes are labels from the derivation trees). This will lead to a much more coarse–grained classification of word forms, but probably also to a less restrictive tree selection. The table below presents our current results, where we computed the average values for the labeled recall and precision results determined in all ten iterations:

Anchor	Cov.	LR(tot.)	LP(tot.)	LR(cov.)	LP(cov.)
LEXICAL LABEL	77.47	57.68	77.07	77.33	78.27
POS	98.12	76.42	78.36	77.92	78.44

where LR(tot.)/LP(tot.) is measured over all sentences, and LR(cov.)/LP(cov.) over the parsed sentences, i.e., for sentences without out–of–vocabulary stems.

6 Discussion

In Neumann [2003] we discussed initial results for our HPSG–DOP approach using an English HPSG grammar and a less detailed analysis. For example instead of a cross–validation analysis, we used 1000 Verbmobil sentences for training, and another 1000 for testing. We did not measure recall and precision, but checked whether it was possible to expand each one, by unifying the feature constraints of the original HPSG grammar. Thus seen, we consider a sentence analysis as valid, if it is consistent with respect to the HPSG constraints (including all lexical constraints, of course). Following this way, 704 sentences were recognized which corresponds to a coverage of 70.4% .

Our current results suggest, that the HPSG–SUPERTAG method is superior compared to our earlier work. It should be clear, that the moderate coverage is basically due to the very specific nature of the tree anchors. The size of the current corpus is small compared to the Penn Treebank, so we assume that it will improve for larger corpora. Note also, that the redundancy in the corpus is limited to partial structures (see also sec. 3.3), which also effects the performance.

Nevertheless, our current results are encouraging, if we compare them with other recent approaches of probabilistic parsing for German. To date, there is only little work on full probabilistic parsing of German from treebanks — mainly using the Negra Treebank (Skut *et al.* [1997]) — and the PARSEVAL measurement. The first probabilistic treebank parser for German (using the Negra Treebank) is presented in Dubey and Keller [2003]. They obtain (for sentence length of ≤ 40): LR=71.32% and LP=70.93% (coverage = 95.9%). Müller *et al.* [2003] also present a probabilistic parser for Negra. They study the consequences that the Negra implies for probabilistic parsing, and concentrate on

the role of two factors (1) lexicalization and (2) grammatical functions. The results they report: LR=71.00% and LP=72.85% (coverage = 100%). Furthermore, Levy and Manning [2004] present experiments on probabilistic parsing using Negra concentrating on non–local dependency reconstruction. Their results also suggest that current state–of–art statistical parsing is far better on Penn Treebank than on the Negra Treebank.

7 Related Work

Here, we will briefly mention and discuss other relevant work, in addition to the already discussed work done for LTAG and German treebank parsing.

Our approach is also related to approaches of grammar specialization based on Explanation-based Learning (EBL), cf. Samuelsson [1994]; Srinivas and Joshi [1995]; Rayner and Carter [1996], and other grammar approximation methods. For example, Krieger [2005], presents an approximation method that specialized an English HPSG–grammar to a probabilistic context–free grammar. In Keselj and Cercone [2002] an interesting approach called "just–in–time subgrammar extraction" is presented, which has some ideas in common with our HPSG–DOP approach, but differs in that they perform a subgrammar extraction in form of a PCFG on–line for a piece of text, rather than off–line for a specific domain.

In Foth *et al.* [2004] a broad–coverage parser for German based on weighted constraint dependency grammar is presented and analysed using the Negra Treebank. In order to evaluate their parsing result with the Negra parse trees, the phrase-based trees are mapped to dependency trees. Then the accuracy is measured by counting the number of correctly computed dependency edges. Using the same subset of Negra sentences for testing as done by Dubey and Keller [2003], they report an labeled edge accuracy of 87%.

Current stochastic approaches for HPSG basically focus on parse tree disambiguation using the English Redwoods Treebank, cf. Oepen *et al.* [2002]. For example, Toutanova *et al.* [2002], present a parse selection method using conditional log-linear models built over the levels of derivation tree, phrase structure tree, and semantic dependency graph in order to analyse the effect of different information levels represented in the Redwoods Treebank. The best reported result (in terms of accuracy) is obtained for the derivation tree representation and by implementing an extended PCFG that conditions each node's expansion on several of its ancestors in the derivation tree (with a manually specified upper bound of 4 ancestors). They report an exact parse accuracy of 81.80% for such an extended PCFG, which was only slightly improved when combining it with a PCFG based on the semantic dependency graph representation (82.65%). In Toutanova and Manning [2002] this work is extended by the integration of automatic feature selection methods based on decision trees and ensembles of decision trees. Using this mechanism, they are able to improve the parse selection accuracy for the derivation tree based PCFG from 81.82% to 82.24%.

8 Conclusion and Future Work

We have presented an approach of extracting supertags from a HPSG–based treebank, and have evaluated the performance of the grammar using a stochastic LTIG parser. In future work, we will consider the following aspects. First, we will explore how the current results can be improved by either adding more information to the tree labels or by generalizing those tree labels which are currently too specific. Second, we will investigate how this technology can be used to provide the N–best derivation trees and to use them as input for the deterministic feature structure expansion step using the HPSG–source grammar. In this way, a preference–based parsing schema for HPSG using a treebank model will function as a filter.

Acknowledgements

The work presented here was partially supported by a research grant from the German Federal Ministry of Education, Science, Research and Technology (BMBF) to the DFKI project `Quetal` (FKZ: 01 IW C02) and the EC–funded project DeepThought. We thank very much David Chiang for making available to us his TIG training and parsing system. We also thank the Redwoods Treebank team for making their tools open–source, and especially Stephan Oepen for his kind support.

A Appendix

The treebank representation of the derivation tree and the phrase tree for the sentence *Wann hätten Sie denn dann noch Zeit?* ('When would you have still time, then?'):

```
(1347 WH-H -2.03203 0 7 (44 PX-ALL_INFL_RULE 1.98206 0 1
 (8 WANN-ADV 0 0 1 ("wann" 0 1)))
 (1137 ADJ-SLASH-INTRO-VINI -3.15029 1 7
  (1133 H-COMP -1.90074 1 7
   (610 H-ADJUNCT 4.23003 1 6
    (492 H-ADJUNCT 6.68481 1 4
     (171 H-SUBJ 2.84761 1 3
      (40 V1-FIN-LRULE-NO-RB 2.49282 1 2
       (39 PERS-SILR-LRULE 3.79787 1 2
        (38 VX-PAST-CONJ-PL-1-3_INFL_RULE 3.79787 1 2
         (21 HABEN-T 0 1 2 ("haetten" 1 2)))))
      (98 PX-ALL_INFL_RULE -2.2774 2 3
        (27 SIE_SH-PIS 0 2 3 ("sie" 2 3))))
     (46 PX-ALL_INFL_RULE 1.63737 3 4
       (29 DENN_SCADV-ADV 0 3 4 ("denn" 3 4))))
    (54 ADJUNCT-H 2.2854 4 6
      (50 PX-ALL_INFL_RULE 1.63737 4 5
      (32 DANN_SCADV-VVPP 0 4 5 ("dann" 4 5)))
     (48 PX-ALL_INFL_RULE 1.63737 5 6
       (33 NOCH_PADV-ADV 0 5 6 ("noch" 5 6)))))
   (860 EMPTY-DET-SG -5.40906 6 7
     (96 NX-FEM-SG_INFL_RULE -1.66928 6 7
       (35 ZEIT-N 0 6 7 ("zeit" 6 7)))))))

(S (ADV
 (ADV (wann)))
 (S/ADV
  (S
   (V (V (V (V (V (V (V (haetten)))))
     (NP-NOM-PL (NP-NOM-PL (sie))))
       (ADV (ADV (denn))))
    (ADV (ADV (ADV (dann)))
      (ADV (ADV (noch)))))
   (NP-ACC-SG (N' (N' (zeit)))))))
```

B Appendix

The top fifty fully encoded symbols together with their frequency counts:

PX-ALL_INFL_RULE 13143 H-COMP 4095 ADJUNCT-H 3499
PERS-SILR-LRULE 3475 H-ADJUNCT 2768 FILLER-H 2293
NX-MAS-NDA-SG_INFL_RULE 1943 DET-NBAR 1940
V1-FIN-LRULE-NO-RB 1890
V1-S 1284 V1-FIN-LRULE-RB 1284 SUBJ-SLASH-INTRO 1244
H-SUBJ 1228 ADV-FLEX-LR-0-0 1167 EMPTY-DET-SG 1157
FULL-PREP-NOUN-TO-VERB-MOD-LRULE 1096 DET-D-DET 1093
EPS-VCOMP-0 1027 ICH-PIS 988 SUBJ-H 982
EMPTY-NOUN-MODIFIER-RULE 895 COMP-H 892
NX-FEM-SG_INFL_RULE 887
NON-FIN-SILR-LRULE 822 VX-PAST-CONJ-SG-1-3_INFL_RULE 803
ADJ-SLASH-INTRO-VINI 775 DX-INFL-EN_INFL_RULE 748
VX-PRES-PL-1-3_INFL_RULE 747 TENSED-NON-FIN-LRULE 741
VX-SUP-BARE_INFL_RULE 727 WIR-PIS 722
INTERJECTION-RULE 711 AX-POS-NULL_INFL_RULE 684
DEF-PREP-NOUN-TO-VERB-MOD-LRULE 661 DAS-NP-NEU-SG-NA 600
CARDINAL_INFL_RULE 519 VX-PRES-IND-SG-3_INFL_RULE 511
ADJ-SLASH-INTRO-VFIN 498 JA_INT-ADV 455
EIN_QUA-ADJA 448 VX-PRES-SG-1_INFL_RULE 447
DA_SCOP 430 DANN_ADV-VVPP 413 ES-NOM-PIS 402
NX-NEU-NDA-SG_INFL_RULE 391 MIR-PIS 391
COMP-SLASH-INTRO 391 AM-APPRART 389
DX-DEF-EN_INFL_RULE 387 DX-DEF-EN_INFL_RULE 369

References

D. Bikel and D. Chiang. Two statistical parsing models applied to the chinese treebank. In *In Proceedings of the Second Chinese Language Processing Workshop*, page 16, 2000.

Rens Bod, Khalil Sima'an, and Remko Scha, editors. *Data Oriented Parsing*. Center for Study of Language and Information (CSLI) Publications, Stanford:CA, USA, 2003.

T. Briscoe and J. Carroll. Generalized probabilistic lr parsing of natural language (corpora) with unification-based grammars. *Computational linguistics*, 19(1):25–59, 1993.

Ulrich Callmeier. PET — a platform for experimentation with efficient HPSG processing techniques. *Journal of Natural Language Engineering*, 6(1):99–108, 2000.

J. Chen and K. Vijay-Shanker. Automated extraction of tags from the penn treebank. In *6th International Workshop on Parsing Technologies (IWPT'2000)*, Trento, Italy, 2000.

D. Chiang. Statistical parsing with an automatically–extracted tree adjoining grammar. In *38th ACL*, Honk Kong, 2000.

D. Chiang. *Evaluating Grammar Formalisms for Applications to Natural Language Processing and Biological Sequence Analysis*. PhD thesis, University of Pennsylvania, 2004.

Michael Collins. Three generative, lexicalised models for statistical parsing. In *ACL*, pages 16–23, 1997.

Ann Copestake. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford, 2001.

Berthold Crysmann. On the efficient implementation of German verb placement in HPSG. In *Proceedings of RANLP 2003*, pages 112–116, Borovets, Bulgaria, 2003.

Berthold Crysmann. Relative clause extraposition in German: An efficient and portable implementation. *Research on Language and Computation*, in press.

Amit Dubey and Frank Keller. Probabilistic parsing for german using sister-head dependencies. In *ACL*, pages 96–103, 2003.

K. Foth, M. Daum, and W. Menzel. A broad-coverage parser for german based on defeasible constraints. In *Konvens, Vienna*, pages 45–52, 2004.

R. Hwa. An empirical evaluation of probabilistic lexicalized tree insertion grammars. In *Proceedings of the 36th ACL and 17th Coling*, Montreal, 1998.

R. Kasper, B. Kiefer, K. Netter, and K. Vijay-Shanker. Compilation of hpsg into tag. In *33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA, 1995.

V. Keselj and N. Cercone. Just-in-time subgrammar extraction for hpsg. In *In Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'01, Kitakyushu, Japan.*, 2002.

B. Kiefer, H. Krieger, and D. Prescher. A novel disambiguation method for unification-based grammars using probabilistic context-free approximations. In *In Proceedings of COLING 2002*, 2002.

Tibor Kiss and Birgit Wesche. Verb order and head movement. In Otthein Herzog and Claus-Rolf Rollinger, editors, *Text Understanding in LILOG*, number 546 in Lecture Notes in Artificial Intelligence, pages 216–240. Springer-Verlag, Berlin, 1991.

Hans-Ulrich Krieger. Grammar approximation as a speed-up technique for unification-based parsing. In Henning Fernau, editor, *Invited talks on Learning of Automata and Grammars at the Workshop on Theoretical Aspects of Grammar Induction (TAGI)*, pages 10–21, 2005.

Roger Levy and Christopher D. Manning. Deep dependencies from context-free statistical parsers: Correcting the surface dependency approximation. In *ACL*, pages 327–334, 2004.

D. Magerman. Statistical decisiontree models for parsing. In *In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 276–283, 1995.

T. Makino, M. Yoshida, K. Torisawa, and J. Tsujii. Lilfes – towards a practical hpsg parser. In *Proceedings of the 36th ACL and 17th Coling*, pages 807 – 811, Montreal, 1998.

Stefan Müller and Walter Kasper. Hpsg analysis of German. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, Artificial Intelligence, pages 238–253. Springer-Verlag, Berlin Heidelberg New York, 2000.

K. Müller, D. Prescher, and K. Sima'an. Grammatical functions and parsing the german negra treebank. In *Slides from CLIN 2003 presentation, http://staff.science.uva.nl/ kmueller/Onlinepapers/CLIN03_slides.pdf*, 2003.

G. Neumann and D. Flickinger. Hpsg-dop: Data-oriented parsing with hpsg. In *Proceedings of the 9th International Conference on HPSG (HPSG-2002), August 8-9*, Kyung Hee University, Seoul, 2002.

G. Neumann. Application of explanation-based learning for efficient processing of constraint-based grammars. In *Proceedings of the 10th IEEE Conference on Artificial Intelligence for Applications, March 1-4*, pages 208–215, San Antonio, Texas, USA, 1994.

G. Neumann. Automatic extraction of stochastic lexicalized tree grammars from treebanks. In *4th workshop on tree-adjoining grammars and related frameworks*, Philadelphia, PA, USA, August 1998.

G. Neumann. Data-driven approaches to head-driven phrase structure grammar. In Rens Bod, Remko Scha, and Khalil Sima'an, editors, *Data Oriented Parsing*. CSLI Publications, University of Chicago Press, Stanford:CA, USA, 2003.

S. Oepen, K. Toutanova, S. Shieber, C. Manning, D. Flickinger, and T. Brants. The lingo redwoods treebank: Motivation and preliminary applications. In *COLING*, 2002.

M. Rayner and D. Carter. Fast parsing using pruning and grammar specialization. In *34th Annual Meeting of the Association for Computational Linguistics*, Morristown, New Jersey, 1996.

C. Samuelsson. Grammar specialization through entropy thresholds. In *Proceedings of the 32nd Annual Meeting of the Association forComputational Linguistics*, pages 188–195, 1994.

Yves Schabes and Stuart M. Shieber. An alternative conception of tree-adjoining derivation. *Computational Linguistics*, 20(1):91–124, 1994.

Y. Schabes and R. Waters. Stochastic lexicalized context-free grammar. In *Proceedings of the 3rd International Workshop on Parsing Technologies (IWPT'93), Tilburg, The Netherlands*, pages pages 257–266, 1993.

Y. Schabes and R. Waters. Tree insertion grammar: A cubic-time parsable formalism that lexicalizes context-free grammar without changing the trees produced. *Computational Linguistics*, 21:479–513, 1995.

Y. Schabes. *Mathematical and Computational Aspects of Lexicalized Grammars*. PhD thesis, University of Pennsylvania, Philadelphia, USA, 1990.

W. Skut, B. Krenn, T. Brants, and H. Uszkoreit. An annotation scheme for free word order languages. In *Proceedings of the fifth conference on Applied natural language processing table of contents Washington, DC*, pages 88–95, 1997.

B. Srinivas and A. Joshi. Some novel applications of explanation-based learning to parsing lexicalized tree-adjoining grammars. In *33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA, 1995.

B. Srinivas. *Complexity of Lexical Restrictions and Its Relevance to Partial Parsing*. PhD thesis, University of Pennsylvania, 1997. IRCS Report 97–10.

B. Srinivas. Localizing dependencies and supertagging. In Rens Bod, Remko Scha, and Khalil Sima'an, editors, *Data Oriented Parsing*. CSLI Publications, University of Chicago Press, Stanford:CA, USA, 2003.

K. Torisawa, N. Nishida, Y. Miyao, and J. Tsujii. An hpsg parser with cfg filtering. *Journal of Natural Language Engineering*, 6(1):63–80, 2000.

K. Toutanova and C. Manning. Feature selection for a rich hpsg grammar using decision trees. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL)*, 2002.

K. Toutanova, C. Manning, S. Shieber, D. Flickinger, and S. Oepen. Parse disambiguation for a rich hpsg grammar. In *In First Workshop on Treebanks and Linguistic Theories (TLT2002)*, pages 253–263, 2002.

G. van Noord. An efficient implementation of the head-corner parser. *Computational Linguistics*, 23:425–456, 1997.

F. Xia. Extracting tree adjoining grammars from bracketed corpora. In *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium(NLPRS-99)*, Beijing, China, 1999.