Beauty and the Beast: What running a broad-coverage precision grammar over the BNC taught us about the grammar — and the corpus

Timothy Baldwin, John Beavers, Emily M. Bender*,
Dan Flickinger, Ara Kim and Stephan Oepen
CSLI, Stanford University
210 Panama St
Stanford CA 94305-4115 USA
*Dept of Linguistics
University of Washington
Seattle WA 98195-4340 USA
{tbaldwin,jbeavers,bender,danf,ara23,oe}@csli.stanford.edu

Typically, broad-coverage precision grammars are based on grammaticality judgment data and syntactic intuition, and corpus data is relegated to secondary status in guiding lexicon and grammar development. On the other end of the scale, shallow grammars are often induced directly from treebank data and make little or no use of grammaticality judgments or intuition. This tends to cause precision grammars to undergenerate and shallow grammars to massively overgenerate. With broad-coverage precision grammars, the issue of undergeneration is solved incrementally by extending coverage. When faced with an as-yet unanalyzed sentence (from a corpus or otherwise), grammar engineers first consult the literature, their intuition, or their office mates in order to map out a space of both grammatical and ungrammatical examples, which serves as the basis for the analysis coded into the grammar. Shallow grammars, on the other hand, tend not to deal with grammaticality and focus instead on selecting the most plausible of the available parses, generally through stochastic means.

In this paper, we take the English Resource Grammar (ERG: Copestake and Flickinger, 2000), an implemented broad-coverage precision Head-driven Phrase Structure Grammar (HPSG; Pollard and Sag (1994)) developed mostly against smaller corpora of informal text (spoken language transcriptions and email) along with manually-generated linguistic examples, and run it over a large-scale corpus of more formal text, namely the written portion of the British National Corpus (BNC: Burnard, 2000). The ERG has been developed for both parsing and generation, and thus places a premium on precision: extraneous parses not only complicate the process of ambiguity resolution, but can lead to ill-formed output from the generator. We discuss how the BNC can be used to constructively road-test and ultimately extend the coverage of the ERG. We also examine limitations in fully corpus-driven grammar development, and motivate

the continued use of judgment data throughout the evolution of a precision grammar.

In order to filter out the effects of lexical coverage and focus on grammatical coverage, we restrict our analysis to sentences for which we have a full lexical span, i.e. which contain only words already licensed by the grammar (including lexical rules)—32% of a random sample of 20,000 BNC strings. The strings were also stripped of most punctuation, and normalized by tokenizing all number expressions and proper names (as identified by a POS tagger) and substituting American spellings.

The grammar was able to parse 57% of these strings. The parses were manually inspected (using a parse selection tool (Oepen et al., 2002a)), and 83% were found to have a correct parse. We then used the grammar to analyze the 43% of sentences with no correct parses by proposing paraphrases until the grammar was able to parse the string. This method allowed us to incrementally diagnose the causes of parse failure by testing linguistic hypotheses rather than resorting solely to implementation details of the ERG. The sources of the parse failures can be classified into four types which we examine in turn: (a) missing lexical entries, (b) missing constructions, (c) ungrammatical strings, and (d) extragrammatical strings.

Despite the restriction to strings with a full lexical span, we were nonetheless confronted by gaps in lexical coverage, which fall into two basic categories: incomplete categorization of existing lexical items and missing multiword expressions (MWEs). Each ERG lexical item is annotated with a specific lexical type which determines its syntactic and semantic behavior. Incomplete categorization of the full range of lexical types for a given word (e.g. the noun *table*, but not the verb) leads to parse failure. Syntactically-marked MWEs—notably verb-particle constructions (e.g. *take off*) and determinerless PPs (e.g. *off screen*)—cause similar problems: The demands of precision grammar engineering dictate that the grammar explicitly license each observed verb-particle pair or determinerless PP rather than letting any particle appear with any verb or any count noun appear immediately after a preposition. In some cases (e.g. mass uses of prototypical count nouns, action verbs+completive particle *up*), it appears that a general process (i.e. the 'universal grinder' (Pelletier, 1979)) is operative, and that the most appropriate way to extend coverage is to add a lexical rule. Other cases simply represent the tip of the iceberg of missing lexical entries.

The BNC data highlighted both lexical gaps which could have been identified through simple introspection (e.g. nominal *attack*), and more subtle ones such as the transitive verb *suffer* and the MWE *at arm's length*. In future work, we expect to leverage the corpus via shallow parsing techniques to bootstrap semi-automatic lexical expansion efforts. We expect there to be limitations to corpus evidence, however, and that quirky constraints on some lexical entries will only be detectable via introspection. For example, the BNC data revealed a lexical gap for the use of *tell* meaning 'discover' or 'find out' in (1). Introspective investigation revealed that this sense of *tell* requires either one of a small set of modals or *how*: see (2). While a subset of the collocations can be found in the BNC, there is no way to automatically detect the full details of

such idiosyncratic constraints on distribution (following Bender and Kathol (2001), we indicate examples from the BNC with $^{@}$).

(1) $^{@}$Not sure how you can tell.

(2) a. Can/could you tell?
b. Are you able to tell?
c. *They might/ought to tell. (ungrammatical on the intended reading)
d. How might you tell?
e. *How ought they to tell? (ungrammatical on the intended reading)

In addition to known difficult problems (e.g. direct quotes, appositives and comparatives), we found many constructions which we believe would be difficult to notice using either a purely intuition-driven or a purely corpus-driven methodology: they're just a little too obscure to arrive at by introspection, and too rare to notice in a corpus without analysis using a broad-coverage precision grammar. We give just a few examples here: (3) illustrates free-relatives where an adjective is pied-piped along with the relativizer *however*; (4) gives an example of the pre-nominal string *hell of a*, which seems to be in the process of grammaticalization (cf. the reduced form *helluva*); and (5) illustrates a use of det+adj strings as NPs, outside the well-known cases such as *the rich*, which are interpreted as referring to groups of people.

(3) $^{@}$However pissed off we might get from time to time, though, we're going to have to accept that Wilko is at Elland Rd. to stay.

(4) $^{@}$He's a good player, a hell of a nice guy too.

(5) $^{@}$The price of train tickets can vary from the reasonable to the ridiculous.

On the boundary between the grammar illuminating the corpus and the corpus illuminating the grammar, we find sentence fragments like (6)–(8). While these are clearly not grammatical sentences, they are grammatical strings, and some even represent idiomatic frames as in (8). We must therefore extend the grammar to include a wider notion of grammaticality, perhaps grounded in what can serve as a stand-alone utterance in a discourse or similar unit in a text (e.g. see Schlangen, 2003).

(6) $^{@}$The Silence of the Piranhas

(7) $^{@}$Mowbray? Not good enough probably

(8) $^{@}$Once a Catholic, always a Catholic

Whereas ungrammatical items in a manually-constructed test suite serve to contrast with minimally different grammatical examples and demarcate the constraints on a

particular construction, naturally occurring ungrammatical items constitute instead haphazard noise. Even in the BNC, much of which is edited text, one finds significant numbers of ungrammatical strings, due to reasons including spelling and string tokenization errors (e.g. *@*...issues they fell should be important...*), typographical inconsistencies, and quoted speech. While NLP systems should incorporate robust processing techniques to extract such information as is possible from ungrammatical strings in the input, a precision grammar should not be adapted to accommodate them. At the same time, such ungrammatical examples can serve as a test for overgeneration that goes far beyond what a grammar writer would think to put in a manually constructed test suite.

Extragrammatical effects were observed to adversely impact parse coverage due to: (a) unhandled phenomena interfacing unpredictably with the grammar, and (b) word tokenization errors. One common currently untreated phenomenon is structural mark-up (e.g. bullets, item numbers, page/section references). Occurrences of structural mark-up had unexpected effects, such as *a* in (9) being misanalyzed as an article, leading to the prediction of ungrammaticality. A pre-processing strategy can be employed here, although simply stripping the mark-up would be insufficient. An interface with the grammar will be required in order to distinguish between structural and lexical usages of *(I)*, e.g. as illustrated in (10) and (11).

(9) @There are five of these general arrest conditions: (a) the name of the person is not known to the police officer and he or she can not “readily ascertain” it.

(10) @(I) That Mrs Simpson could never be Queen.

(11) @“(I) rarely took notes during the thousands of informal conversational interviews.

In our current grammar implementation, we tokenize proper names based on the output of a POS tagger. Tagging errors and subsequent mistokenization—such as tagging *Whilst* in (12) as a proper noun—contributed to parse failures.

(12) @Whilst doing this you need to avoid the other competitors.

Our treebank annotation strategy successfully identified a large number of sentences and fragments in the BNC for which the current ERG was unable to provide a correct analysis, even where it did offer some (often many) candidate analyses. The paraphrase proposal worked well in diagnosing the specific source of the parse failure. The undergraduate annotator (previously unfamiliar with the ERG) using these techniques was able to correctly identify, diagnose, and document often subtle errors for about 100 BNC examples per day. The annotator’s analysis was evaluated and extended in an item-by-item discussion of 510 such errors with the grammar writers. This precise, detailed classification of errors and their frequency in the subcorpus provides important

guidance to the ERG developers both in setting priorities for hand-coded lexical and syntactic extensions to the grammar, and also in designing methods for semi-automatic acquisition of lexical items on a much larger scale.

Recent advances in processing techniques (Oepen et al., 2002b) make broad-coverage precision grammars even more attractive as components of NLP systems and as testbeds for grammatical hypotheses, both specific (analyses of particular constructions) and general (properties of grammatical formalisms). In the development of a broad-coverage precision grammar, corpora can play a key role, but corpus data alone is not sufficient for at least two reasons: first, precision grammar development requires negative examples to guide the design of constraints that make correct predictions; and second, data sparseness means that certain interesting examples can only be located in a large corpus with the assistance of an existing deep grammar.

References

Bender, E. M. and A. Kathol (2001). Constructional effects of *Just Because ... Doesn't Mean ...* In *Proc. of the 27th Annual Meeting of the Berkeley Linguistics Society*. Berkeley, USA.

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

Copestake, A. and D. 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.

Oepen, S., D. Flickinger, K. Toutanova, and C. D. Manning (2002a). LinGO Redwoods: A rich and dynamic treebank for HPSG. In *Proc. of The First Workshop on Treebanks and Linguistic Theories (TLT2002)*. Sozopol, Bulgaria.

Oepen, S., D. Flickinger, H. Uszkoreit, and J. Tsujii, eds. (2002b). *Efficiency in Unification-Based Processing*. CSLI Publications, Stanford, USA.

Pelletier, F. J. (1979). Non-singular reference: Some preliminaries. In F. J. Pelletier, ed., *Mass Terms: Some Philosophical Problems*, pp. 1–14. Dordrecht, Reidel.

Pollard, C. and I. A. Sag (1994). *Head-driven Phrase Structure Grammar*. The University of Chicago Press, Chicago, USA.

Schlangen, D. (2003). *A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. Ph.D. thesis, University of Edinburgh.