Evaluating and Extending the Coverage of HPSG Grammars: A Case Study for German

Jeremy Nicholson†‡, Valia Kordoni†, Yi Zhang†, Timothy Baldwin‡, Rebecca Dridan†

†Dept of Computational Linguistics and DFKI GmbH, Saarland University, Germany
‡Dept of Computer Science and Software Engineering and NICTA, University of Melbourne, Australia
{jeremy,kordoni,yzhang,rdrid}@coli.uni-sb.de
tim@csse.unimelb.edu.au

Abstract

In this work, we examine and attempt to extend the coverage of a German HPSG grammar. We use the grammar to parse a corpus of newspaper text and evaluate the proportion of sentences which have a correct attested parse, and analyse the cause of errors in terms of lexical or constructional gaps which prevent parsing. Then, using a maximum entropy model, we evaluate prediction of lexical types in the HPSG type hierarchy for unseen lexemes. By automatically adding entries to the lexicon, we observe that we can increase coverage without substantially decreasing precision.

1. Introduction

Deep lexical grammars have been used in applications such as machine translation and information extraction, where they have been useful because they produce semantic structures which provide more information than shallower tools such as chunkers and dependency parsers. However, as many deep grammars tend to emphasise precision over recall, their coverage can be low, hence they are not considered practical for certain applications.

DELPH-IN (Oepen et al., 2002) is an initiative which attempts to develop and enhance broad–coverage "precision grammars" based on the Head–driven Phrase Structure Grammar formalism (HPSG: Pollard and Sag (1994)). We examine and attempt to improve the coverage of HPSG grammars, building on the work of Baldwin et al. (2004), who used the English Resource Grammar (ERG: Copestake and Flickinger (2000)) to parse a fragment of the British National Corpus (BNC: Burnard (2000)). We repeat the experiment using GG, a German HPSG grammar (Müller and Kasper, 2000; Crysmann, 2005) showing that their results are also applicable to German. Using this information, we carry out deep lexical acquisition experiments and evaluate their efficacy.

2. Background

Aiming to isolate the causes of parse failure and identify the "low–hanging fruit" for automatic improvement of the grammar, Baldwin et al. (2004) used the ERG[1] to parse a random sample of 20K sentences from the written component of the BNC. Because parsing fails when a lexical item in the sentence is not attested in the lexicon, this sample was taken from the 32% of the corpus which had a full lexical span. The grammar generates at least one parse for 57% of the sentences with lexical span, and 83% of these have at least one correct parse attested.

Baldwin et al. then proceed to analyse the lexical gaps of open–class words with an eye toward lexical expansion. They observe that nouns present the most promising class, especially in the context of the rich structure of the lexical type hierarchy. Finally, they conclude that some of the information can be obtained automatically.

This form of automatic extension of a deep lexical resource is referred to as deep lexical acquisition. In our case, we are interested in extending the lexicon for an HPSG grammar; however, none of the methods used are specific to that formalism. Baldwin (2005) looked at *in vitro* and *in vivo* methods for lexical type prediction of unknown words. Zhang and Kordoni (2006) attempted to mine errors in the lexicon of the ERG using the method of van Noord (2004). They go on to add lexical entries for likely multi–word expressions in Zhang et al. (2006), increasing coverage for sentences having this phenomenon. The wider impact of lexical acquisition on grammar coverage for corpus text, however, remains unclear.

3. Evaluating GG

Like the ERG, the German HPSG grammar GG[2] has been in development for the better part of a decade. We ran the parser over a corpus of the Frankfurter Rundschau, for about 612K sentences consisting of between 5 and 20 tokens. The number of sentences with full lexical span was about 28%, with about 42% of these having at least one parse.

	No span	Span, no parse	≥ 1 parse
GG	72%	16%	12%
ERG	68%	14%	18%

Table 1: Comparison of the sentence distribution when parsing using the two HPSG grammars.

The proportion of sentences with a full lexical span is reasonably close. We did an analysis of the lexical gaps for GG over a randomly selected set of 1000 sentences, and discovered that the gaps fell into a number of categories: missing lexical entries, proper nouns, unattested noun compounds, punctuation and tokenisation errors, and garbage strings. The error distribution is shown in Table 2.

[1] As of July, 2003.

[2] As of March, 2007.

Error Type	Proportion
lexical entries	33%
proper nouns	22%
noun compounds	30%
tokenisation	12%
garbage strings	2%

Table 2: Distribution of the types of lexical gaps in the GG lexicon.

Unlike English, proper nouns are difficult to identify in German because all nouns are capitalised. Also, noun compounds are multi–word expressions in English, but simplex constructions in German. The lexicalisation of noun compounds in the grammar means that their presence artificially inflates the proportion of lexical gaps[3].

We proceeded to analyse the causes of gaps in the parse coverage, that is, what phenomena in the grammar would cause a grammatical sentence to have no parses found. These also came in a number of flavours: a grammatical construction was missing from the grammar, a lexical item was attested in the lexicon, but without the requisite lexical type (e.g. a noun only had a verbal entry), the sentence had a multi–word expression unattested in the lexicon, old–style spelling was used, or the sentence was a fragment. The distribution of these is shown in Table 3.

Error type	Proportion
constructional gap	39%
lexical item gap	47%
multi–word expression	7%
spelling	4%
fragment	3%

Table 3: The distribution of parse gaps for GG over the corpus.

These values generally agree with those observed in Baldwin et al. (2004), where about 40% of errors were observed for both missing constructions and missing lexical entries. Furthermore, manual evaluation showed a correct parse was attested in 85% of sentences that had at least one parse: also comparable to the results observed on the ERG.

4. Lexical Acquisition

We then attempted to hypothesise lexical entries for GG in the manner of Baldwin (2005). This is construed as a classification task, where a feature set is constructed for a lexeme, and a class is predicted from the set of leaf lexical types for the grammar. A fragment of the lexical type hierarchy above the count noun class is shown in Figure 1.

Lexical type classification is similar to POS tagging, but with a much more refined tagset, which accounts for various syntactic and semantic categorisations. For example, the noun family partly shown below has subtypes for count nouns (as distinct from mass nouns), deverbal nouns, adpositional information like PPs or compounding modifiers, and numerous proper noun classes including names of people, places, times, dates, and holidays. Morphological information like grammatical gender and number is typically included within the feature structure.

Baldwin contrasted a range of morphological, syntactic, and semantic properties for class prediction of lexical types for both word types and tokens, and concluded that they were all comparable. We also discovered that a range of methods resulted in similar performance; hence we used a simple feature set similar to the one in Zhang and Kordoni (2006).

Affix features	K, Ka, Kat, Katz, e, ze, tze, atze
Context features	Die, DT, ist, VBZ, schwarz, JJ

Table 4: Features for the token *Katze* in the sentence *Die Katze ist schwarz und hübsch.* Lexical types are shown as Penn–style POS tags for simplicity.

The features correspond to prefixes and suffixes of the lexeme from length 1 to 4, and two tokens of context to the left and right, where available. We also include the lexical types for the contextual tokens, which may be available when prediction is only required for single unknown words. These are shown as POS tags in the table above, but actually look more like `adj-prd-le` for the predicating adjective *schwarz*; the target lexical type for *Katze* above is the count noun leaf type `count-noun-le`.

We examined prediction of open–class tokens within the treebank associated with the given version of GG. The treebank consisted of about 11K sentences comprising about 75K tokens, of which about 28K tokens were open–class. 10–fold cross–validation was used to evaluate a maximum entropy model developed using the OpenNLP package[4]. The tokens were randomly allocated to 10 equally sized folds, and each fold was used as a test set with the remaining 9 folds as training data. The ten prediction accuracies were then macro–averaged. As tokens which appeared in both the training and test data were artificially easy to classify, we restricted evaluation to "unknown words," that is, those tokens in the test set whose wordform–lexical type pair was not attested in the training data. Each test fold presented about 100 unknown words.

The accuracy results are shown in Table 5, where 0 words of context is contrasted with 2 tokens of context, and 2 tokens with lexical types (like `adj-prd-le`), and affixes up to length 0, 2, and 4. Increasing the feature space was not observed to improve results further.

Affix Length	Context		
	0	2 tokens	2 types
0	-	0.38	0.42
2	0.40	0.48	0.54
4	0.50	0.55	0.58

Table 5: Cross–validation accuracy when predicting lexical types in German

These results are slightly higher than those observed for the ERG in a corresponding experiment by Zhang and Ko-

[3] Using a sophisticated tokeniser, like TreeTagger (Schmid, 1994), better estimates could be found using more accurate tokenisation of compounds and proper nouns.

[4] http://maxent.sf.net

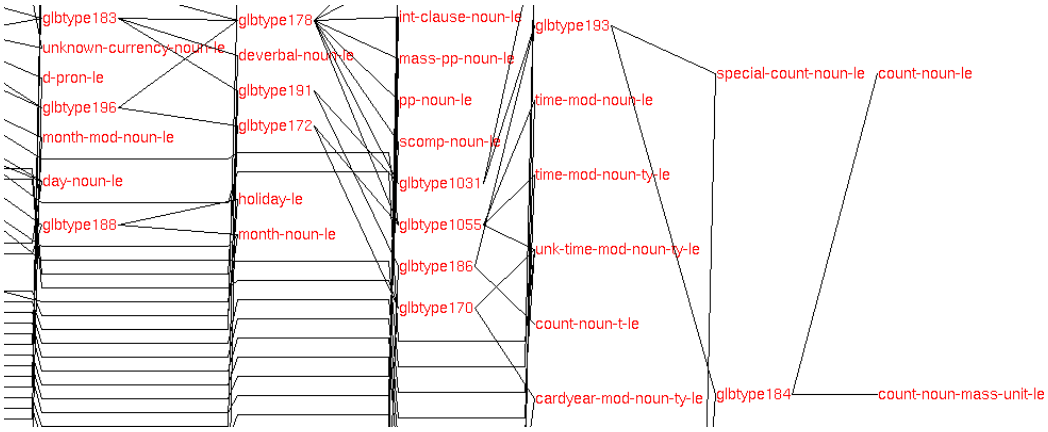

Figure 1: A fragment of the lexical type hierarchy of GG.

rdoni (2006). One possible reason for this is the number of classes (leaf lexical types), which is about 800 for the ERG and only about 400 for GG.

5. Lexicon Extension

One persistent question for deep lexical acquisition is the real impact on the deep lexical resource. While we can observe close to 60% type–level accuracy on the difficult task of assigning one of hundreds of lexical types to a lexeme, it is unclear how this impacts the grammar itself. To analyse this, we used the method of deep lexical acquisition described in Section 4 on the German corpus data analysed in Section 3.

Briefly, for the sentences in the segment of the Frankfurter Rundschau that lacked a full lexical span in our analysis, we generated features as in Table 4. Lexical types were not trivially available, so we used only the wordforms as contextual features. This was then used as test data for the maximum entropy model, and the entire treebank was used as training data, admittedly coming from a different domain. We assumed that all unknown words observed in the corpus were from open classes: determiners that were missing from the lexicon, for example, would not be predicted by the model.

The corpus data was not marked for HPSG lexical types, so we cannot directly evaluate the accuracy of the lexical type prediction, but we expect it to be similar to the results in Section 4. To see how this low accuracy affected the end parsing result, we took the most likely tag as output by the machine learner, thresholded at 10% likelihood, and generated lexical entries where possible for the predicted lexemes. This resulted in about 1130 entries which we added to the initial lexicon (of about 35K entries), out of about 1400 unknown tokens in the sentences.

The net effect was that a further 87 sentences (about 9%) displayed at least one parse. Further examination showed that 83% of these sentences had at least one correct parse within the added parses: only slightly less than that of the original lexicon. Errors were divided fairly evenly between the wrong lexical type being predicted and parser failure in spite of a seemingly correct lexical type.

So, deep lexical acquisition raised the coverage of the parser from about 12% at 85% precision to about 20% at 84% precision (the fraction of sentences having at least one correct parse). Considering the low expected accuracy of the lexical type predictor, this performance is remarkably high — indicating that it is, in fact, the easy sentences that are being recovered. Closer inspection shows that this is indeed the case: the extra sentences were mostly short and simple (with a missing entry for a count noun, for example). These results complement those found in Zhang et al. (2007).

6. Conclusion

We observe a striking similarity with the results of Baldwin et al. (2004), reinforcing our impressions of the language–independence of the methods. Indeed, we feel that the observations are widely applicable for other languages and formalisms; the analysis and features we use are not reliant on German, HPSG, or maximum entropy models.

We observe that the current GG has much in common with the ERG at the time of the study by Baldwin et al. (2004). Since then, the ERG has undergone considerable extension (constituting at least an increase in coverage), spurred on by a number of conclusions in that work. The observations here can be used to similarly improve GG, and to some extent lexical gaps in deep grammars in general.

Lexical type prediction, despite having moderate intrinsic accuracy, was shown to improve the performance of the grammar: the coverage nearly doubled with only a slight drop in precision. Although the token–wise improvements were observed primarily on simple sentences, the resulting increase to the lexicon may potentially lead to a resource which is more useful for future parsing.

There are numerous possible extensions, beginning with the obvious extension of the scope of the study, in terms of analysing a larger proportion of the parsed sentences, and aiding the grammar writer in covering lexical and constructional gaps. A similar analysis for the Japanese HPSG grammar, JACY (Siegel and Bender, 2002), again with external newspaper corpora, could validate the observations for a very different language with little morphology to aid class prediction.

In terms of the lexical type prediction analysis, we could examine using POS tags instead of lexical types for the contextual tokens, as the latter are not always readily available.

Examining only left context would be more in the spirit of online supertagging.

Another rich source of data for grammar expansion would be error mining, where constructions which are difficult to parse could be extracted from the sentences which had a lexical span, but no parses attested by the parser.

7. References

Timothy Baldwin, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2004. Road-testing the English Resource Grammar over the British National Corpus. In *Proc. of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2047–50, Lisbon, Portugal.

Timothy Baldwin. 2005. Bootstrapping deep lexical resources: Resources for courses. In *Proc. of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pages 67–76, Ann Arbor, USA.

Lou Burnard. 2000. *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Ann Copestake and Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proc. of the Second Conference on Language Resources and Evaluation*, Athens, Greece.

Berthold Crysmann. 2005. Relative clause extraposition in German: An efficient and portable implementation. *Research on Language and Computation*, 3(1):61–82.

Stefan Müller and Walter Kasper. 2000. HPSG analysis of German. In *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 238–253. Springer, Berlin, Germany.

Stephan Oepen, Dan Flickinger, Jun'ichi Tsujii, and Hans Uszkoreit, editors. 2002. *Collaborative Language Engineering. A Case Study in Efficient Grammar-Based Processing*. CSLI Publications, Stanford, USA.

Carl Pollard and Ivan Sag. 1994. *Head–Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, USA.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proc. of the 1st International Conference on New Methods in Language Processing*, Manchester, UK.

Melanie Siegel and Emily Bender. 2002. Efficient deep processing of Japanese. In *Proc. of the 2002 COLING Workshop on Asian Language Resources and International Standardisation*, Taipei, Taiwan.

Gertjan van Noord. 2004. Error mining for wide-coverage grammar engineering. In *Proc. of the 42nd Annual Meeting of the ACL*, pages 446–453, Barcelona, Spain.

Yi Zhang and Valia Kordoni. 2006. Automated deep lexical acquisition for robust open texts processing. In *Proc. of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 275–280, Genoa, Italy.

Yi Zhang, Valia Kordoni, Aline Villavicencio, and Marco Idiart. 2006. Automated multiword expression prediction for grammar engineering. In *Proc. of the COLING/ACL 2006 Workshop on Multiword Expressions*, pages 36–44, Sydney, Australia.

Yi Zhang, Timothy Baldwin, and Valia Kordoni. 2007. The corpus and the lexicon: Standardising deep lexical acquisition evaluation. In *Proc. of the ACL 2007 Workshop on Deep Linguistic Processing*, pages 152–159, Prague, Czech Republic.