The Corpus and the Lexicon: Standardising Deep Lexical Acquisition Evaluation

Yi Zhang† and **Timothy Baldwin**‡ and **Valia Kordoni**†
† Dept of Computational Linguistics, Saarland University and DFKI GmbH, Germany
‡ Dept of Computer Science and Software Engineering, University of Melbourne, Australia
{yzhang,kordoni}@coli.uni-sb.de
tim@csse.unimelb.edu.au

Abstract

This paper is concerned with the standardisation of evaluation metrics for lexical acquisition over precision grammars, which are attuned to actual parser performance. Specifically, we investigate the impact that lexicons at varying levels of lexical item precision and recall have on the performance of pre-existing broad-coverage precision grammars in parsing, i.e., on their coverage and accuracy. The grammars used for the experiments reported here are the LinGO English Resource Grammar (ERG; Flickinger (2000)) and JACY (Siegel and Bender, 2002), precision grammars of English and Japanese, respectively. Our results show convincingly that traditional F-score-based evaluation of lexical acquisition does not correlate with actual parsing performance. What we argue for, therefore, is a recall-heavy interpretation of F-score in designing and optimising automated lexical acquisition algorithms.

1 Introduction

Deep processing is the process of applying rich linguistic resources within NLP tasks, to arrive at a detailed (=deep) syntactic and semantic analysis of the data. It is conventionally driven by deep grammars, which encode linguistically-motivated predictions of language behaviour, are usually capable of both parsing and generation, and generate a high-level semantic abstraction of the input data. While enjoying a resurgence of interest due to advances in parsing algorithms and stochastic parse pruning/ranking, deep grammars remain an underutilised resource predominantly because of their lack of coverage/robustness in parsing tasks. As noted in previous work (Baldwin et al., 2004), a significant cause of diminished coverage is the lack of lexical coverage.

Various attempts have been made to ameliorate the deficiencies of hand-crafted lexicons. More recently, there has been an explosion of interest in deep lexical acquisition (DLA; (Baldwin, 2005; Zhang and Kordoni, 2006; van de Cruys, 2006)) for broad-coverage deep grammars, either by exploiting the linguistic information encoded in the grammar itself (*in vivo*), or by using secondary language resources (*in vitro*). Such approaches provide (semi-)automatic ways of extending the lexicon with minimal (or no) human interference.

One stumbling block in DLA research has been the lack of standardisation in evaluation, with commonly-used evaluation metrics including:

- *Type precision*: the proportion of correctly hypothesised lexical entries
- *Type recall*: the proportion of gold-standard lexical entries that are correctly hypothesised
- *Type F-measure*: the harmonic mean of the type precision and type recall
- *Token Accuracy*: the accuracy of the lexical entries evaluated against their token occurrences in gold-standard corpus data

It is often the case that the different measures lead to significantly different assessments of the quality of DLA, even for a given DLA approach. Additionally, it is far from clear how the numbers generated by these evaluation metrics correlate with actual parsing performance when the output of a given DLA method is used. This makes standardised comparison among the various different approaches to DLA very difficult, if not impossible. It is far from clear which evaluation metrics are more indicative of the true “goodness” of the lexicon. The aim of this research, therefore, is to analyse how the different evaluation metrics correlate with actual parsing performance using a given lexicon, and to work towards

Proceedings of the 5th Workshop on Important Unresolved Matters, pages 152–159,
Ann Arbor, June 2005.

a standardised evaluation framework for future DLA research to ground itself in.

In this paper, we explore the utility of different evaluation metrics at predicting parse performance through a series of experiments over two broad coverage grammars: the English Resource Grammar (ERG; Flickinger (2000)) and JACY (Siegel and Bender, 2002). We simulate the results of DLA by generating lexicons at different levels of precision and recall, and test the impact of such lexicons on grammar coverage and accuracy related to gold-standard treebank data. The final outcome of this analysis is a proposed evaluation framework for future DLA research.

The remainder of the paper is organised as follows: Section 2 reviews previous work on DLA for the robust parsing task; Section 3 describes the experimental setup; Section 4 presents the experiment results; Section 5 analyses the experiment results; Section 6 concludes the paper.

2 Lexical Acquisition in Deep Parsing

Hand-crafted large-scale grammars are error-prone. An error can be roughly classified as *undergenerating* (if it prevents a grammatical sentence from being generated/parsed) or *overgenerating* (if it allows an ungrammatical sentence to be generated/parsed). Hence, errors in deep grammar lexicons can be classified into two categories: i) a lexical entry is missing for a specific lexeme; and ii) an erroneous lexical entry enters the lexicon. The former error type will cause the grammar to fail to parse/generate certain sentences (i.e. undergenerate), leading to a loss in coverage. The latter error type will allow the grammar to parse/generate inappropriate sentences (i.e. overgenerate), potentially leading to a loss in accuracy. In the first instance, we will be unable to parse sentences involving a given lexical item if it is missing from our lexicon, i.e. coverage will be affected assuming the lexical item of interest occurs in a given corpus. In the second instance, the impact is indeterminate, as certain lexical items may violate constraints in the grammar and never be licenced, whereas others may be licenced more liberally, generating competing (incorrect) parses for a given input and reducing parse accuracy. It is these two competing concerns that we seek to quantify in this research.

Traditionally, errors in the grammar are detected manually by the grammar developers. This is usually done by running the grammar over a carefully designed test suite and inspecting the outputs. This procedure becomes less reliable as the grammar gets larger. Also we can never expect to attain complete lexical coverage, due to language evolution and the effects of domain/genre. A static, manually compiled lexicon, therefore, becomes inevitably insufficient when faced with open domain text.

In recent years, some approaches have been developed to (semi-)automatically detect and/or repair the lexical errors in linguistic grammars. Such approaches can be broadly categorised as either symbolic or statistical.

Erbach (1990), Barg and Walther (1998) and Fouvry (2003) followed a unification-based symbolic approach to unknown word processing for constraint-based grammars. The basic idea is to use underspecified lexical entries, namely entries with fewer constraints, to parse whole sentences, and generate the "real" lexical entries afterwards by collecting information from the full parses. However, lexical entries generated in this way may be either too general or too specific. Underspecified lexical entries with fewer constraints allow more grammar rules to be applied while parsing, and fully-underspecified lexical entries are computationally intractable. The whole procedure gets even more complicated when two unknown words occur next to each other, potentially allowing almost any constituent to be constructed. The evaluation of these proposals has tended to be small-scale and somewhat brittle. No concrete results have been presented relating to the improvement in grammar performance, either for parsing or for generation.

Baldwin (2005) took a statistical approach to automated lexical acquisition for deep grammars. Focused on generalising the method of deriving DLA models on various secondary language resources, Baldwin used a large set of binary classifiers to predict whether a given unknown word is of a particular lexical type. This data-driven approach is grammar independent and can be scaled up for large grammars. Evaluation was via type precision, type recall, type F-measure and token accuracy, resulting in different interpretations of the data depending on the evaluation metric used.

Zhang and Kordoni (2006) tackled the robustness problem of deep processing from two aspects. They employed error mining techniques in order to semi-automatically detect errors in deep grammars. They then proposed a maximum entropy model based lexical type predictor, to generate new lexical entries on the fly. Evaluation focused on the accuracy of the lexical type predictor over unknown words, not the overall goodness of the resulting lexicon. Similarly to Baldwin (2005), the methods are applicable to other constraint-based lexicalist grammars, but no direct measurement of the impact on grammar performance was attempted.

van de Cruys (2006) took a similar approach over the Dutch Alpino grammar (cf. Bouma et al. (2001)). Specifically, he proposed a method for lexical acquisition as an extension to automatic parser error detection, based on large amounts of raw text (cf. van Noord (2004)). The method was evaluated using type precision, type recall and type F-measure. Once again, however, these numbers fail to give us any insight into the impact of lexical acquisition on parser performance.

Ideally, we hope the result of DLA to be both accurate and complete. However, in reality, there will always be a trade-off between coverage and parser accuracy. Exactly how these two concerns should be balanced up depends largely on what task the grammar is applied to (i.e. parsing or generation). In this paper, we focus exclusively on the parsing task.[1]

3 Experimental Setup

In this research, we wish to evaluate the impact of different lexicons on grammar performance. By grammar performance, we principally mean coverage and accuracy. However, it should be noted that the efficiency of the grammar—e.g. the average number of edges in the parse chart, the average time to parse a sentence and/or the average number of analyses per sentence—is also an important performance measurement which we expect the quality of the lexicon to impinge on. Here, however, we expect to be able to call on external processing optimisations[2] to dampen any loss in efficiency, in a way which we cannot with coverage and accuracy.

3.1 Resources

In order to get as representative a set of results as possible, we choose to run the experiment over two large-scale HPSGs (Pollard and Sag, 1994), based on two distinct languages.

The *LinGO English Resource Grammar* (ERG; Flickinger (2000)) is a broad-coverage, linguistically precise HPSG-based grammar of English, which represents the culmination of more than 10 person years of (largely) manual effort. We use the *jan-06* version of the grammar, which contains about 23K lexical entries and more than 800 leaf lexical types.

JACY (Siegel and Bender, 2002) is a broad-coverage linguistically precise HPSG-based grammar of Japanese. In our experiment, we use the November 2005 version of the grammar, which contains about 48K lexical entries and more than 300 leaf lexical types.

It should be noted in HPSGs, the grammar is made up of two basic components: the grammar rules/type hierarchy, and the lexicon (which interfaces with the type hierarchy via leaf lexical types). This is different to strictly lexicalised formalisms like LTAG and CCG, where essentially all linguistic description resides in individual lexical entries in the lexicon. The manually compiled grammars in our experiment are also intrinsically different to grammars automatically induced from treebanks (e.g. that used in the Charniak parser (Charniak, 2000) or the various CCG parsers (Hockenmaier, 2006)). These differences sharply differentiate our work from previous research on the interaction between lexical acquisition and parse performance.

Furthermore, to test the grammar precision and accuracy, we use two treebanks: Redwoods (Oepen et al., 2002) for English and Hinoki (Bond et al., 2004) for Japanese. These treebanks are so-called dynamic treebanks, meaning that they can be (semi-)automatically updated when the grammar is updated. This feature is especially useful when we want to evaluate the grammar performance with different lexicon configurations. With conventional treebanks, our experiment is difficult (if not impossible) to perform as the static trees in the treebank cannot be easily synchronised to the evolution of the grammar, meaning that we cannot regenerate gold-standard parse trees relative to a given lexicon (especially when for reduced recall where there is no guarantee we will be able to produce all of the parses in the 100% recall gold-standard). As a result, it is extremely difficult to faithfully update the statistical models.

The Redwoods treebank we use is the *6th growth*,

[1] In generation, we tend to have a semantic representation as input, which is linked to pre-existing lexical entries. Hence, lexical acquisition has no direct impact on generation.

[2] For example, (van Noord, 2006) shows that a HMM POS tagger trained on the parser outputs can greatly reduce the lexical ambiguity and enhance the parser efficiency, without an observable decrease in parsing accuracy.

which is synchronised with the *jan-06* version of the ERG. It contains about 41K test items in total.

The Hinoki treebank we use is updated for the November 2005 version of the JACY grammar. The *"Rei"* sections we use in our experiment contains 45K test items in total.

3.2 Lexicon Generation

To simulate the DLA results at various levels of precision and recall, a random lexicon generator is used. In order to generate a new lexicon with specific precision and recall, the generator randomly retains a portion of the gold-standard lexicon, and generates a pre-determined number of erroneous lexical entries.

More specifically, for each grammar we first extract a subset of the lexical entries from the lexicon, each of which has at least one occurrence in the treebank. This subset of lexical entries is considered to be the gold-standard lexicon (7,156 entries for the ERG, 27,308 entries for JACY).

Given the gold-standard lexicon L, the target precision P and recall R, a new lexicon L' is created, which is composed of two disjoint subsets: the retained part of the gold-standard lexicon G, and the erroneous entries E. According to the definitions of precision and recall:

$$P = \frac{|G|}{|L'|} \quad (1) \qquad R = \frac{|G|}{|L|} \quad (2)$$

and the fact that:

$$|L'| = |G| + |E| \tag{3}$$

we get:

$$|G| = |L| \cdot R \tag{4}$$

$$|E| = |L| \cdot R \cdot (\frac{1}{P} - 1) \tag{5}$$

To retain a specific number of entries from the gold-standard lexicon, we randomly select $|G|$ entries based on the combined probabilistic distribution of the corresponding lexeme and lexical types.[3] We obtain the probabilistic distribution of lexemes from large corpora (BNC for English and Mainichi Shimbun [1991-2000] for Japanese), and the distribution of lexical types from the corresponding treebanks. For each lexical entry $e(l,t)$ in the gold-standard lexicon with lexeme l and lexical type t, the combined probability is:

$$p(e(l,t)) = \frac{C_L(l) \cdot C_T(t)}{\sum_{e'(l',t') \in L} C_L(l') \cdot C_T(t')} \tag{6}$$

The erroneous entries are generated in the same way among all possible combinations of lexemes and lexical types. The difference is that only open category types and less frequent lexemes are used for generating new entries (e.g. we wouldn't expect to learn a new lexical item for the lexeme *the* or the lexical type `d_-_the_le` in English). In our experiment, we consider lexical types with more than a predefined number of lexical entries (20 for the ERG, 50 for JACY) in the gold-standard lexicon to be open-class lexical types; the upper-bound threshold on token frequency is set to 1000 for English and 537 for Japanese, i.e. lexemes which occur more frequently than this are excluded from lexical acquisition under the assumption that the grammar developers will have attained full coverage of lexical items for them.

For each grammar, we then generate 9 different lexicons at varying precision and recall levels, namely 60%, 80%, and 100%.

3.3 Parser Coverage

Coverage is an important grammar performance measurement, and indicates the proportion of inputs for which a correct parse was obtained (adjudged relative to the gold-standard parse data in the treebanks). In our experiment, we adopt a weak definition of coverage as "obtaining at least one spanning tree". The reason for this is that we want to obtain an estimate for novel data (for which we do not have gold-standard parse data) of the relative number of strings for which we can expect to be able to produce at least one spanning parse. This weak definition of coverage actually provides an upper bound estimate of coverage in the strict sense, and saves the effort to manually evaluate the correctness of the parses. Past evaluations (e.g. Baldwin et al. (2004)) have shown that the grammars we are dealing with are relatively precise. Based on this, we claim that our results for parse coverage provide a reasonable estimate indication of parse coverage in the strict sense of the word.

In principle, coverage will only decrease when the lexicon recall goes down, as adding erroneous entries should not invalidate the existing analyses. However, in practice, the introduction of erroneous entries increases lexical ambiguity dramati-

[3] For simplicity, we assume mutual independence of the lexemes and lexical types.

P \ R	0.6			0.8			1.0		
	C	E	A	C	E	A	C	E	A
0.6	4294	2862	7156	5725	3817	9542	7156	4771	11927
0.8	4294	1073	5367	5725	1431	7156	7156	1789	8945
1.0	4294	0	4294	5725	0	5725	7156	0	7156

Table 1: Different lexicon configurations for the ERG with the number of correct (C), erroneous (E) and combined (A) entries at each level of precision (P) and recall (R)

P \ R	0.6			0.8			1.0		
	C	E	A	C	E	A	C	E	A
0.6	16385	10923	27308	21846	14564	36410	27308	18205	45513
0.8	16385	4096	20481	21846	5462	27308	27308	6827	34135
1.0	16385	0	16385	21846	0	21846	27308	0	27308

Table 2: Different lexicon configurations for JACY with the number of correct (C), erroneous (E) and combined (A) entries at each level of precision (P) and recall (R)

cally, readily causing the parser to run out of memory. Moreover, some grammars use recursive unary rules which are triggered by specific lexical types. Here again, erroneous lexical entries can lead to "fail to parse" errors.

Given this, we run the coverage tests for the two grammars over the corresponding treebanks: Redwoods and Hinoki. The maximum number of passive edges is set to 10K for the parser. We used `[incr tsdb()]` (Oepen, 2001) to handle the different lexicon configurations and data sets, and `PET` (Callmeier, 2000) for parsing.

3.4 Parser Accuracy

Another important measurement of grammar performance is accuracy. Deep grammars often generate hundreds of analyses for an input, suggesting the need for some means of selecting the most probable analysis from among them. This is done with the parse disambiguation model proposed in Toutanova et al. (2002), with accuracy indicating the proportion of inputs for which we are able to accurately select the correct parse.

The disambiguation model is essentially a maximum entropy (ME) based ranking model. Given an input sentence s with possible analyses $t_1 \dots t_k$, the conditional probability for analysis t_i is given by:

$$P(t_i|s) = \frac{\exp \sum_{j=1}^{m} f_j(t_i)\lambda_j}{\sum_{i'=1}^{k} \exp \sum_{j=1}^{m} f_j(t_{i'})\lambda_j} \qquad (7)$$

where $f_1 \dots f_m$ are the features and $\lambda_1 \dots \lambda_m$ are the corresponding parameters. When ranking parses, $\sum_{j=1}^{m} f_j(t_i)\lambda_j$ is the indicator of "goodness". Drawing on the discriminative nature of the ME models, various feature types can be incorporated into the model. In combination with the dynamic treebanks where the analyses are (semi-)automatically disambiguated, the models can be easily re-trained when the grammar is modified.

For each lexicon configuration, after the coverage test, we do an automatic treebank update. During the automatic treebank update, only those new parse trees which are comparable to the active trees in the gold-standard treebank are marked as correct readings. All other trees are marked as inactive and deemed as overgeneration of the grammar. The ME-based parse disambiguation models are trained/evaluated using these updated treebanks with 5-fold cross validation. Since we are only interested in the difference between different lexicon configurations, we use the simple *PCFG-S* model from (Toutanova et al., 2002), which incorporates PCFG-style features from the derivation tree of the parse. The accuracy of the disambiguation model is calculated by top analysis exact matching (i.e. a ranking is only considered correct if the top ranked analysis matches the gold standard prefered reading in the treebank).

All the Hinoki *Rei* noun sections (about 25K items) were used in the accuracy evaluation for JACY. However, due to technical limitations, only the *jh* sections (about 6K items) of the Redwoods Treebank were used for training/testing the disambiguation models for the ERG.

4 Experiment Results

The experiment consumes a considerable amount of computational resources. For each lexicon config-

P \ R	0.6	0.8	1.0
0.6	44.56%	66.88%	75.51%
0.8	42.18%	65.82%	75.86%
1.0	40.45%	66.19%	76.15%

Table 3: Parser coverage of JACY with different lexicons

P \ R	0.6	0.8	1.0
0.6	27.86%	39.17%	79.66%
0.8	27.06%	37.42%	79.57%
1.0	26.34%	37.18%	79.33%

Table 4: Parser coverage of the ERG with different lexicons

uration of a given grammar, we need to i) process (parse) all the items in the treebank, ii) compare the resulting trees with the gold-standard trees and update the treebank, and iii) retrain the disambiguation models over 5 folds of cross validation. Given the two grammars with 9 configurations each, the entire experiment takes over 1 CPU month and about 120GB of disk space.

The coverage results are shown in Table 3 and Table 4 for JACY and the ERG, respectively.[4] As expected, we see a significant increase in grammar coverage when the lexicon recall goes up. This increase is more significant for the ERG than JACY, mainly because the JACY lexicon is about twice as large as the ERG lexicon; thus, the most frequent entries are still in the lexicons even with low recall.

When the lexicon recall is fixed, the grammar coverage does not change significantly at different levels of lexicon precision. Recall that we are not evaluating the correctness of such parses at this stage.

It is clear that the increase in lexicon recall boosts the grammar coverage, as we would expect. The precision of the lexicon does not have a large influence on coverage. This result confirms that with DLA (where we hope to enhance lexical coverage relative to a given corpus/domain), the coverage of the grammar can be enhanced significantly.

The accuracy results are obtained with 5-fold cross validation, as shown in Table 5 and Table 6

[4]Note that even with the lexicons at 100% precision and recall level, there is no guarantee of 100% coverage. As the contents of the Redwoods and Hinoki treebanks were determined independently of the respective grammars, rather than the grammars being induced from the treebanks e.g., they both still contain significant numbers of strings for which the grammar cannot produce a correct analysis.

P-R	#ptree	Avg.	σ
060-060	13269	62.65%	0.89%
060-080	19800	60.57%	0.83%
060-100	22361	59.61%	0.63%
080-060	14701	63.27%	0.62%
080-080	23184	60.97%	0.48%
080-100	27111	60.04%	0.56%
100-060	15696	63.91%	0.64%
100-080	26859	61.47%	0.68%
100-100	31870	60.48%	0.71%

Table 5: Accuracy of disambiguation models for JACY with different lexicons

P-R	#ptree	Avg.	σ
060-060	737	71.11%	3.55%
060-080	1093	63.94%	2.75%
060-100	3416	60.92%	1.23%
080-060	742	70.07%	1.50%
080-080	1282	61.81%	3.60%
080-100	3842	59.05%	1.30%
100-060	778	69.76%	4.62%
100-080	1440	60.59%	2.64%
100-100	4689	57.03%	1.36%

Table 6: Accuracy of disambiguation models for the ERG with different lexicons

for JACY and the ERG, respectively. When the lexicon recall goes up, we observe a small but steady decrease in the accuracy of the disambiguation models, for both JACY and ERG. This is generally a side effect of change in coverage: as the grammar coverage goes up, the parse trees become more diverse, and are hence harder to discriminate.

When the recall is fixed and the precision of the lexicon goes up, we observe a very small accuracy gain for JACY (around 0.5% for each 20% increase in precision). This shows that the grammar accuracy gain is limited as the precision of the lexicon increases, i.e. that the disambiguation model is remarkably robust to the effects of noise.

It should be noted that for the ERG we failed to observe any accuracy gain at all with a more precise lexicon. This is partly due to the limited size of the updated treebanks. For the lexicon configuration $060-060$, we obtained only 737 preferred readings/trees to train/test the disambiguation model over. The 5-fold cross validation results vary within a margin of 10%, which means that the models are still not converging. However, the result does confirm that there is no significant gain in grammar accuracy with a higher precision lexicon.

Finally, we combine the coverage and accuracy scores into a single F-measure ($\beta = 1$) value. The results are shown in Figure 1. Again we see that the difference in lexicon recall has a more significant impact on the overall grammar performance than precision.

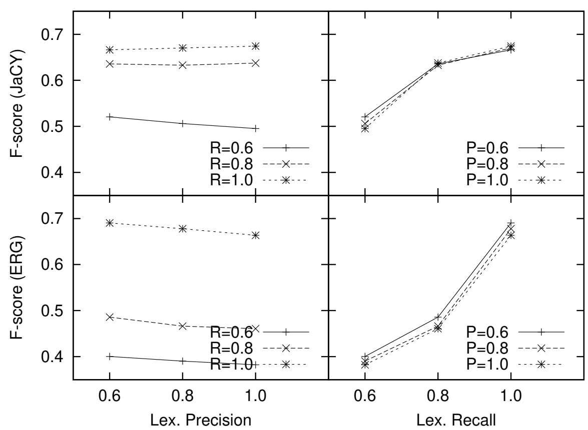

Figure 1: Grammar performance (F-score) with different lexicons

5 Discussion

5.1 Is F-measure a good metric for DLA evaluation?

As mentioned in Section 2, a number of relevant earlier works have evaluated DLA results via the unweighted F-score (relative to type precision and recall). This implicitly assumes that the precision and recall of the lexicon are equally important. However, this is clearly not the case as we can see in the results of the grammar performance. For example, the lexicon configurations $060 - 100$ and $100 - 060$ of JACY (i.e. 60% precision, 100% recall vs. 100% precision, 60% recall, respectively) have the same unweighted F-scores, but their corresponding overall grammar performance (parser F-score) differs by up to 17%.

5.2 Does precision matter?

The most interesting finding in our experiment is that the precision of the deep lexicon does not appear to have a significant impact on grammar accuracy. This is contrary to the earlier predominant belief that deep lexicons should be as accurate as possible. This belief is derived mainly from observation of grammars with relatively small lexicons. In such small lexicons, the closed-class lexical entries and frequent entries (which comprise the "core" of the lexicon) make up a large proportion of lexical entries. Hence, any loss in precision means a significant degradation of the "core" lexicon, which leads to performance loss of the grammar. For example, we find that the inclusion of one or two erroneous entries for frequent closed-class lexical type words (such as *the*, or *of* in English, for instance) may easily "break" the parser.

However, in state-of-the-art broad-coverage deep grammars such as JACY and ERG, the lexicons are much larger. They usually have more or less similar "cores" to the smaller lexicons, but with many more open-class lexical entries and less frequent entries, which compose the "peripheral" parts of the lexicons. In our experiment, we found that more than 95% of the lexical entries belong to the top 5% of the open-class lexical types. The bigger the lexicon is, the larger the proportion of lexical entries that belong to the "peripheral" lexicon.

In our experiment, we only change the "peripheral" lexicon by creating/removing lexical entries for less frequent lexemes and open-class lexical types, leaving the "core" lexicon intact. Therefore, a more accurate interpretation of the experimental results is that the precision of the *open type* and *less frequent* lexical entries does not have a large impact on the grammar performance, but their recall has a crucial effect on grammar coverage.

The consequence of this finding is that the balance between precision and recall in the deep lexicon should be decided by their impact on the task to which the grammar is applied. In research on automated DLA, the motivation is to enhance the robustness/coverage of the grammars. This work shows that grammar performance is very robust over the inevitable errors introduced by the DLA, and that more emphasis should be placed on recall.

Again, caution should be exercised here. We do *not* mean that by blindly adding lexical entries without worrying about their correctness, the performance of the grammar will be monotonically enhanced – there will almost certainly be a point at which noise in the lexicon swamps the parse chart and/or leads to unacceptable levels of spurious ambiguity. Also, the balance between precision and recall of the lexicon will depend on various expectations of the grammarians/lexicographers, i.e. the linguistic precision and generality, which is beyond the scope of this paper.

As a final word of warning, the absolute grammar performance change that a given level of lexicon type precision and recall brings about will obviously depend on the grammar. In looking across two grammars from two very different languages, we are confident of the robustness of our results (at least for grammars of the same ilk) and the conclusions that we have drawn from them. For any novel grammar and/or formalism, however, the performance change should ideally be quantified through a set of experiments with different lexicon configurations, based on the procedure outlined here. Based on this, it should be possible to find the optimal balance between the different lexicon evaluation metrics.

6 Conclusion

In this paper, we have investigated the relationship between evaluation metrics for deep lexical acquisition and grammar performance in parsing tasks. The results show that traditional DLA evaluation based on F-measure is not reflective of grammar performance. The precision of the lexicon appears to have minimal impact on grammar accuracy, and therefore recall should be emphasised more greatly in the design of deep lexical acquisition techniques.

References

Timothy Baldwin, Emily Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2004. Road-testing the English Resource Grammar over the British National Corpus. In *Proc. of the fourth international conference on language resources and evaluation (LREC 2004)*, pages 2047–2050, Lisbon, Portugal.

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

Petra Barg and Markus Walther. 1998. Processing unknown words in HPSG. In *Proc. of the 36th Conference of the ACL and the 17th International Conference on Computational Linguistics*, pages 91–95, Montreal, Canada.

Francis Bond, Sanae Fujita, Chikara Hashimoto, Kaname Kasahara, Shigeko Nariyama, Eric Nichols, Akira Ohtani, Takaaki Tanaka, and Shigeaki Amano. 2004. The Hinoki treebank: a treebank for text understanding. In *Proc. of the first international joint conference on natural language processing (IJCNLP04)*, pages 554–562, Hainan, China.

Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. Alpino: wide-coverage computational analysis of Dutch. In *Computational linguistics in the Netherlands 2000*, pages 45–59, Tilburg, the Netherlands.

Ulrich Callmeier. 2000. PET – a platform for experimentation with efficient HPSG processing techniques. *Natural Language Engineering*, 6(1):99–107.

Eugene Charniak. 2000. A maximum entropy-based parser. In *Proc. of the 1st Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL2000)*, Seattle, USA.

Gregor Erbach. 1990. Syntactic processing of unknown words. IWBS Report 131, IBM, Stuttgart, Germany.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28.

Frederik Fouvry. 2003. Lexicon acquisition with a large-coverage unification-based grammar. In *Proc. of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pages 87–90, Budapest, Hungary.

Julia Hockenmaier. 2006. Creating a CCGbank and a wide-coverage CCG lexicon for German. In *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 505–512, Sydney, Australia.

Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger, and Thorsten Brants. 2002. The LinGO Redwoods treebank: Motivation and preliminary applications. In *Proc. of the 17th international conference on computational linguistics (COLING 2002)*, Taipei, Taiwan.

Stephan Oepen. 2001. [incr tsdb()] — competence and performance laboratory. User manual. Technical report, Computational Linguistics, Saarland University, Saarbrücken, Germany.

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

Melanie Siegel and Emily Bender. 2002. Efficient deep processing of Japanese. In *Proc. of the 3rd Workshop on Asian Language Resources and International Standardization*, Taipei, Taiwan.

Kristina Toutanova, Christoper Manning, Stuart Shieber, Dan Flickinger, and Stephan Oepen. 2002. Parse ranking for a rich HPSG grammar. In *Proc. of the First Workshop on Treebanks and Linguistic Theories (TLT2002)*, pages 253–263, Sozopol, Bulgaria.

Tim van de Cruys. 2006. Automatically extending the lexicon for parsing. In *Proc. of the eleventh ESSLLI student session*, pages 180–191, Malaga, Spain.

Gertjan van Noord. 2004. Error mining for wide-coverage grammar engineering. In *Proc. of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 446–453, Barcelona, Spain.

Gertjan van Noord. 2006. At Last Parsing Is Now Operational. In *Actes de la 13e conference sur le traitement automatique des langues naturelles (TALN06)*, pages 20–42, Leuven, Belgium.

Fei Xia, Chung-Hye Han, Martha Palmer, and Aravind Joshi. 2001. Automatically extracting and comparing lexicalized grammars for different languages. In *Proc. of the 17th International Joint Conference on Artificial Intelligence (IJCAI-2001)*, pages 1321–1330, Seattle, USA.

Yi Zhang and Valia Kordoni. 2006. Automated deep lexical acquisition for robust open texts processing. In *Proc. of the fifth international conference on language resources and evaluation (LREC 2006)*, pages 275–280, Genoa, Italy.