In K-Y. Su, J. Tsujii, J-H. Lee, et al. (eds.) Natural Language Processing – IJCNLP 2004: First International Joint Conference, Hainan Island, China, March 22–24, 2004, Revised Selected Papers. Springer Lecture Notes in Computer Science, Volume 3248, 2005 ©Springer-Verlag

The Automatic Acquisition of Verb Subcategorisations and their Impact on the Performance of an HPSG Parser

John Carroll and Alex C. Fang

Department of Informatics, University of Sussex, Falmer, Brighton BN1 9RH, UK
{johnca, alexf}@sussex.ac.uk

Abstract. We describe the automatic acquisition of a lexicon of verb subcategorisations from a domain-specific corpus, and an evaluation of the impact this lexicon has on the performance of a "deep", HPSG parser of English. We conducted two experiments to determine whether the empirically extracted verb stems would enhance the lexical coverage of the grammar and to see whether the automatically extracted verb subcategorisations would result in enhanced parser coverage. In our experiments, the empirically extracted verbs enhance lexical coverage by 8.5%. The automatically extracted verb subcategorisations enhance the parse success rate by 15% in theoretical terms and by 4.5% in practice. This is a promising approach for improving the robustness of deep parsing.

1 Introduction

Typed unification-based grammatical frameworks such as head-driven phrase-structure grammars (HPSG; [1]) typically make use of two levels of lexical descriptions: a generic feature-based description for the word class in general and a set of lexical entries with specific features for each item. The grammar uses these lexical descriptions to assign precise syntactic and semantic analyses. Verb subcategorisation information, in particular, plays a vital role in the correct identification of complements of predicates. Consider:

> *I'm thinking of buying this software too but the trial version doesn't seem to have the option to set priorities on channels.*

The two prepositional phrases can only be attached correctly if the lexicon specifies that the verb *think* is complemented by the *of* phrase and that the complex transitive verb *set* can take a prepositional phrase headed by *on* as the object complement.

To date, verb subcategorisation lexicons have generally been constructed by hand. As with most handcrafted linguistic resources, they tend to excel in terms of precision of description but suffer from omissions and inconsistencies. A reason for this is that linguistically sophisticated descriptions of natural language are complex and expensive to construct and there is always some disparity between the computational grammarian's introspection and realistic input to the system. As an example, the LinGO English Resource Grammar (ERG; [2]) is a large HPSG grammar that contains around 2,000 lexical entries for over 1,100 verbs[1] selected to cover a number

[1] We use version 1.4 (April 2003) of the LinGO ERG.

of application-orientated corpora. When tested with 10,000 randomly selected sentences from a corpus of mobile phone-related discussions – a new domain for the ERG and containing many lexical items not present in standard dictionaries – it achieved a lexical coverage of only about 86% in terms of tokens and 42.3% in terms of types (see Section 4 below).

In this paper we describe research that aims to enhance the lexical coverage (and hence the parser success rate) of the LinGO ERG, using purely automatic techniques. We describe the construction and pre-processing of a domain-specific corpus, the extraction of verb subcategorisations from it using shallow processing, and experimental results on the impact of empirically extracted verb subcategorisations on the performance of the deep grammar/parser combination.

The approach we take is novel in that it combines unsupervised learning of language with manually constructed linguistic resources, working towards robust deep language processing.

2 Corpus Construction

We take as our domain and text type emails about models of mobile phones. In selecting a source for building a corpus of material of this type, practical considerations include:

- The source be unencumbered by copyright.
- The source be rich enough to allow for a sizable sample.
- The source be diversified enough to include a mixture of both spontaneous discussions and more formal prose (e.g. press releases).

As a result of these considerations, we decided to use the following mobile phone news groups, as archived by Google:

- alt.cell-phone.tech
- alt.cellular-phone-tech
- alt.cellular sub-groups:
  alltel, attws, bluetooth, cingular, clearnet, data, ericsson, fido, gsm, motorola, nextel, nokia, oki, rogersatt, sprintpcs, tech, telephones, umts, verizon

We downloaded 16,979 newsgroup postings, covering the period from 27 December 2002 to 4 April 2003.

2.1 Pre-processing

Each posting was automatically segmented into a header, quoted text, body text, and signature, based on the HTML markup and formatting clues. We made sure that only the body text was retained as part of the corpus, with the header, any quotation and signature automatically removed. Because of the planned evaluation of the subcategorisation lexicon, it was necessary to divide the corpus into training, development, and testing sections. From the corpus, 1,000 articles were set aside as the development set, 2,000 were held out for testing, and the rest (13,979 articles in all) were used for the training of the subcategorisation lexicon. Ten sets of 1,000 sentences were randomly selected from the testing section, to be used for evaluation purposes.

2.2 Language Processing

We next applied the RASP system [3] sentence boundary detector, tokeniser and part-of-speech (PoS) tagger to the training corpus. The 13,979 articles gave rise to 251,805 sentences. On this corpus we estimate sentence segmentation and tagging accuracy to both be around 95%, and the tagger's guessing of unknown words to be 80% accurate. The lexical component of the ERG is based around word stems rather than inflected word forms. We therefore applied the RASP morphological analyser [4] to reduce inflected verbs and nouns to their base forms.

We used the RASP parser to syntactically analyse the corpus. The parser uses a manually written "shallow" unification grammar of PoS tags and contains a statistical disambiguation model that selects the structurally most plausible syntactic analysis. The parser by default does not contain any word co-occurrence information, which makes it suitable for use in a system that acquires lexical information. The parser can recover from extra-grammaticality by returning partial analyses, which is essential for the processing of real-world data. The parser produces output that indicating the verb frame for each clause and the heads of the complements of the verb. Prepositional complements are also represented if they occur. After parsing, 165,852 of the 251,805 sentences in the training section of the corpus produced at least one verb pattern. This means that about 65.9% of the corpus was useful data for the extraction of verb subcategorisation patterns.

3 Lexicon Construction

This phase consists of two stages: extraction of subcategorisation frames, and then mapping them to the ERG scheme.

3.1 Extraction of Subcategorisation Frames

This stage involves the extraction of all the observed frames for any particular verb. For this purpose, we used the subcategorisation acquisition system of Briscoe and Carroll [5] as enhanced by Korhonen [6] and applied it to all the verbal pattern sets extracted from the training section of the parsed corpus. There are thus three subcategorisation representations: the RASP grammar subcategorisation values used in the parsed corpus, the Briscoe and Carroll (B&C) classes produced by the acquisition system, and the ERG lexical types for the target grammar.

From the training section of the corpus, a total of 16,371 such frames were extracted with 4,295 unique verb stems. On average, each verb has 3.8 different frames.

3.2 Mapping between RASP and ERG

The final stage in constructing the new ERG verb subcategorisation lexicon is the mapping of subcategorisation frames from the B&C scheme to the ERG scheme. The B&C scheme comprises a total of 163 possible subcategorisations, and the ERG scheme, 216. A B&C-to-ERG translation map was manually drawn up[2] and automatically applied to the acquired subcategorisation lexicon. 145 of the B&C subcategorisation frames map to 70 unique ERG lexical types, indicating a considerable degree of many-to-one matching. In the current mapping, for example, 9 different B&C frames are mapped to v_empty_prep_trans_le, the ERG type for verbs complemented by a prepositional phrase:

NP-FOR-NP	She bought a book for him.
NP-P-ING-OC	She accused him of being lazy.
NP-P-ING-SC	She wasted time on combing her hair.
NP-P-ING-AC	She told him about going to the park.
NP-P-NP-ING	She blamed it on no one buying it.
NP-P-POSSING	She asked him about his missing the train.
NP-P-WH-S	She asked whether she should go.
NP-P-WHAT-S	She asked what she should do.
NP-PP	She added salt to the food.

Conversely, 48 ERG lexical types do not have a B&C counterpart. Both schemes encode syntactic and semantic distinctions but it is evident that they do this in different ways.

The eventual lexicon contains 5,608 entries for 3,864 verb stems, an average of 1.45 entries per verb. See Figure 1 for entries acquired for the verb *accept* (where the lexical type v_np_trans_le represents a transitive entry and v_unerg_le an intransitive one).

```
Accept_rasp_v_np_trans_le := v_np_trans_le &
        [ STEM < “accept” > ] .
Accept_rasp_v_unerg_le := v_unerg_le &
        [ STEM < “accept” > ] .
```

Fig. 1. Entries Acquired for *accept*.

4 Lexical Coverage

We carried out a number of experiments to measure the impact of the subcategorisation lexicon on the performance of the HPSG parser. First of all, we wanted to determine whether the empirically extracted verb stems would enhance the lexical coverage of the grammar. Secondly, we wanted to see whether the empirically extracted verb entries would result in better parsing success rate, i.e., more sentences receiving an analysis by the parser. A third possibility is that the use of empirically extracted

[2] We are grateful to Dan Flickinger for providing us with this mapping.

verb subcategorisations, when applied to text of the same subject domain, would produce more accurate analyses. However, our experiments to date have been limited to the evaluation of the first two. In addition, we investigated the impact of the acquired lexicon on system performance, as measured by parsing time and space.

For the first experiment, we collected and unified all the verb stems from three machine-readable lexicons: 5,453 from the machine-readable Oxford Advanced Learner's Dictionary (OALD), 5,654 from ComLex [7], and 1,126 from the ERG lexicon. These three lexicons jointly yielded a total of 6,341 verbs. Comparing the 3,864 mobile phone verbs with this joint list, there are a number of items that are not represented which thus can be considered to be particular to the mobile phone domain. Manual inspection of the list indicates that there seem to be three groups of verbs. First of all, verbs like *recognise*, *realise*, and *patronise* may have crept into the list as a result of British English spelling not being represented in the three sources. Secondly, we observe some neologisms such as *txt* and *msg*, which are almost exclusively used within the mobile phone domain. Finally, we observe verbs that have been derived from free combinations of prefixes and verb stems such as *re-install* and *xtnd-connect*. These observations suggest the importance of the inclusion of variant spellings, and empirical corpus-based selection of lexical items.

To measure the lexical coverage of these verb stems, ten sets of 1,000 randomly selected sentences each were used as testing material. For comparison, all the verb stems (1,126 in all) were extracted from the manually coded ERG lexicon. The results show that the hand-selected verb stems have an average coverage of 85.9% for verb tokens and 42.3% for verb types in the 10 test sets. In comparison, the empirically selected list has a much higher coverage of 94.4% in terms of tokens and 62.9% in terms of types. The average token coverage of the lexicon that would be produced by taking the union of the verbs in OALD and Comlex is 91%. There is considerable variation in the coverage by ERG verb stems across the ten test sets (SD=1.1) while the empirically selected verbs have a much more consistent coverage with a standard deviation of only 0.34. The high level and consistency of coverage by the empirically selected verb stems demonstrate the advantages of a corpus-based approach to lexical selection.

5 Parse Success Rate

In order to achieve tangible results through comparison, we compared the original version of the ERG with an enhanced version that incorporates the empirically extracted verb subcategorisations. For these experiments we used the efficient HPSG parser PET [8], launched from [incr tsdb()], an integrated package for evaluating parser and grammar performance on test suites [9]. For all of the experiments described in the following sections, we set a resource limit on the parser, restricting it to produce at most 40,000 chart edges.

The evaluation aimed at two scenarios: theoretical enhancements from the empirically extracted verb subcategorisations and realistic improvements. The former was an attempt to ascertain an upper bound on parser coverage independent of practical limitations of parser timeouts as a result of the increase in lexical ambiguity in the grammar. The latter was performed in order to establish the real enhancements given practical constraints such as available parsing space and time. The following sections describe these two experiments.

5.1 Theoretical Enhancement

Deep grammars tend to require large amounts of memory when parsing to accommodate rich intermediate representations and their combination. High levels of lexical ambiguity often result in parser timeouts and thus could affect the correct estimate of the enhancements afforded by the extracted verb subcategorisations. Whether PET produces an analysis for a test item is decided by several factors. While the coverage of the grammar plays a central role, lexical complexities may cause the parser to run out of memory. We therefore attempted to factor out effects caused by increased lexical ambiguity in the acquired lexicon. For this purpose, a set of 1,133 sentences were specially selected from the test corpus with the condition that all the verbs were represented in the ERG grammar. We then manipulated this sub-corpus, replacing all of the nouns by *sense*, adjectives by *nice*, and adverbs by *nicely*. In doing so, it was guaranteed that failure to produce an analysis by the parser was not due to lexical combinatorial problems. Any observable improvements could be unambiguously attributed to the extracted verb subcategorisations.

The average parse success rate of the ERG for the test set is 48.5%, with sentences of fewer than 5 words receiving the highest percentage of analysis (60.8%). Lexical ambiguity is about 3.6 entries per word (33.71/9.32). In contrast, the enhanced ERG achieved a parse success rate of 63.5%, an increase of 15 basis points over that achieved by the original grammar. This time, 85.7% of the sentences of fewer than 5 words have at least one analysis, an almost 25% increase over the coverage of the same group by the original grammar. This improvement was achieved at a considerable increase in lexical ambiguity, 6.98 entries per word versus 3.6 for the original grammar. Parsing speed dropped from one second per sentence to about 4 seconds per sentence[3], and the average space requirement increased from 4.8 MB to 13.4 MB.

It should be noted that many ‘sentences’ in the original test set do not conform to standard English grammar; the version of PET that we used does not contain a robustness component, so such sentences failed to get a parse. In addition, lexical substitution was applied only to words tagged by RASP as either adverbs, adjectives, and nouns; there remain still a large number of ill-formed words (e.g. **that**, *dfgh*, *17c*) that eventually caused parse failures. Moreover, the words we used for lexical substitution are not grammatical in all contexts so in some cases substitution actually reduced parser coverage. Therefore, the absolute coverage figures should **not** be taken as definitive. However, since both setups used the same set of substituted sentences these failures do not affect our measurement of the difference in coverage, the subject of this experiment.

[3] There are 10 sentences which disproportionally increase the average. This appears to be due to errors in the subcategorisations for a few common words such as*ask*.

5.2 Realistic Enhancement

For the second experiment, we used a test set of 1,000 sentences randomly selected from the original test corpus. Unlike the previous experiment, the sentences were not lexically reduced. Since PET requires that every input token be known to the grammar, we supplemented the grammar with an openclass lexicon of nouns, adjectives and adverbs. These additional openclass items were extracted from the training section of the mobile phone corpus (as tagged by RASP) and consisted of 29,328 nouns, 7,492 adjectives, and 2,159 adverbs, totalling 38,797 stems. To cater for lexical items in the test set that are still not represented in the extended lexicon, the RASP tagging system was used as a PoS pre-processor for the test set. Each input sentence is thus annotated with PoS tags for all the tokens. For any lexical item unknown to the grammar, PET falls back to the PoS tag assigned automatically by RASP and applies an appropriate generic lexical description for the unknown word.

For this experiment, again, we have two grammars. The baseline statistics were obtained with the original handcrafted ERG grammar supplemented by the additional openclass items in the noun, adjective, and adverb classes. The enhanced grammar is additionally supplemented with the extracted verb subcategorisations. As indicated in Table 1[4], the original ERG grammar scored an average parse success rate of 52.9% with the space limit of 40,000 edges. Four test items had to be deleted from the test set since they consistently caused the parser to crash. The actual success rate is 52.7%.

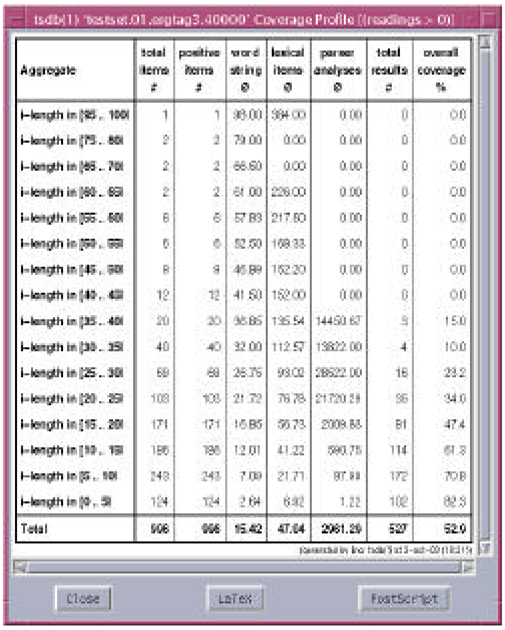
The same version of the ERG grammar was then integrated with the automatically extracted verb subcategorisations and the augmented version was applied to the same set of test sentences with the same system settings. As shown in Table 2, the enhanced grammar had 57.3% coverage, an increase of 4.4% over the original grammar without the automatically extracted verb subcategorisations.

We thus calculate the overall coverage enhancement as 4.4%. As in our previous experiment, we observed an expected increase in lexical ambiguity due to the increase of lexical items in the lexicon, up from 3.05 entries per word to 4.87. CPU time increased from 9.78 seconds per string for the original grammar to 21.78 for the enhanced grammar, with space requirements increasing from 45 MB to 55 MB.

[4] The colums in Tables 2 and 3 contain the following information:

• Aggregate	sentence length breakdown
• Total items	number of sentences
• Positive items	sentences seen by the parser
• Word string	average number of words
• Lexical items	average number of lexical entries
• Parser analyses	average number of analyses
• Total results	total number of successfully parsed sentences
• Overall coverage	percentage of successfully parsed sentences

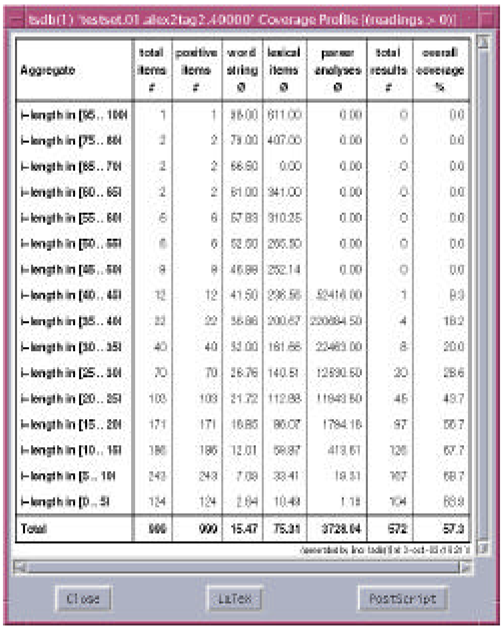
Table 1. Realistic Coverage of the ERG.


tsdb(1) 'testset.01.ergtag3.40000' Coverage Profile [(readings > 0)]
Close
LaTeX
PostScript

Aggregate	total items #	positive items #	word string Ø	lexical items Ø	parser analyses Ø	total results #	overall coverage %
i-length in [95 .. 100|	1	1	98.00	394.00	0.00	0	0.0
i-length in [75 .. 80|	2	2	79.00	0.00	0.00	0	0.0
i-length in [65 .. 70|	2	2	66.50	0.00	0.00	0	0.0
i-length in [60 .. 65|	2	2	61.00	226.00	0.00	0	0.0
i-length in [55 .. 60|	6	6	57.83	217.50	0.00	0	0.0
i-length in [50 .. 55|	6	6	52.50	169.33	0.00	0	0.0
i-length in [45 .. 50|	9	9	46.89	152.20	0.00	0	0.0
i-length in [40 .. 45|	12	12	41.50	152.00	0.00	0	0.0
i-length in [35 .. 40|	20	20	36.85	135.54	14450.67	3	15.0
i-length in [30 .. 35|	40	40	32.00	112.57	13822.00	4	10.0
i-length in [25 .. 30|	69	69	26.75	93.02	28622.00	16	23.2
i-length in [20 .. 25|	103	103	21.72	76.78	21720.21	35	34.0
i-length in [15 .. 20|	171	171	16.86	56.73	2009.88	81	47.4
i-length in [10 .. 15|	186	186	12.01	41.22	590.75	114	61.3
i-length in [5 .. 10|	243	243	7.09	21.71	87.94	172	70.8
i-length in [0 .. 5|	124	124	2.64	6.92	1.22	102	82.3
Total	996	996	15.42	47.04	2061.29	527	52.9

(generated by [incr tsdb()] at 3-oct-03 (18:21))

Table 2. Realistic Coverage of the Enhanced Grammar.


tsdb(1) 'testset.01.alex2tag2.40000' Coverage Profile [(readings > 0)]
Close
LaTeX
PostScript

Aggregate	total items #	positive items #	word string Ø	lexical items Ø	parser analyses Ø	total results #	overall coverage %
i-length in [95 .. 100|	1	1	98.00	611.00	0.00	0	0.0
i-length in [75 .. 80|	2	2	79.00	407.00	0.00	0	0.0
i-length in [65 .. 70|	2	2	66.50	0.00	0.00	0	0.0
i-length in [60 .. 65|	2	2	61.00	341.00	0.00	0	0.0
i-length in [55 .. 60|	6	6	57.83	310.25	0.00	0	0.0
i-length in [50 .. 55|	6	6	52.50	265.50	0.00	0	0.0
i-length in [45 .. 50|	9	9	46.89	252.14	0.00	0	0.0
i-length in [40 .. 45|	12	12	41.50	236.58	52416.00	1	8.3
i-length in [35 .. 40|	22	22	36.86	200.67	220684.50	4	18.2
i-length in [30 .. 35|	40	40	32.00	181.66	22463.00	8	20.0
i-length in [25 .. 30|	70	70	26.76	140.51	12930.50	20	28.6
i-length in [20 .. 25|	103	103	21.72	112.88	11943.50	45	43.7
i-length in [15 .. 20|	171	171	16.85	86.07	1794.16	97	56.7
i-length in [10 .. 15|	186	186	12.01	58.97	413.61	126	67.7
i-length in [5 .. 10|	243	243	7.08	33.41	19.31	167	68.7
i-length in [0 .. 5|	124	124	2.64	10.48	1.18	104	83.9
Total	999	999	15.47	75.31	3728.04	572	57.3

(generated by [incr tsdb()] at 3-oct-03 (18:21))

6 Conclusion

We presented a novel approach to improving the robustness of deep parsing, through the unsupervised acquisition of a verb subcategorisation lexicon. The acquired information helps deep parsing to produce detailed logical form representations that shallow analysers are unable to.

We described the construction of a corpus in our target mobile phone domain, and the acquisition of verb subcategorisations from the corpus through shallow processing with the RASP system. The empirically selected verb stems show enhanced lexical coverage of 94.4% against the 85.9% achieved by an existing handcrafted list of verb stems.

We then reported experiments to establish theoretical and practical gains from the use of extracted verb subcategorisations. When tested with 1,133 lexically reduced sentences, we observed that the enhanced grammar had 15% better coverage than the original grammar. Under practical constraints of parsing space and time, the enhanced grammar had an overall parse success rate of 57.3%, a 4.4% increase over the original grammar.

In our experiments, we used a completely automatic process to augment an existing hand-crafted lexicon. If manual effort is available, a good strategy would be for a linguist to check the acquired entries before they are added to the lexicon.

In future work we will implement a process that intelligently filters verb entries for words that already have entries in the lexicon; this should lead to a smaller lexicon, higher parse success rates and reduced parsing time and space. We will also refine acquired entries that require the explicit declaration of the heads of prepositional phrase complements. We also intend to investigate whether automatically extracted verb subcategorisations are capable of improving the accuracy of analyses proposed by the parser, perhaps taking advantage of the frequency information that is also collected in the acquisition process.

Acknowledgements

This research was supported by EU project Deep Thought IST-2001-37836. We would like to thank the Deep Thought project team, and in particular Ann Copestake, Dan Flickinger and Stephan Oepen for helpful suggestions and comments on earlier drafts. We are also grateful to Stephan Oepen for technical assistance, and to Anna Korhonen for providing us with her subcategorisation acquisition system.

References

1. Pollard, C. and I. Sag. 1994. *Head-Driven Phrase Structure Grammar.* Chicago University Press.
2. Copestake, A. and D. Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of LREC 2000*, Athens, Greece.
3. Briscoe, E. and J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation,* Las Palmas, Gran Canaria. 1499–1504.
4. Minnen, G., J. Carroll and D. Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering,* 7(3). 207-223.
5. Briscoe, E. and J. Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing,* Washington, DC. 356–363.
6. Korhonen, A. 2002. *Subcategorization Acquisition*. PhD thesis published as *Techical Report UCAM-CL-TR-530*. Computer Laboratory, University of Cambridge.
7. Grishman, R., C. Macleod and A. Meyers. 1994. Comlex syntax: Building a computational lexicon. In *Proceedings of the 15th International Conference on Computational Linguistics,* Kyoto, Japan. 268–272.
8. Callmeier, U. 2000. PET – A platform for experimentation with efficient HPSG processing techniques. *Natural Language Engineering*, 6(1) (Special Issue on Efficient Processing with HPSG):99–108.
9. Oepen, S. 1999. *[incr tsdb()]: Competence and Performance Laboratory: User & Reference Manual*, Computational Linguistics, Saarland University, Saarbrücken, Germany.