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

> "… every corpus I've had a chance to examine, however small, has taught me facts I couldn't imagine finding out about in any other way"
>
> Chuck Fillmore (1992:35)

1 Introduction

The relative merits of corpus and native speaker judgment data is a topic of long-standing debate in linguistics (Labov 1972; Fillmore 1992, *inter alia*). In this paper, we approach the question from the perspective of grammar engineering, and argue that (unsurprisingly to some, cf. Fillmore) these sources of data are best treated as complementary to one another. Further, we argue that encoding native speaker intuitions in a broad-coverage precision implemented grammar and then using the grammar to process a corpus is an effective way to explore the interaction between the two sources of data, while illuminating both. We discuss how the corpus can be used to constructively road-test such a grammar and ultimately extend its coverage. 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.

Our use of corpus data is limited to evaluating the grammar and exposing gaps in its lexical and constructional coverage, where actual grammar development is based on the combination of corpus and judgment data. In this sense, we distinguish ourselves from the research of, for example, Hockenmaier and Steedman (2002) wherein grammar development is exclusively corpus data-driven in an attempt to enhance coverage over a given corpus (i.e. the Penn Treebank). In this style of approach, only those lexical items observed in the corpus are added to the lexicon, and constructional coverage is tailored to the given corpus. We claim that this approach leads to bias in coverage and restricts the generality of grammatical analyses.

In §2, we review some of the arguments for and against both corpus and intuition data. In §3 we introduce the particular resources we used, viz. the English Resource Grammar (ERG; Copestake and Flickinger 2000) and a portion of the British National Corpus (BNC; Burnard 2000), and outline our methodology for combining the two sources of evidence. In §4 we present our results: a categorization of areas for improvement in the grammar as well as a categorization of sources of 'noise' in the corpus properly treated as outside the domain of the grammar. In §5 we discuss how these results can inform both future grammar development and syntactic theory.

2 Background

While it might seem to be common sense that corpus data and judgment data are complementary sources of evidence, the recent history of the field of linguistics (certainly since the rise of Chomskyan generative grammar) has tended to relegate each of them to competing modes of investigation. Early 20th century American structuralists such as Boas and Sapir relied on both philological sources and elicited data. However, the modern notion of grammaticality (as representative of underlying grammatical principles) was absent from such work, a methodological stance partly due to the behaviorist assumption that mental structure was either non-existent or at least beyond the realm of exploration with empirical data (cf. Bloomfield 1933). It was not until Chomsky's groundbreaking work in generative grammar that the notion of an inherent grammatical structure in the minds of speakers, and thereby an inherent mental structure to the language faculty, (re)entered mainstream modern linguistics (see in particular Chomsky 1957, 1959, 1965). With this new paradigm of linguistic inquiry came also the distinction between "competence" and "performance", i.e. the knowledge a speaker has about his or her language vs. how that knowledge is used (see Chomsky 1964 for an early discussion). The study of competence has since received paramount importance, and native speaker judgments of grammaticality/acceptability are now frequently seen as the only means of investigating it. Corpora are instead (somewhat dismissively) relegated to studies of language use and deemed uninteresting to most generative grammarians, on the grounds that:

- Corpora are limited in size and therefore may not reflect the full range of grammatical constructions.
- Corpora are full of errors due to processing and reflect other extra-grammatical factors (not part of competence).
- Corpora can only provide positive (attested) examples. Without information on contrasting ungrammatical examples, one cannot achieve a complete understanding of competence.

The competence/performance distinction and consequent division of types of data has survived in some form in every version of Chomskyan generative grammar.[1] However, a significant (albeit somewhat dispersed) amount of literature calls into question the primacy of native speaker intuitions as linguistic data. The main arguments are the general slipperiness of grammaticality data, primarily highlighted by the following objections:[2]

- Grammaticality is neither homogeneous nor categorical, but instead represents a cline of relative acceptabilities that vary from speaker to speaker.
- Grammaticality judgments are frequently formed in unnatural contextual vacuums (thereby producing unnatural judgments).
- Social/cultural biases color judgments (and for that matter so do biases of linguists toward their own theories).
- Relying solely on intuitions limits linguists to only the data they have the imagination to think up.

While few linguists have completely given up grammaticality judgments, their tenuousness has given much cause for reevaluation. Some researchers have tried to reduce acceptability judgments to other properties of the language faculty (see e.g. Boersma and Hayes 2001 and Boersma 2004 on the prototype/frequency basis of grammaticality in Stochastic OT). Others have argued instead for more controlled, experimental methods of judgment collection and interpretation to increase the quality of intuition data (Labov 1975, 1996; Schütze 1996; Cowart 1997, Keller and Asudeh 2000; Wasow

2002; Wasow and Arnold to appear), although these techniques are not necessarily practical in all circumstances (see fn. 4). However, a sizable number of linguists have in practice adopted the middle ground between more traditional introspection and corpus-based methods. Fillmore (1992) in particular argues for a methodology of linguistic analysis using corpora as a means of maintaining authenticity as well as a way of discovering new types of expressions, while augmenting this data with (informal) native speaker intuitions as a way of filling out paradigms, exploring possible analyses, and drawing semantic generalizations. This approach solves many of the supposed problems of using corpora (the sparseness of data and lack of a basis for relative acceptability) while tempering the biases inherent in free-for-all introspection (see Svartvik 1992 for a collection of papers including Fillmore's work arguing for and applying this approach). Similarly, descriptive grammars such as Quirk et al. (1985), Sinclair (1990), Biber et al. (1999) and Huddleston and Pullum (2002) have used corpus data in varying degrees to trace out the structure of the English language and unearth generalities, and intuition to fill in the boundaries of grammaticality.

One can also find a contrast between corpus- and judgment-based methods in NLP research. This difference constitutes one of the underlying differences between broad-coverage precision grammars and shallow statistical parsers. 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 (e.g. Copestake and Flickinger 2000; Bouma et al. 2001; Bond et al. 2004). Shallow and/or statistical grammars, however, are often induced directly from treebank/corpus data and make little or no use of grammaticality judgments or intuition (Brill and Marcus 1992; Gaizauskas 1995). Their respective limitations are revealing of the philosophical debates between judgment-based vs. corpus linguistics: precision grammars tend to undergenerate—particularly when presented with novel constructions or lexical items—and shallow grammars to massively overgenerate. With broad-coverage precision grammars, the issue of undergeneration is addressed incrementally by grammar writers working with judgment data and analyses published in the linguistic literature to extend coverage. Developers of shallow grammars, on the other hand, tend not to deal with grammaticality and focus instead on selecting the most plausible of the available parses given the knowledge derived from the corpus.

Following directly on this discrepancy between shallow and deep parsing, we illustrate in this paper how the hybrid approach advocated by Fillmore applies in the world of grammar engineering. We present a methodology for building a broad-coverage precision grammar using corpora as a primary source of data, enhancing and expanding that data with native speaker judgments in order to fully flesh out the paradigms in the corpora while staying true to their authenticity. We outline our methodology in the next section.

3 Methodology

3.1 The English Resource Grammar

The ERG is an implemented open-source broad-coverage precision Head-driven Phrase Structure Grammar (HPSG; Pollard and Sag 1994) developed for both parsing and generation. It has been engineered primarily in the context of applications involving genres such as conversations about meeting scheduling and email regarding e-commerce transactions. While these domains are relatively open-ended, their task-orientation leads to a 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 20,500 distinct word forms.[3] Each lexical item consists of a unique identifier, a lexical type (one of roughly 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 a Minimal Recursion Semantics framework (MRS; Copestake et al. 2003). Of the 10,500 lexical items, roughly 3,000 are multiword expressions (MWEs; Sag et al. 2002).

Development of the ERG has been corpus-driven in the sense that coverage is expanded according to the phenomena which appear in the corpora from the domains to which the ERG has been applied. However, the grammar is not a simple reflection of what has been found in the corpus. Rather, when a corpus example illustrates a previously untreated phenomenon, the grammar engineers construct a space of similar examples drawn from the corpora, then consult the linguistic literature, their intuitions, and other informants in order to map out a space of both grammatical and ungrammatical examples. The total of these investigations serve as the basis for the analyses coded in the grammar. It is in this sense that the ERG stands as an encoding of linguistic intuitions, albeit driven primarily by data found in corpora.[4]

Finally, we would like to emphasize that the ERG is a deep, precision grammar. By this we mean that it relates surface strings not only to syntactic structures but also to explicit, elaborated semantic representations, and further that it attempts to encode a sharp notion of grammaticality: only well-formed strings representing linguistic phenomena analyzed by the grammar will be parsed. Contrasting ill-formed examples will not. Avoiding ungrammaticality cuts down on spurious ambiguity in parsing, simplifying somewhat the problem of parse selection, and is crucial in avoiding ill-formed output in generation. This precision contrasts with shallow approaches to parsing which, as noted above, tend to deal with selecting plausibly parses (generally through stochastic means) rather than grammaticality.

3.2 The BNC Sample

To investigate domain portability, we tested the coverage of the ERG over a random sample of 20,000 strings from the written component of the BNC. Here, the term "string" is used to refer to a "sentence" token according to the original BNC tokenization, and intended to reflect the fact that significant numbers of such tokens are not syntactic sentences (see §4); the random sample was extracted from the 4.6m strings contained in the written portion of the BNC by iteratively selecting a random string from the set of non-selected BNC strings based on the scaled output of a random number generator.

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 seem to have a full lexical span, i.e. which contain only word forms already known to the grammar. An important point to note for the discussion of results in §4 below is that our notion of lexical span still leaves plenty of room for lexical gaps, e.g. where a form may be included in the lexicon with only a subset of its appropriate parts of speech, subcategorization frames, or other idiosyncratic lexical properties. In order to apply this filter to the data, we first tagged the strings for part-of-speech and stripped away any punctuation not handled by the grammar (e.g. commas and periods). 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. 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.

3.3 Combining the Sources of Evidence

We used the ERG to analyze the BNC sample in two ways. In the first instance, we used the ERG to effectively sift out the interesting new cases from the previously analyzed ones. Rather than looking at the raw corpus, we focused on those sentences in the sample which we were not able to parse. This significantly increased the signal-to-noise ratio, where the signal we were interested in was syntactic and lexical gaps in our grammar. We were also able to use the ERG as an aid in analyzing the unparsed sentences, by manually proposing paraphrases until the grammar was able to parse the string. The differences between the parsed paraphrase(s) and the original string indicate the phenomena which need to be added to the grammar or else excluded from it if ungrammatical or extragrammatical (see §4 below).

We illustrate the application of the paraphrase method by way of the following sentence, which the ERG is unable to produce an analysis for:[5]

(1) $^{@}$Always exercise gently to begin with, building up gradually over a period of time and remembering that there is never any need to strain yourself.

We diagnosed the cause(s) of parse failure by first breaking the sentence down into unit clauses and isolating possible sources of error through a depth-first paraphrase process. The resultant unit clauses in the case of (1) are:

(2) a. Always exercise gently to begin with.
b. It builds up gradually over a period of time.
c. Remember that there is never any need to strain yourself.

Applying the paraphrase method, we fed each sentence in (2) into the grammar one by one. The ERG failed to parse (2a), so we then stripped the clause of sentential modifiers, producing *Always exercise gently*. This too failed, whereupon we looked up *exercise* in the lexicon and found it lacked an entry as an intransitive verb. We then tried a paraphrase of the clause using the known intransitive verb *walk*, with and without *to begin with*. *Always walk gently* parsed whereas *Always walk gently to begin with* did not. This suggested that *to begin with* was a MWE currently missing in the ERG, and thus another source of parse failure. Turning to (2b), this expression likewise failed to parse. Once again, we tried stripping the sentential modifiers and proposed *It builds up*. This also produced parse failure, revealing the absence of a lexical entry for the intransitive verb particle construction *build up*. We then verified that *It eats gradually over a period of time* parses, indicating no further problems within this clausal unit. Finally, (2c) also failed to parse, causing us to test the sentence again without the adverb *never*, i.e. *Remember that there is a need to strain yourself*. Since this paraphrase parsed, we concluded *never* was missing a lexical entry that would license this particular construction (as type `adv_vp_aux`, i.e. a polar adverbial licensed by an auxiliary verb). In total, therefore, we were able to identify 4 lexical gaps in (1). This methodology was similarly applied to other parse failures to identify a wide range of lexical and constructional gaps. Note that one advantage of this method is that it does not require an advanced knowledge of the workings of the ERG, only the ability to test linguistic hypotheses.

4 Results

Of the strings with full lexical span, the grammar was able to generate at least one parse for 57%. The parses were manually inspected (using a parse selection tool; Oepen et al. 2002). Of these, 83% of the strings were found to have been assigned a correct (i.e. preferred) parse. At first sight, the absolute coverage figures reported for parsing the BNC with the ERG 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 a large, general-purpose lexicon, a 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 average ambiguity rates of around 100 analyses per input. A recent experiment in manually adding vocabulary for a 300-item excerpt from tourism brochures gave the ERG an initial coverage of above 90% (at an average ambiguity of 187 anal-

Cause of Parse Failure	*Frequency*
Missing lexical entry	41%
Missing construction	39%
Fragment	4%
Preprocessor error	4%
Parser resource limitations	4%
Ungrammatical string	6%
Extragrammatical string	2%

Table 1: Breakdown of causes of parse failure

yses for an average string length of 13 words). In all three scenarios manual parse inspection of ERG outputs confirms analysis correctness measures of at least 90%. The somewhat lower average ambiguity over the BNC data presumably reflects the incomplete lexical coverage diagnosed below (§4.1). The ambiguity levels in each case contrast sharply with the thousands or even millions of 'distinct' analyses typically delivered by treebank-derived statistical parsers (Charniak 1997).

We then turned to the 43% of the original sample which did not receive any parse, and used the methodology described in §3.3 above to diagnose and classify the cause(s) of parse failure. This analysis was carried out over a sampled subset of the original data set, 1190 items, or approximately 14%. In our analysis, we found seven categories of causes of parse failure, as detailed in Table 1. The frequencies in Table 1 were calculated by itemizing the causes of parse failure for each string which did not receive a parse, and summing up the frequency of occurrence of each cause across all strings. Note that the Fragment, Preprocessor error, Parser resource limitations, Ungrammatical string and Extragrammatical string categories apply at the string level. A single string can thus produce (at most) one count for each of these categories. The Missing lexical entry and Missing construction categories, on the other hand, operate at the word/constituent level. We made every attempt to exhaustively identify and sub-classify every such occurrence within a given string, resulting in the possibility for a single string to be counted multiple times in our statistics.

The first two causes of parse failure represent clear lacunae in the grammar, and we argue that the third does as well. Preprocessor errors and parser resource limitations involve other components of the system (the preprocessor and the parser, respectively) failing, and don't necessarily reflect on either the grammar or the corpus. Finally, the last two categories represent noise in the corpus which should not be accommodated in a precision grammar. In the remainder of this section, we illustrate each type of cause in turn, and then evaluate the strategy as a whole.

4.1 Missing Lexical Entries

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). Incomplete categorization refers to missing lexical types for a given word token. While each ERG lexical item is annotated with a specific lexical type which determines its syntactic and semantic behavior, a gap in the full paradigm of types for a given orthographic form (e.g. the noun *table*, but not the verb) leads to parse failure. In some cases it appears that a general process is involved (e.g. a 'universal grinder' treatment of mass uses of prototypical count nouns as in Pelletier 1979), such that the most appropriate way to extend coverage is to add a lexical rule, but many more cases don't seem amenable to this kind of treatment.

Second, syntactically-marked MWEs—notably verb-particle constructions (e.g. *take off*) and determinerless PPs (e.g. *off screen*, *at arm's length*)—cause similar problems. Once again, we find general processes, such as valence patterns for action verbs with and without the completive particle *up*. However, such general processes hardly account for the full range of idiosyncrasies and partial generalizations of MWE. Frequently, then, 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. The flip-side of requiring explicit licensing is a susceptibility to lexical gaps. The frequency with which MWEs appear in the data underscores the fact that they are not a minor annoyance, to be relegated to the periphery. To truly achieve broad-coverage and adequate semantic representations, a precision grammar must treat them as first class entities. Verb-particle constructions, for example, are estimated to account for 1.6% of word token occurrences in the BNC, and determinerless PPs 0.2% (Baldwin et al. to appear).

Regardless of the class of lexical gap, the BNC data highlighted both lexical gaps which could easily 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 intend 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 (3). Introspective investigation revealed that this sense of *tell* requires either one of a small set of modals or *how* (see (4)). While a subset of the collocations can be found in the BNC, there is no obvious way to automatically detect the full details of such idiosyncratic constraints on distribution.

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

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

Further investigation of the corpus revealed instances of *how could (one) tell* and *how does (one) tell*, but not alternative modal collocates such as *how might/would (one) tell*. Thus, having been alerted to the presence of this expression in actual use, we used linguistic intuition in order to determine its full variability (see Fillmore 1992 and Fillmore and Atkins 1992 for a similar hybrid approach to the distribution and semantics of *risk* and *home*).

4.2 Missing Constructions

In addition to known difficult problems (e.g. direct quotes, appositives and comparatives), we found many constructions which were more obscure, and might not have occurred to us as something to analyze without the aid of both the corpus (presenting the examples) and the grammar itself (sifting away all of the previously analyzed phenomena). We present a few such examples here, aiming not to provide full analyses but rather to motivate their interest.

The first example (5) involves the pied-piping of an adjective by its degree specifier in a free relative construction. Such examples were not parsed by the ERG since it explicitly coded the expectation that adjectives were not allowed to pied-pipe in this context.

(5) $^{@}$*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.

At first glance, it appeared that this particular configuration might be restricted to concessive uses of free relatives like (5). However, further investigation into another corpus (the Web, via Google) turned up examples like (6), indicating that this is in fact a general pattern for free relatives.

(6) @The actual limit is *however big* your file system will let the file be.

The second example (7) involves a class of expressions which one might call quasi-partitives.

(7) a. @He's a good player, a hell of a nice guy too.
b. That's a bitch of a problem to solve.
c. *He's a hell on wheels/hell and a half/hell beneath us of a guy.
d. *The hell of a guy that I met at the party last night...

In addition to *hell of a* (and its reduced forms *helluva*/*hella*), one also finds *bitch of a* (7b) and perhaps others. It appears that nothing can intervene between *hell* and *of* (7c). This construction presents a neat little semantic puzzle. First, note that it appears that the construction is restricted to indefinite NPs (7d). Thus it appears that *hell of* is attaching to the NP *a nice guy*, or perhaps *hell of a* is attaching to the $\overline{\text{N}}$ *nice guy*. On the other hand, the *of* can be directly followed by a noun or by an adjective and then a noun. When there is an adjective present, *hell of a* seems to be acting semantically as an intensifier of the adjective. Given ordinary assumptions about semantic composition, it is not immediately clear how an element attaching syntactically to an NP/$\overline{\text{N}}$ could semantically modify an adjective modifier inside that NP/$\overline{\text{N}}$.

The next example (8a) involves exocentric NPs of the form [Det Adj], but (surprisingly, if one believes the textbooks) not restricted to referring to generic classes of humans (cf. *the rich*, *the famous*).[6]

(8) a. @The price of train tickets can vary from the reasonable to the ridiculous.
b. The range of airfares includes the reasonable and the ridiculous.
c. An exhibit of the grotesque is on display at the museum today.
d. *My collection of toy cars include the red and the blue.

Further reflection brought us to examples like (8b), which joins the two exocentric NPs with a conjunction rather than the *from ... to* construct and (8c) which involves only one exocentric NP. The infelicity of (8d) indicates that this construction isn't available with all adjectives, or perhaps with all construals of the resulting NP. We believe that the corpus example (8a) motivates an investigation into the classes of adjectives which can appear in this construction, the classes of referents the resulting NPs can have, and the relationship (if any) between adjective class and resulting potential referent classes.

Our final example (9) involves a construction which licenses the use of any common noun as a title, paired with an enumerator from an ordered list (e.g. numbers, letters, *alpha/bravo/charlie/...*).

(9) @This sort of response was also noted in the sample task for criterion 2.

This example appears to involve a construction somewhat similar to the one that pairs a title like *Prof.* or *Dr.* with a personal name, and raises the question of whether that family of constructions might not include a few other members, again with slightly varied constraints. It is worth noting here that this example also represents a class of phenomena (including number names, quotatives, and time/date expressions) which are relatively frequent and commonplace in corpus data, but tend to go unnoticed in linguistic investigations which are not rooted in corpora. We speculate that this is because they are somehow more context-dependent and are therefore unlikely to crop up in the sort of decontextualized sentence generation which is typical in syntactic research.

We take this to be a validation of our methodology: corpora are a rich source of largely unnoticed lexical items and construction types, some of which are context-dependent in a way which makes them unlikely to be noticed through introspection but still frequent enough to pose a problem for any parser. However, the inherent biases in corpora (e.g. frequency of some uses over others) might mask the underlying paradigms governing the distribution of these items, calling for a broader approach to updating a grammar like the ERG involving introspective analysis. Furthermore, using the existing grammar to analyze the corpus enriches the data sample presented to human analysts, thus enhancing the usefulness of the corpus.

4.3 Fragments

On the boundary between the grammar illuminating the corpus and the corpus illuminating the grammar, we find sentence fragments like (10a–c). While these are clearly not grammatical sentences, they are grammatical strings, and some even represent idiomatic frames as in (10c).

(10) a. $^{@}$The Silence of the Piranhas
b. $^{@}$Mowbray? Not good enough probably
c. $^{@}$Once a Catholic, always a Catholic

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 for a detailed analysis of a wide range of sentence fragments within this framework).

4.4 Preprocessor Errors and Parser Resource Limitations

Preprocessor errors involve common nouns or other elements (e.g. *whilst*) in (11) being mistagged as proper nouns[7] or vice versa, causing errors in tokenization, leading in turn to unparsable inputs.

(11) $^{@}$Whilst doing this you need to avoid the other competitors.

Also, a small number of remaining British spellings caused parse failure in some cases. While these do not reflect directly on the ERG, they do illustrate one kind of noise in the corpus. That is, in any practical application, a precision grammar will have to contend with both inherent corpus noise (see §4.5) and noise added by other components of the NLP system.

Parser resource limitations refer to instances where 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 or modifier/attachment ambiguity. This problem can be mitigated to some degree at the hardware level by increasing the memory, or resolved more substantively through the adoption of a beam search-based parse selection facility. Beam search would take the form of dynamic pruning of improbable edges, determined relative to a treebank constructed from successfully parsed examples (Oepen et al. 2002). With such a facility, 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. For the moment we ignore these limitations (which affected only a small number of candidate sentences).

4.5 Ungrammatical Strings

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 larger NLP systems (into which a precision grammar may be embedded) should incorporate robust processing techniques to extract such information as is possible from ungrammatical strings in the input, the precision grammar per se should not be adapted to accommodate them.[8]

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. This underscores the importance of the treebank annotation step of our methodology. Having a human annotator effectively vet the grammar's analyses also turns up any ungrammatical examples that the grammar (mistakenly) assigned an analysis to.

4.6 Extragrammatical Strings

Extragrammatical effects involve unhandled non-linguistic or quasi-linguistic phenomena, associated with written presentation, interfacing unpredictably with the grammar. A prime example is structural mark-up, which can lead to unexpected effects, such as *a* in (12) being misanalyzed as an article, instead of stripped off the sentence. If *a* is taken as an article, the grammar correctly predicts the string to be ungrammatical. 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 (13) and (14).

(12) $^{@}$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.

(13) $^{@}$(I) That Mrs Simpson could never be Queen.

(14) $^{@}$"(I) rarely took notes during the thousands of informal conversational interviews.

4.7 Evaluation and Summary

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, across all of the types: lexical gaps, constructional gaps, fragments, ungrammatical strings and extragrammatical strings.

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.

5 Conclusions

We have explored the interaction of two types of evidence (corpus data and grammaticality judgments) from the perspective of grammar engineering. Combining the two sources of linguistic evidence as we did—encoding intuitions in a broad-coverage precision grammar and using this grammar to process the corpus—allowed us to explore their interaction in detail.

The corpus provides linguistic variety and authenticity, revealing syntactic constructions which we had not previously considered for analysis, including many which fall outside the realm likely to be explored in the context of decontextualized example generation. Analyzing the corpus with the grammar allowed us to efficiently focus on the new territory, neatly sweeping away the well-known constructions which we have already incorporated. Since the as-yet unanalyzed constructions tend to be lower frequency, this ability to enrich the data that must be gone through by hand is crucial. Insisting on maintaining a notion of grammaticality in our precision grammar (rather than aiming to analyze every string in the corpus) leads us to recognize and categorize the noise in the corpus. Finally, as the corpus examples inspire us to add further analyses to the grammar, we incorporate additional intuition-based evidence as well as attested examples from other corpora gleaned from targeted searches. This is in fact required by the precision grammar approach: If we were to rely only on attested examples to craft our analyses (and especially examples from a single corpus or genre), they would be a very poor match to the actual state of the language indeed. We believe that any such attempt would necessarily end up being too permissive (leading to massive ambiguity problems and ill-formed output in generation) or incoherent, as one tried to incorporate unnatural constraints to match the attested examples too closely.

In illustrating our methodology and providing a taste of the kind of results we find, we hope to have shown that precision grammar engineering serves both as a means of linguistic hypothesis testing and as an effective way to bring new data into the arena of syntactic theory.

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, Stanford University. It has been presented at the International Conference on Linguistic Evidence (Tübingen, Germany, January 2004) and the Fourth International Conference on Language Resources and Evaluation (Lisbon, Portugal, May 2004). We would like to thank Francis Bond, Ivan Sag, Tom Wasow, and anonymous reviewers and audience members from both conferences for their valuable input.

CSLI, Stanford University
210 Panama St
Stanford CA 94305 USA

†*CSSE*
University of Melbourne
Victoria 3010 Australia

**Dept of Linguistics*
University of Washington
Seattle WA 98195 USA

{tbaldwin,jbeavers,bender,danf,ara23,oe}@csli.stanford.edu

Notes

[1] See e.g. Chomsky (2001) and Newmeyer (2003) for recent discussions.

[2] See Labov (1972, 1975, *inter alia*), for early discussion of some of these points; see Schütze (1996) for a detailed summary of critiques of grammaticality.

[3] All statistics and analysis relating to the ERG in this paper are based on the version of 6 June, 2003.

[4] As discussed in §2, a more rigorous alternative to standard introspection would be to use judgment data collected via experimental techniques. However, we find that in the development cycle of a project such as ours, it is not practical to carry out full-scale grammatical surveys for each contrast we want to encode. Thus we continue to use informal methods to collect introspective data (where more sophisticated surveys are not available in the literature) and rely on the corpus to

show us when these methods have gone astray.

5 Following Bender and Kathol (2001), we indicate attested examples with $^{@}$. Unless otherwise noted, all attested examples cited in this paper are from the BNC.

6 Such cases of so-called $\overline{\text{N}}$-ellipsis are of course quite common in a number of other languages (Beavers 2003).

7 In this case, the capitalization might have been one factor in the mistagging.

8 We note, however, that it is possible to adapt a precision grammar to handle ungrammaticality (while recognizing it as such) by incorporating a combination of robustness root conditions, “mal-rules” and error-predictive lexical entries, and still produce a well-formed semantic representation (Bender et al. 2004).

References

Baldwin, Timothy, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger, and Ivan A. Sag

To appear In search of a systematic treatment of determinerless PPs. In Patrick Saint-Dizier, (ed.), *Computational Linguistic Dimensions of Syntax and Semantics of Prepositions*. Kluwer, Dordrecht, Germany.

Baldwin, Timothy, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen

2004 Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 2047–2050. Lisbon, Portugal.

Beavers, John

2003 More heads and less categories: A new look at noun phrase structure. In *Proceedings of the 2003 HPSG Conference*, pp. 47–67. CSLI Publications, Stanford, USA.

Bender, Emily M., Dan Flickinger, Stephan Oepen, Annemarie Walsh, and Tim Baldwin

2004 Arboretum: Using a precision grammar for grammar checking CALL. In *Proceedings of the InSTIL/ICALL Symposium: NLP and Speech Technologies in Advance Language Learning Systems*, pp. 83–86. Venice, Italy.

Bender, Emily M. and Andreas Kathol

2001 Constructional effects of *just because ... doesn’t mean ...* . In *BLS 27*.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan
1999 *Longman Grammar of Spoken and Written English*. Longman, London, UK.

Bloomfield, Leonard
1933 *Language*. Holt, New York, USA.

Boersma, Paul
2004 A Stochastic OT account of paralinguistic tasks such as grammaticality and prototypicality judgments. Rutgers Optimality Archive 648.

Boersma, Paul and Bruce Hayes
2001 Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32: 45–86.

Bond, Francis, 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 *Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*, pp. 554–559. Hainan Island, China.

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

Brill, Eric and Mitchell Marcus
1992 Automatically acquiring phrase structure using distributional analysis. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pp. 155–159. Pacific Grove, USA.

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

Charniak, Eugene
1997 Statistical techniques for natural language parsing. *AI Magazine*, 18: 33–44.

Chomsky, Noam
1957 *Syntactic Structures*. Mouton, The Hague, Netherlands.
1959 A review of BF Skinner's *Verbal Behavior*. *Language*, 35: 26–58.
1964 *Current Issues in Linguistic Theory*. Mouton, The Hague, Netherlands.
1965 *Aspects of the Theory of Syntax*. MIT Press, Cambridge, USA.
2001 *New Horizons in the Study of Language and Mind*. Cambridge University Press, Cambridge, UK.

Copestake, Ann and Dan Flickinger
2000 An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pp. 591–600. Athens, Greece.

Copestake, Ann, Daniel P. Flickinger, Carl Pollard, and Ivan A. Sag
2003 Minimal Recursion Semantics. An introduction. Unpublished ms., `http://www.cl.cam.ac.uk/~acc10/papers/newmrs.ps`.

Cowart, Wayne
1997 *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. SAGE Publications, Thousand Oaks, USA.

Fillmore, Charles J.
1992 "Corpus linguistics" or "computer-aided armchair linguistics". In Jan Svartvik, (ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August, 1991*, pp. 35–60. Mouton de Gruyter, Berlin, Germany.

Fillmore, Charles J. and Beryl T.S. Atkins
1992 Towards a frame-based lexicon: The semantics of risk and its neighbors. In Adrienne Lehrer and Eva Kittay, (eds.), *Frames, Fields, and Contrasts*, pp. 75–102. Erlbaum Publishers, Hillsdale, USA.

Gaizauskas, Rob
1995 Investigations into the grammar underlying the Penn Treebank II. Technical report, Research Memorandum CS-95-25, University of Sheffield.

Hockenmaier, Julia and Mark Steedman
2002 Acquiring compact lexicalized grammars from a cleaner treebank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pp. 1974–1981. Las Palmas, Canary Islands.

Huddleston, Rodney and Geoffrey K. Pullum
2002 *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

Keller, Frank and Ash Asudeh
2000 Constraints on linguistic reference: An experimental investigation of exempt anaphors. Unpublished ms., University of Edinburgh and Stanford University.

Labov, William

1972 *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia, USA.

1975 What is a linguistic fact? In R. Austerlitz, (ed.), *The Scope of American Linguistics*, pp. 77–133. Peter de Ridder, Lisse, Netherlands.

1996 When intuitions fail. In Lisa McNair, Kora Singer, Lise M. Dobrin, and Michelle M. Aucoin, (eds.), *CLS 32: Papers from the Parasession on Theory and Data in Linguistics*, pp. 76–106.

Newmeyer, Frederick J.

2003 Grammar is grammar and usage is usage. *Language*, 79: 679–681.

Oepen, Stephan, Kristina Toutanova, Stuart Shieber, Chris Manning, Dan Flickinger, and Thorsten Brants

2002 The LinGO Redwoods treebank. Motivation and preliminary applications. In *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 1253–1257. Taipei, Taiwan.

Pelletier, F. J.

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

Pollard, Carl and Ivan A. Sag

1994 *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. The University of Chicago Press and CSLI Publications, Chicago, USA and Stanford, USA.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik

1985 *A Comprehensive Grammar of the English Language*. Longman, London, UK.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger

2002 Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pp. 1–15. Mexico City, Mexico.

Schlangen, David

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

Schütze, Carson

1996 *The Empirical Base of Linguistics*. University of Chicago Press, Chicago, USA.

Sinclair, John, (ed.)
1990 *Collins COBUILD English Grammar*. Harper Collins, London, UK.

Svartvik, Jan, (ed.)
1992 *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August, 1991*. Mouton de Gruyter, Berlin, Germany.

Wasow, Thomas
2002 *Postverbal Behavior*. CSLI Publications, Stanford, USA.

Wasow, Thomas and Jennifer Arnold
To appear Intuitions in linguistic argumentation. *Lingua*.