Road-testing the English Resource Grammar over the British National Corpus

Timothy Baldwin[*]**, Emily M. Bender**[†]**, Dan Flickinger**[*]**,**
Ara Kim[*]**, and Stephan Oepen**[‡*]

[*]Center for the Study of Language and Information
Stanford University
Stanford, CA 94305 (USA)

[†]Department of Linguistics
University of Washington
Seattle WA 98195 (USA)

[‡]Institutt for Lingvistiske Fag
Universitetet i Oslo
0317 Oslo (Norway)

{tbaldwin|danf|ara23|oe}@csli.stanford.edu ebender@u.washington.edu

Abstract

This paper addresses two questions: (1) when a large deep processing resource developed for relatively closed domains is run over open text, what coverage does it have, and (2) what are the most effective and time-efficient ways of consolidating gaps in the coverage of such as resource?

1. Introduction

Deep processing has entered the mainstream of applied NLP research in recent years due to the cumulative effects of Moore's Law, advances in algorithm efficiency, and the maturation of deep processing resources (Uszkoreit, 2002; Oepen et al., 2002a). Here, we define *deep processing* to be an umbrella term for methods which are based on full grammatical analysis, and generally grounded in semantics; *shallow processing*, on the other hand, refers to methods which make use of a diminished level of linguistic precision.

While exponential advances have been made in processing time for deep processing systems, language resource (LR) development has tended to take a more linear path. This is perhaps inevitable as deep processing LRs tend to be the source of the precision, making their development labor-intensive. One way in which the issue of restricted coverage has been defused is to focus on limited domains or make closed world assumptions about vocabulary and syntax. However, if we are to aim for broad-coverage deep processing without any domain assumptions, alternate strategies are clearly necessary.

This paper addresses two fundamental questions: (1) when a large deep processing resource developed for relatively closed domains is run over open text, what coverage does it have, and (2) what are the most effective and time-efficient ways of consolidating gaps in the coverage of such LRs? In attempting to answer these questions, we take the English Resource Grammar (ERG: Copestake and Flickinger (2000), Flickinger (2002)) as our deep processing LR, and carry out a detailed evaluation of its coverage over the written component of the British National Corpus (BNC: Burnard (2000)). We then apply these results in postulating approaches for narrowing the coverage gap over the BNC.

In the following sections, we first give a detailed description of the ERG and BNC, and the methodology for evaluating coverage (§2.). Next, we classify different causes of gaps in coverage over the BNC and present a breakdown of the frequency of each (§3.). Based on these results, we postulate an approach for achieving rapid coverage expansion (§4.).

2. Resources and preprocessing

The ERG is an open-source[1] broad-coverage precision HPSG grammar developed for parsing and generation. It has been engineered primarily based on corpus data in informal genres such as conversations about scheduling and email regarding e-commerce. While these domains are relatively open-ended, their task orientation leads to significant bias in their lexical and constructional composition. Also, both are informal genres based on either transcribed speech or informal text, raising questions about the portability of the ERG to more formal corpora such as the BNC.

The ERG contains roughly 10,500 lexical items, which, when combined with 59 lexical rules, compile out to around 12,500 distinct word forms.[2] Each lexical item consists of a unique identifier, a lexical type (one of some 600 leaf types organized into a type hierarchy with a total of around 4,000 types), an orthography and a semantic relation. The grammar also contains 77 phrase structure rules which serve to combine words and phrases into larger constituents, and compositionally relate such structures to semantic representations in the Minimal Recursion Semantics framework (MRS: Copestake et al. (2003)). Of the 10,500 lexical items, roughly 3,000 are multiword expressions (MWEs).

We test the coverage of the ERG over a random sample of 20,000 strings from the written component of the BNC (based on the BNC sentence tokenization). At present, unknown word handling in the ERG is restricted to number expressions and proper names. An input containing any word which does not fall into these classes or is not explicitly described as a lexical item therefore leads to parse failure. In order to filter out the effects of unknown words and focus on constructional coverage and the syntactic coverage of known words, we restricted our attention to strings for which we have a full lexical span, i.e. which contain only words already licensed by the grammar (including lexical rules). In order to apply this filter to the data, we first POS-tagged the strings and stripped away any punctuation not handled by the grammar (e.g. commas and periods).

[1]The ERG can be downloaded from `http://lingo.stanford.edu/erg.html`.

[2]All statistics and analysis relating to the ERG in this paper are based on the version of 1 July, 2003.

Based on the tagger output, we tokenized proper names and number expressions (both cardinal and ordinal), and finally used a table of British–American spelling variants to translate any British spellings into their American equivalents.[3] After tokenization and spelling normalization, the proportion of strings for which the ERG had full lexical span was 32%. This analysis was done by building a lattice of simplex words and multiword expressions licensed by the grammar, and looking for the existence of a spanning path through the lattice.

3. Coverage analysis

Of the strings with full lexical span, the grammar was able to generate a parse for 57%. Of these, 83% were found to have a correct parse. We diagnosed the cause(s) of parse failure in the unparsed sentences (43% of the total) by postulating possible causes, validating them by proposing minimal paraphrases and testing whether the grammar could assign them a correct parse. We then classified the cause(s) of parse failure into six categories, briefly described below: (a) missing lexical entries (40%), (b) missing constructions (39%), (c) preprocessor errors (4%), (d) fragments (4%), (e) parser failures (4%), and (f) garbage strings (11%).

1. There were two varieties of lexical gaps: incomplete categorization of existing lexical items (32% of total errors) and missing multiword expressions (MWEs, 8% of total errors). Missing MWEs cause parse failure if the MWE is syntactically marked—such as verb-particle constructions (e.g. *take off*) and determiner-less PPs (e.g. *off screen*)—such that the ordinary grammatical rules and lexical entries can't generate the string. In incomplete categorization the lexical entries for an orthographic form do not instantiate the full range of lexical types necessary, e.g. the noun *table*, but not the verb. In some cases, we observed that the gaps reflected general processes (e.g. the universal grinder converting countable nouns to uncountable nouns (Pelletier, 1979)) which could be handled by lexical rules. Mostly, however, the effect was simply one of patchy coverage.

2. Parse errors due to missing constructions were either known gaps in the constructional coverage of the ERG (e.g. vocative uses of NPs: 23% of total errors), instances of overly restrictive constraints on implemented construction types (e.g. particular types of coordination: 16% of total errors), or previously unconsidered constructions (e.g. *He's a hell of a nice guy.*).

3. Preprocessor errors involved common nouns or other elements (e.g. *Whilst*) being mistagged as proper nouns or vice versa, causing errors in tokenization, leading in turn to unparseable inputs. Also, a small number of remaining British spellings caused parse failure in some cases.

4. While the ERG handles some fragments (e.g. *next Wednesday*) by allowing specific kinds of phrases other than complete sentences as stand-alone utterances, certain fixed-phrase fragments such as *Small world* fell outside the kinds of phrases thus treated so far.

5. Parser failures occurred when the parser ran out of chart edges before creating any spanning parses which satisfied the root conditions. This occurred particularly for strings with a high level of coordination, modifier or attachment ambiguity. We believe this problem can be resolved by training the parse search algorithm on a treebank constructed from successfully parsed examples Oepen et al. (2002b). With such training, the parser should be able to find spanning edges even for very long and ambiguous sentences, whereas in the experiments here it was always attempting to parse exhaustively within the limits given.

6. Finally, garbage strings were either ungrammatical (either as full sentences or as fragments) or were simply strings of names and numbers (e.g. addresses), better treated via a more extended preprocessing mechanism than within the grammar itself.

Based on the distribution of the above effects, the two obvious places to extend the coverage of the ERG are lexical and constructional coverage. The easier by far is lexical coverage as we can expect to largely map new words onto existing lexical types. Extending constructional coverage, on the other hand, requires designing new analyses and implementing them in such a way that they are compatible with the rest of the grammar. Additionally, even if we were to implement analyses of the missing constructions, we can only expect an increment in coverage of roughly 39% of the 43% of strings which were unparseable. As we only attempted to parse the strings with full lexical span, this equates to only 12% of the total BNC sample. With missing lexical items, on the other hand, we can expect to make inroads into both 40% of the subset of unparseable strings with lexical span and also the remaining 68% of strings *without* lexical span. Lexical expansion is thus the first step in the proposed way forward. Figure 1 depicts coverage and parser performance measures over the relevant sub-corpus.

4. Expanding the Coverage of the ERG

In order to get a better sense of how to approach the issue of lexical expansion, we carried out more detailed analysis of lexical gaps, focusing on nouns, verbs and adjectives occurring as unknown words or lexemes with incomplete type coverage in the ERG lexicon. For the strings *without* full lexical span (68% of our original 20,000 strings) we used the tagger output to analyze the distribution of the parts-of-speech of each unknown word token. We found that 61% of unknown words are nouns, 22% are adjectives 13% are verbs. Additionally, an analysis of the strings with full lexical span showed that, of the lexemes with incomplete type coverage, 56% were nouns with only verbal or adjectival entries, 11% adjectives with only nominal or verbal entries, and 10% verbs with only nominal or adjectival entries. Another 15% were words with the right parts of speech, but incomplete sets of subcategorization frames.

[3] The ERG is based on American English, whereas the BNC is, unsurprisingly, a sample of British English.

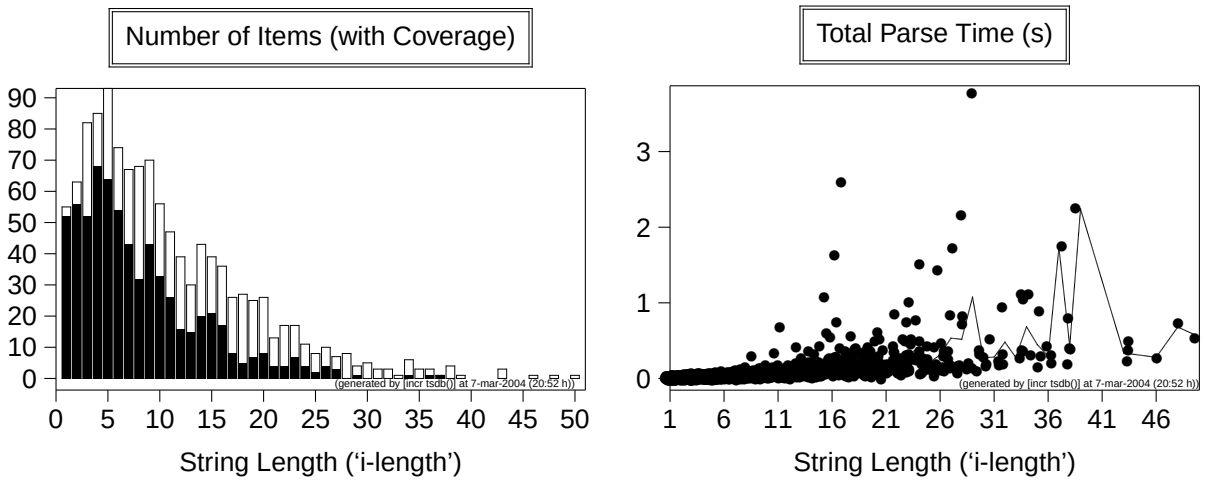

Figure 1: Summaries of grammatical coverage (left) over the relevant BNC sub-set and corresponding parser performance (right). The coverage bar chart gives a graphical impression of the proportion of items in each aggregate that received one or more analyses (the solid bar) contrasted with the total number of inputs of the specific length. As should be expected, analysis coverage deteriorates in input length, as the likelihood of hitting a lexical or constructional gap increases. Although in the present study we focus on the experimental assessment of grammatical coverage and error types, it is reassuring to observe that parse times—using the high-efficiency PET parser (Callmeier, 2002)—for the vast majority of sentences are well below one second.

Turning next to the structure of the existing ERG lexicon, we observe that the number of lexical leaf types for verbs is 131, more than double the number for adjectives or nouns. Additionally, the distribution of verbal lexical types across lexemes is relatively flat, with the top-10 lexical types accounting for only around 77% of verbal lexical items; for nouns, this figure is 96%, for adjectives 90%. In this sense, the task of the lexicographer or a (semi-)automated lexical acquisition technique is much simpler for nouns than either adjectives or verbs. Thus we conclude that the most efficient way forward in terms of incrementally expanding the coverage of the ERG is to focus on lexical acquisition, beginning with nouns.

However—even in the relatively vanilla-looking universe of nouns—a precision grammar like the ERG requires a range of lexical distinctions that exceed the information traditionally available from part of speech taggers and computational dictionaries. The ERG classifies its lexical types for nouns along several potentially cross-cutting parameters, including: countability (1), nominal argument structure (2), proper vs. common nouns (3), the ability to project into PP-like modifiers (4), and various idiosyncrasies (e.g. pluralia tantum (5)).

(1) a. an/*ϕ/*much aperitif

b. much/ϕ/*a leisure

c. much/ϕ/a wine

(2) a. a bike (*of a vacation/*that Kim slept/...)

b. an accumulation of debt/*that Kim slept/*in photography

c. an interest in photography/*of photography/*that Kim slept

d. the belief that Kim slept/*of photography/in photography

(3) a. I always prefer (*the) Amtrak.

b. I always prefer *(the) train.

(4) a. Kim arrived the first *day* of March.

b. Kim arrived *Tuesday*.

c.*Kim arrived *afternoon*.

(5) Kim's got the goods/*good

Although Baldwin and Bond (2003) offer hope of acquiring some of the relevant information automatically, in order to avoid extensive overgeneration (which reduces the utility of the grammar in both parsing and generation), hand-inspection of candidate lexical entries will be necessary. We believe that even this hand-inspection can be expedited by using the ERG to generate (somewhat generic) test sentences illustrating the range of potential contexts each lexical type predicts. An annotator would then only need to glance over a set of sample sentences for each word, and confirm or correct the indicated grammaticality judgments. If the corrected pattern fit one predicted by a particular set of lexical entries for the word, such entries could be automatically generated. If it did not, the lexical item would be flagged for future reference as perhaps motivating a new lexical type. We predict, however, that the vast majority of unknown words will be accommodated by our existing lexical types.

5. Conclusion

At first sight, the absolute coverage figures reported for parsing the BNC with the LinGO ERG, an HPSG implementation that has been under continuous development at CSLI since 1993, must seem disappointingly low. At the same time, we felt reasonably content with the outcome of this first, out-of-the box experiment: obtaining close to 60% grammatical coverage from applying to the BNC a hand-built precision grammar that was originally developed for informal, unedited English in limited domains (and lacks both a large, general-purpose lexicon, refined treatment of unknown words, and any kind of robustness facilities) seemed like a respectable outcome. Furthermore, the 83% correctness measure that we found in treebanking the analyses produced by the grammar appears to confirm the semantically precise nature of the grammar; as does an average ambiguity of 64 analyses per sentence for strings of length 10 to 20 words. To put these results into perspective, typical coverage figures for the ERG on new data from the closed (spoken) appointment scheduling and (email) e-commerce domains tend to range upwards of 80%, with ambiguity rates of around 100 analyses on average per input. A recent experiment in manually adding vocabulary from a word list for a 300-item excerpt from tourism brochures gave the ERG an initial coverage of above 90% (at an average ambiguity of 187 analyses for an average string length of 13 words). In all three scenarios treebanking confirms parse correctness measures of at least 90%.

Clearly, our results confirm that additional effort would be required to develop the ERG into a domain-independent, wide-coverage LR for collections of (mostly) edited text like the BNC (or the Penn Treebank). Although parsing open-domain edited text is, actually, not among our immediate goals (see below), the coverage assessments reported here seem to suggest that there is quite a bit of low-hanging fruit in porting a hand-built grammar to a new task. Lexical gaps account for by far the largest fraction of parse failures found in our experiments, and we envision a semi-automated lexical acquisition set-up where candidate lexical entries are generated from various sources and manually reviewed prior to inclusion in the ERG lexicon. The ERG, including its integrated semantics, and an associated generation component, could be leveraged to reduce the need for trained lexicographers. This technology could be used, as described above, to automatically produce schematic 'test' sentences using candidate words—for count vs. mass nouns, for example—that could be confidently judged by non-linguists.

However, to enable embedding of a semantic precision grammar in applications requiring one to parse arbitrary running text, we actively pursue hybrid NLP approaches, combining both shallow and deeps LRs, and using stochastic methods to cope with ambiguity. The LinGO ERG, as reviewed here, is one of several building blocks in this set-up.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. BCS-0094638 and also the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, and CSLI at Stanford University. We would like to thank John Beavers, Ivan Sag, Tom Wasow and the three anonymous reviewers for their valuable input on this research.

6. References

Timothy Baldwin and Francis Bond. 2003. Learning the countability of English nouns from corpus data. In *Proc. of the 41st Annual Meeting of the ACL*, pages 463–70, Sapporo, Japan.

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

Ulrich Callmeier. 2002. PET – a platform for experimentation with efficient HPSG processing techniques. In (Oepen et al., 2002a).

Ann Copestake and Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage english grammar using HPSG. In *Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece.

Ann Copestake, Dan Flickinger, Ivan A. Sag, and Carl Pollard. 2003. Minimal recursion semantics: An introduction. Under review.

Dan Flickinger. 2002. On building a more efficient grammar by exploiting types. In (Oepen et al., 2002a).

Stephan Oepen, Dan Flickinger, Jun'ichi Tsujii, and Hans Uszkoreit, editors. 2002a. *Collaborative Language Engineering*. CSLI Publications, Stanford, USA.

Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger, and Thorsten Brants. 2002b. The LinGO Redwoods Treebank: Motivation and preliminary applications. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 1253–7, Taipei, Taiwan.

F. Jeffrey Pelletier. 1979. Non-singular reference: Some preliminaries. In F. Jeffrey Pelletier, editor, *Mass Terms: Some Philosophical Problems*, pages 1–14. Dordrecht, Reidel.

Hans Uszkoreit. 2002. New chances for deep linguistic processing. In *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan.