1

Accuracy vs. Robustness in Grammar Engineering

Dan Flickinger

Preface

The conceptual origins of the English Resource Grammar go back to the days of GPSG grammar development in the Natural Language Project at Hewlett-Packard Laboratories beginning in the early 1980s, with Tom Wasow serving as one of the initiators and guiding forces of that research group, which this author joined starting as a summer intern in 1983. It was at HP Labs that we developed the methodology of building and using test suites exhibiting core linguistic phenomena, for regression testing and measurement of progress as we extended grammar coverage while maintaining a high level of linguistic accuracy.

1.1 Introduction

The implementation of a computational grammar for a natural language is an extended exercise in the art of compromise, since the emerging grammar will strive to excel on several measures which are in competition for primacy. The ideal grammar would produce a completely accurate result for every input presented to it, with a mimimum of computational effort. But short of that ideal, any existing grammar will necessarily either emphasize robustness at the expense of accuracy, or favor accuracy while conceding some limitation in robustness. Many modern broad-coverage grammars maximize robustness, often for good practical reasons, but the inevitable corresponding sacrifices in accuracy can be difficult to quantify, since public standards for testing

Readings in Cognitive Science: Papers in Honor of Tom Wasow (Working Title).
Emily M. Bender and Jennifer Arnold.

and comparing grammars are either inadequate or lacking altogether. Hence the relative benefits and costs of using a more robust grammar vs. a more accurate one are often judged instead by task-based success rates within applications. Such 'black-box' measures are not helpful in predicting the success of a grammar in a new application, nor do they afford direct illumination of the linguistic shortcomings of the grammar, insights which could guide its further development.

Robustness and efficiency are relatively easy to measure, but for some applications accuracy is of equal importance, and better methods and annotated corpora will be necessary to enable its effective evaluation. This paper examines some of the engineering trade-offs that have been made in the development of one broad-coverage grammar over the course of its fifteen-year development, with the aim of contributing to the design of more effective grammar evaluation standards. Greater clarity about the nature of the compromises embodied in a grammar should help in designing annotation schemes for test data which reveal the consequences of these choices for accuracy, and thus enable better evaluation of suitability for a given task, and more fine-grained comparison across grammars.

1.2 English Resource Grammar

The English Resource Grammar (ERG: Flickinger, 2000, Flickinger, Copestake, & Sag, 2000, Copestake & Flickinger, 2000) is a broad-coverage grammar which was started in 1994,[1] and which has been under continuous development since then within the Linguistic Grammars Online (LinGO) laboratory at CSLI (Center for the Study of Language and Information, Stanford University).

As an implementation within the theoretical framework of Head-driven Phrase Structure Grammar (HPSG: Pollard & Sag, 1994), the ERG has since its inception encoded both morphosyntactic and semantic properties of English, in a declarative representation that enables

[1]The first version of the English Resource Grammar was designed and implemented at CSLI by Rob Malouf; several other Stanford graduate students also contributed to early implementation work, most notably Emily Bender. The grammar has also benefited significantly over the years from the suggestions, critique, and wealth of syntactic expertise that Tom brought to our weekly meetings, matched by LinGO director Ivan Sag (also a co-founder of the HP Labs NLP effort), and assisted by a steady stream of visiting scholars to the LinGO lab at CSLI. Ann Copestake and Stephan Oepen each authored software platforms central to the ERG's implementation (LKB: Copestake, 2002, [incr tsdb()]: Oepen & Carroll, 2000) and both continue as vital contributors to its development. Broader support now comes from the international research network DELPH-IN (cf. `www.delph-in.net`). An online interface is available at `www.delph-in.net/erg`.

both parsing and generation. While development has always taken place in the context of one or more applications at a time, primary emphasis in the ERG has consistently been on the linguistic accuracy of the resulting analyses, at some expense to robustness. Its initial use was for generation within the German-English machine translation prototype developed in the Verbmobil project (Wahlster, 2000), so constraining the grammar to avoid overgeneration was a necessary design requirement that fit well with the broader aims of its developers.

The ERG consists of a rich hierarchy of types encoding regularities both in the lexicon and in the syntactic constructions of English. As of 2010, the lexicon contains 35,000 manually constructed lexeme entries, each assigned to one of 980 lexical types at the leaves of this hierarchy, where the types encode idiosyncracies of subcategorization, modification targets, exceptional behavior with respect to lexical rules, etc. The grammar also includes 70 derivational and inflectional rules which apply to these lexemes (or to each other's outputs) to produce the words as they appear in text. The grammar provides 200 syntactic rules which admit either unary or binary phrases; these include a relatively small number of highly schematic rules which license ordinary combinations of heads with their arguments and modifiers, and a larger number of construction-specific rules both for frequently occurring phrase types such as coordinate structures or appositives, as in (1):

(1) Kim, my colleague, has arrived.

and for phrase types that occur much less frequently in most corpora, such as vocatives, as in (2):

(2) Kim, can you wait for me?

Statistical models trained on some of the treebanks discussed below are used both in parsing (Toutanova, Manning, Shieber, Flickinger, & Oepen, 2002) and in generation (Velldal, 2008) to rank the relative likelihoods of the outputs, to address the issue of disambiguation which is central to the use of any broad-coverage grammar for almost any task.

1.3 Accuracy Measures via Treebanking

While the measure of *coverage* of a grammar over a corpus is often simply the percentage of items in the corpus for which the grammar assigns at least one analysis, this is a relatively uninformative measurement taken alone, revealing little about either the linguistic adequacy of the analyses or their utility in a given application. For almost any use of an implemented grammar, the accuracy of these analyses is crucially

important, whether measured in terms of the phrasal structures or the semantic dependencies that are assigned to each sentence.

The notion of accuracy for ERG analyses has been determined on the basis of sentence-by-sentence human judgments, with local experts in syntax and semantics meeting weekly for most of the past fifteen years to assist in designing and judging analyses of linguistic phenomena as they appear in application-specific corpora, or in hand-built collections of test sentences (cf. Bender, Flickinger, & Oepen, this volume). The central aims in the design of the grammar are that it will assign one fully correct syntactic structure relating a sentence and its meaning representation, and that all other analyses that the grammar licenses should be linguistically defensible even if pragmatically dispreferred. Correctness of syntactic structures and their corresponding meaning representations is of course theory-dependent, variable in granularity, and subject to lively debate for all but the most basic phenomena, but over time a steadily growing collection of sentences and their preferred ERG analyses has been manually validated, and has been further tested by the use of these analyses in a variety of applications. These annotations of test suites and naturally occurring corpora are recorded in *dynamic treebanks* using the methodology described in Oepen, Flickinger, Toutanova, & Manning, 2004. Of course, there is a more ambitious notion of accuracy in parsing, where the correct analysis is not only produced, but identified as the most likely one out of all the competing analyses licensed by the grammar. These issues of disambiguation and parse ranking are taken up below.

Given the primary emphasis on accuracy, where every word in a sentence (and even every punctuation mark) must be explicitly licensed by some rule of the grammar, some sentences in any naturally occurring corpus of reasonable size will exhibit linguistic phenomena which fall outside the capabilities of the current grammar. These shortcomings of the grammar can be for several reasons: (1) no theoretically sound analysis of the phenomenon in sufficient detail is known; (2) implementation of an existing analysis has so far proved unworkable, due to limitations either of the formalisms employed, or of the ingenuity of the grammarian; (3) adding an available analysis to the grammar would lead to an unacceptable overall increase in ambiguity or in processing costs. It is clear that continued efforts can overcome these shortcomings for many phenomena over time, but Zipf's law (Zipf, 1949) governing the distribution of word frequencies may hold as well for syntactic phenomena in a corpus of sufficient size (cf. Culy, 1998). Since there are many phenomena that occur with relatively low frequency even in very large corpora, any grammar which insists on a high degree of linguistic

accuracy will inevitably encounter obstacles to full robustness.

Fifteen years of development of the ERG have led to a grammar which now consistently assigns fully correct syntactic and semantic analyses to more than 75% of the sentences in previously unseen English texts of many types. There are of course specialized genres which can prove more challenging, such as online newsgroup discussions, chemistry research articles, or technical manuals authored by non-native writers. But recent experiments using the ERG to parse such corpora still consistently give accurate coverage rates above 65%, indicating that the flip side of Zipf's law works here in favor of the grammar: providing analyses of enough of the relatively frequent phenomena will enable relatively robust coverage even for specialized genres. Note that the *observed* coverage numbers for the ERG on any corpus will inevitably be higher, since the grammar can sometimes assign semantically or pragmatically flawed analyses to sentences whose correct analysis would require treatment of phenomena which the grammar does not yet include. For example, the intended meaning of the sentence

(3) Abrams didn't write as many essays as you did poems.

compares the number of essays to the number of poems, but the current ERG, lacking an analysis of comparative sub-deletion (Bresnan, 1973), only assigns the logically possible but unwanted analysis where "as you did poems" is interpreted as "while you wrote poems". Such a sentence in a corpus would count as being parsed by the ERG in the *observed* coverage number, but would be excluded from the *verified* coverage total, after manual annotation of the corpus to construct the treebank.

To date, ERG treebanks identify exactly one analysis (or none) as 'correct' for each sentence, even though for some sentences, the grammar may assign multiple analyses which can be judged correct, even in context. It can be that two syntactically distinct analyses correspond to the same underspecified semantic representation assigned by the ERG, which uses Minimal Recursion Semantics (MRS: Copestake, Flickinger, Pollard, & Sag, 2005) as its formalism. For example, the sentence

(4) They took a nap while he spoke.

might have the subordinate clause "while he spoke" attach either to the verb phrase "took a nap", or to the whole main clause "they took a nap", but the MRS representation will be the same on both attachment decisions. Alternatively, two syntactic analyses may correspond to distinct semantic representations which are pragmatically difficult to resolve, as in some noun-noun-noun compounds such as "airline reservation counter", where it generally doesn't matter whether reference

is to a counter for airline reservations, or a reservation counter for an airline. Such 'spurious' ambiguity can in principle be reduced in a grammar, either by increasing the expressivity of the semantic formalism to allow more underspecification, or by more fine-grained syntactic constraints on the interactions among phenomena. But in practice any broad-coverage grammar implementing a linguistic theory will give rise to instances of spurious ambiguity in any sizeable corpus. At present, this kind of ambiguity is resolved in ERG treebanks via a set of a few dozen heuristics (such as "Attach subordinate clauses as high as possible") which the annotators have negotiated and apply at attachment choice points when treebanking in order to arrive consistently at a single best parse. An alternative approach, not yet investigated for the ERG, would be to leave spurious ambiguity unresolved in the treebank, so some sentences would have multiple analyses all annotated as correct.

Table 1 summarizes the success rates of the current ERG in parsing a variety of collections of English text which have formed the development corpora for NLP projects over the lifespan of the grammar to date. Each of these data sets was parsed and then fully treebanked manually as described above.

Table 1 ERG Treebanks

Corpus type	Number of items	Av. item length	Observed coverage	Verified coverage
Meeting scheduling	11660	7.5	96.8%	93.8%
E-commerce	5392	8.0	96.1%	93.0%
Norwegian tourism	10834	15.0	94.2%	90.1%
SemCor (partial)	2501	18.0	91.8%	82.0%
Wikipedia (CmpLng)	11558	19.5	87.4%	80.0%
Online user forum	578	12.5	85.5%	77.5%
Dictionary defs.	10000	6.0	81.2%	75.5%
Essay	769	21.6	83.2%	69.4%
Chemistry papers	637	27.0	87.8%	65.3%
Technical manuals	4000	12.5	86.8%	61.9%

Each row of the table records

- the total number of individual sentences in a corpus
- the average number of tokens per item in the corpus
- the *observed* coverage: the number of items for which the parser assigned at least one syntactic analysis
- the *verified* coverage, where a correct analysis was identified from among these candidates.

For the first three treebanks, the manually constructed lexicon was extended to ensure that all words used in the corpus have corresponding lexical entries in the ERG. For the remainder, default lexical entries were added automatically for unknown words while parsing, guided by part-of-speech tags assigned by the TnT tagger (Brants, 2000). Brief descriptions of each of these treebanks can be found in the appendix to this chapter.

The parser used in constructing these treebanks was the PET parser (Callmeier, 2000), a bottom-up exhaustive chart parser with packing which employs a statistical model to compute the relative likelihood of each candidate analysis, and selective unpacking to present these analyses in ranked order. Since a few difficult sentences could consume a disproportionate share of the total time and memory required to parse a given corpus, resource limits were imposed on the parser when constructing these treebanks (up to 60 CPU seconds per sentence, or 100K chart edges, or one gigabyte of memory). Some longer sentences in a corpus hit one of these resource limits during parsing, halting before any analyses were found, even though the grammar might well be capable of analyzing such sentences given more time or memory. The negative practical effect of these limitations is most noticeable in the chemistry corpus, where up to 10% of the sentences failed to parse within the resource limits imposed. These limits illustrate one rather obvious but significant compromise between the aim of robustness (treebanking as many sentences as possible in a corpus), and the need for efficiency (constructing the parsed corpus on available hardware in the available time).

Unsurprisingly, the 'survival' rate of treebanked items in a corpus parsed by the ERG is largely correlated with the average sentence length in a corpus, in part simply because longer sentences carry with them a greater likelihood of encountering an occurrence of a linguistic phenomenon outside the scope of the grammar. One other factor bringing down this survival rate is a consequence of the method of preparing these treebanks, involving the strategy employed to contend with highly ambiguous sentences when treebanking.

The Redwoods (Oepen et al., 2004) platform used for treebanking presents the 'forest' of candidate parse trees to the annotator in the form of binary *discriminants* (Carter, 1997), each of which divides the parse forest into one set of trees which have a given property and the complement set which do not. While this approach enables efficient and consistent annotation, its practical use requires that some upper bound be imposed on the number of candidate analyses (the size of the parse forest) recorded for any one sentence. Depending on how well the sta-

tistical model used in parse ranking matches the linguistic phenomena observed in a given corpus, the intended analysis for some sentence may be within the scope of the grammar, yet be unhappily ranked beyond the limit imposed on the number of analyses the annotator considers when treebanking.

This second factor is partly responsible for the more marked contrasts between 'observed' coverage and 'verified' coverage in the treebank of chemistry articles, and the one for the essay "The Cathedral and the Bazaar". The statistical model used when parsing these smaller corpora had been trained on annotations of the Norwegian tourism corpus, which did not provide enough training instances of some linguistic phenomena observed more frequently in these additional corpora. Training new corpus-specific statistical models will very likely lead to a reduction in the damage caused by this mismatch between training data and parsed corpus, enabling more success when treebanking, but this dependence on customized parse-ranking models presents a minor but appreciable obstacle when treebanking a previously unseen corpus. Here the relevant compromise is between the desire on the one hand for both coverage and accuracy, and on the other for minimizing the manual customization costs when treebanking a new text corpus.

1.4 Expanding Grammar Coverage

Throughout its development, the ERG has both benefited and suffered from the fundamental design decision to make primary the accuracy of its linguistic analyses. One significant benefit is the relative ease of applying the grammar to the task of generating well-formed and natural-sounding sentences of English from input meaning representations (MRSs), enabling its use as a generator in the machine-translation systems of Verbmobil (German/English) and LOGON (Norwegian/English: Lønning et al., 2004). Indeed, the ability to generate has proven to be valuable in grammar development itself, since the generator is quick to reveal syntactic structures erroneously licensed by the grammar. Observing and diagnosing such overgeneration often leads to quick and rewarding improvements in the implementation, with the additional benefit of reducing unwanted ambiguity when parsing.

A second important benefit has been the ability to sustain discussions with linguists, through years of grammar development, consulting on the detailed design and evaluation of syntactic and semantic analyses of phenomena implemented in the ERG. By ensuring that the structures licensed by the grammar correspond well to the expectations of the theoretician, the grammar engineer can continue to co-design

computationally tractable treatments of new phenomena with (fellow) theoreticians, even as the grammar's complexity inexorably increases.

However, one ongoing consequence of this emphasis on linguistic accuracy is the lack of analyses for some portion of the sentences in any naturally occurring corpus of English text. While this percentage of unanalyzed sentences steadily shrinks as development of the ERG continues, the remaining shortcomings in robustness are nontrivial, and practical applications using the grammar may require hybrid processing strategies which include an additional and more robust if less accurate analysis engine, in an architecture of the kind studied in Schäfer, 2007. An alternative strategy already used by other linguistically rich broad-coverage grammars such as the PARC LFG English grammar (Butt, Dyvik, King, Masuichi, & Rohrer, 2002) produces a partial analysis when no full analysis is available; an investigation of this approach using the ERG is reported in Zhang & Kordoni, 2008.

As already noted, some kind of hybrid strategy will always be necessary when using linguistically rich grammars if every sentence in a corpus must get some analysis. But the grammarian's aim in development is to continuously reduce the work load for such robustness safety nets. While much of the development of a broad-coverage grammar comes in quite small increments, the ERG has seen several more noteworthy steps in its steady quest for more robust coverage.

One important step was a careful diagnosis of the classes of errors made by the ERG in parsing a nontrivial set of sentences extracted from the British National Corpus (BNC). This study (Baldwin et al., 2005) provided a baseline of coverage and accuracy for the ERG on data from a large corpus, and highlighted the need for a more effective treatment of open-class vocabulary, and in particular multi-word expressions.

A second step in the development of the ERG which significantly improved its robustness focused on an intensive expansion of the manually constructed lexicon, driven by the observation that predicting suitable lexical entries for unknown nouns and adjectives is typically easier than for verbs, which exhibit more variation in the kinds of complements they select. Manually constructed lexical entries were added to the ERG's lexicon for all words which occurred as verbs more than 100 times in the 100-million word BNC. The working hypothesis is that infrequently occurring verbs are more likely to be either simple intransitive or transitive verbs, so the creation of on-the-fly lexical entries for remaining unknown verbs encountered in a text should be straightforward. With this addition of some 2000 verb entries, plus another 3000 entries for other high-frequency words identified by Zhang, 2007's application of van Noord, 2004's error-mining technique, the 'observed'

coverage for the ERG on one subset of the BNC roughly doubled from less than 20% to above 40% when coupled with a simple unknown word predictor based on part-of-speech tags again using the TnT tagger. Evaluation of the accuracy of these analyses has not been carried out for the BNC corpus, but a smaller-scale evaluation of the efficacy of this method (among others) was conducted by Zhang & Kordoni, 2006 using the manually-annotated 2005 Redwoods corpus.

A third and more recent advance in robustness for the ERG has come in the form of a preprocessing component which enables declarative statements of grammar-specific tokenization and normalization rules (Adolphs et al., 2008). These rules cope with phenomena such as time and date expressions, telephone numbers, measure phrases, integers, ratios, and web addresses, while also defining the triggering conditions for introducing on-the-fly lexical entries for proper names and for open-class words which lack the necessary lexical entries in the manually constructed ERG lexicon. While the ERG has employed a preprocessing component for several years already, including accommodation for unknown words, this more recent chart-based approach enables greater consistency and more fine-grained interactions between token-level properties such as mixed case (capitalization) and morphosyntactic or semantic properties defined in the ERG lexicon. One example of improved robustness using this approach is the ability to posit a generic proper name entry for a capitalized word which is already included in the ERG lexicon as some other part of speech, even though in general native lexical entries block the addition of on-the-fly entries to avoid massive spurious ambiguity. Thus in sentence (5), a proper name entry is now created for "Grumpy" even though an adjective entry already exists in the ERG lexicon, so this sentence parses correctly.

(5) We saw Grumpy at Disneyland.

Such re-assignment of words for use as proper names is particularly prevalent in scientific texts that we have analyzed, including the SciBorg chemistry articles and the Wikipedia articles on computational linguistics.

These three improvements to the ERG not only increased the overall robustness of the grammar, but interestingly also improved its accuracy. A richer lexical inventory of frequently occurring verbs avoids shortcomings in part-of-speech tag-based unknown word prediction, leading to an increase in the number of sentences that receive correct analyses, since the lack of a correct lexical entry for even one token in a sentence will prevent the grammar from assigning the right analysis to that sentence. Likewise, better integration of preprocessing with the constraints

of the core grammar enables a closer match in tokenization to the expectations of the grammar, and more accurate on-the-fly lexical entries for unknown words. The result has been a modest but visible improvement of some 5% in the 'survival' rate in the treebanks where not all of the vocabulary is included in the manually constructed lexicon, since a correct parse is now more often available for sentences that previously only got spurious analyses due to erroneous unknown word guesses or incorrect tokenization.

1.5 Limitations to Accuracy

As already noted, giving a precise characterization of accuracy in linguistic analyses has proven to be an elusive goal. Reaching consensus on the correct syntactic structures for sentences of a corpus is difficult even within a single project, and simply not possible across linguistic frameworks, since the phrasal structures assigned are too theory-dependent. Agreement on the semantic dependencies expressed within a sentence may be more achievable across frameworks, and there have been some constructive moves in this direction such as the COLING parser evaluation shared task and workshop (Bos et al., 2008).

Even with the emphasis on linguistic accuracy in the analyses licensed by the ERG, including semantic representations for each sentence, many desirable elements of a 'full' analysis are notably lacking. While the assertions included in the MRS representation assigned by the ERG as the correct analysis of a sentence should all be true, they are far from complete. For one, the ERG currently draws few distinctions in lexical semantics, instead assigning almost every lexical entry its own semantic predicate, and thus failing to express semantic commonality for regularities such as the non-productive nominalization of "arrive" as "arrival", or for synonyms like "buy" and "purchase". Some productive category-changing regularities are expressed in the grammar, such as nominalization with *-ing* ("walk/walking") and the *-ly* adverbial suffix ("quick/quickly"), as well as some productive derivational prefixes like *re-* as in "re-hire".

Idioms are accommodated to some extent in the grammar, adapting the approaches of Nunberg, Wasow, & Sag, 1994 and Riehemann, 2001, but they are only sparsely included as illustrative examples of the formal mechanisms designed to represent them, since they only appear with any significant frequency in one of the corpora treebanked to date, namely SemCor, which includes works of fiction that employ many idioms. Thus the grammar, if it lacked a particular entry in the idiom lexicon, would assign a plausible-looking but incorrect analysis

to a sentence like the following, where in context it is the idiomatic reading that is intended:

(6) They kept tabs on him.

More generally, the ERG expresses few constraints on interpretation that are imposed on a sentence by its linguistic context, instead treating each sentence as an isolated expression. Thus the grammar currently makes no attempt to bind ordinary pronouns to their antecedents, not even when the antecedent is present in the same sentence. Likewise, no attempt is made to constrain the interpretation of elided verb phrases like the "should" used in the sentence in (7).

(7) They didn't even try to win, but we should.

These shortcomings in annotation detail do not sharply distinguish the ERG from other broad-coverage grammars, many of which provided comparable dependency representations for a small set of sentences for the COLING 2008 shared task workshop. Yet the lack of such detail will inevitably limit the utility of these grammars for some tasks and for some linguistic investigations that might otherwise benefit from the analyses the grammars provide.

In the case of the ERG, some of this lack of annotation in its semantic representations is due to quite practical considerations. For example, after observing the productive regularity employing the notion of a "Universal Grinder" (Pelletier, 1975) which, for example, relates words for animals to morphologically identical words which denote the "meat" sense of those animals, the grammarian might be expected to add a derivational rule which captures this regularity, enabling a successful parse of the sentence in (8).

(8) They had dog for lunch.

However, such a rule would effectively double the number of lexical entries introduced into the parse chart for every count noun in a given sentence, leading to an additional computational cost that is rarely repaid, since instances of such "grinding" are very rare in many corpora, including the ones discussed above. This awkward tension is compounded by a corresponding rule for the "Universal Sorter" (Bunt, 1985) which relates mass nouns to derived ones with a countable sense, as in (9).

(9) This is an excellent wine.

Adding a further derivational rule for this regularity would mean that in fact every noun, whether originally count or mass, would now introduce two entries into the parse chart, compounding the computational

costs for every sentence, and again providing only rare benefits in increased robustness and accuracy. The current ERG once again compromises, simply listing in the lexicon a small number of the most frequent nouns which exhibit these two alternations, like "chicken" and "wine", sacrificing some small degree of robustness and accuracy in favor of a substantial gain in processing efficiency.

1.6 Conclusion

Since every broad-coverage grammar implementation will make distinct design choices in the face of the inevitable tensions among the goals of accuracy, robustness, and efficiency, any effective evaluation of the resulting analyses must be fine-grained enough to reveal the consequences of these compromises. As noted, measurements for robustness and for efficiency are relatively straightforward and widely reported as parsing results. But it is ultimately the accuracy of these resulting analyses which determines the effectiveness of a grammar for a given task, and thus better methods and annotated corpora for measuring linguistic accuracy would be welcome. The profile presented here of one grammar's compromises in balancing a desire for high accuracy with a steady push toward more robustness will perhaps contribute to the design and production of these improved measures of linguistic analyses.

Appendix: ERG Treebanks

Meeting/hotel scheduling: VerbMobil

The VerbMobil project (Wahlster, 2000) developed, among its many results, a collection of transcriptions of spoken dialogues each of which reflected a negotiation either to schedule a meeting, or to plan a hotel stay. One dialogue usually consists of 20–30 turns, with most of the utterances relatively short, including greetings and closings, and not surprisingly with a high frequency of time and date expressions as well as questions and sentence fragments. A typical example from this corpus (where commas are often used by the transcribers to indicate short pauses in the recorded dialogue):

> *Looks like we, need to schedule another meeting, in the next couple of weeks*

Discussion of this treebank, along with reports on the development and evaluation of statistical models trained on it, can be found in Oepen et al., 2002, Toutanova et al., 2002, and Toutanova & Manning, 2002.

E-commerce: YY Software

While the ERG was being used in a commercial software product developed by the YY Software Corporation for automated response to customer emails, a corpus of training and test data was constructed and made freely available, consisting of email messages composed by people pretending to be customers of a fictional consumer products online store. The messages in the corpus fall into four roughly equal-sized categories: Product Availability, Order Status, Order Cancellation, and Product Return. A typical example from the corpus:

Don't ship the order and send me a refund immediately.

Like the Verbmobil corpus, this data consists of relatively short utterances, including a high frequency of sentence fragments, and some questions, but also a much more frequent use of commands.

Norwegian tourism: LOGON

The Norwegian/English machine translation research project LOGON (Lønning et al., 2004) acquired for its development and evaluation corpus a set of tourism brochures originally written in Norwegian and then professionally translated into English. The project paid for additional professional English translations of these brochures to enable better evaluation studies, producing a sentence-aligned pair of freely available data sets, with the English corpus consisting of 9000 sentences. These are augmented with another 1300 English sentences taken from public-domain Norwegian tourism web sites. The corpus, not surprisingly, consists almost entirely of declarative sentences and many sentence fragments, where the average number of tokens per item is higher than in the Verbmobil and e-commerce data. A typical example:

If you would rather go fishing, there are opportunities in both Øvre Sjodalsvatn and Bessvatn.

More information on the LOGON project can be found at the web site `www.emmtee.net`.

SemCor

The freely available SemCor corpus (Miller, Leacock, Tengi, & Bunker, 1993) consists of 230,000 words of text extracted primarily from the one-million-word Brown corpus (Kucera & Francis, 1967), and tagged with WordNet senses. Work is now underway in collaboration with researchers at the University of Melbourne to construct a treebank for the subset of SemCor which is fully sense-tagged. At present 2500 sentences are included in this emerging treebank, whose average sentence length is greater than in the LOGON texts. A typical example:

Anyone's identification with an international struggle, whether warlike or peaceful, requires absurd oversimplification and intense emotional involvement.

Wikipedia: Computational Linguistics

In collaboration with researchers at Oslo University, we have constructed a treebank for 100 Wikipedia articles on Computational Linguistics and closely related topics, for use in studies including information extraction and parse selection (Ytrestøl, Flickinger, & Oepen, 2009). The treebank of 11558 sentences comprises 13 of the 16 sets of articles, with the remaining three sets held out for testing. The corpus contains mostly declarative, relatively long sentences, along with some fragments. The original wiki markup is preserved in the treebank, accommodated in the ERG by a small number of wiki-specific preprocessing rules. A typical example:

'''Computational linguistics''' is an [[interdisciplinary]] field dealing with the [[Statistics—statistical]] and/or rule-based modeling of [[natural language]] from a computational perspective.

Online user forum: ILIAD

Again in collaboration with the University of Melbourne, construction is underway on a treebank of data extracted from Linux user web forums, as part of the ILIAD (Improved Linux Information Access by Data Mining) project. Only a few hundred sentences have been treebanked so far, and the mix of non-native English and highly informal usage presents an engaging challenge for a high-precision grammar like the ERG. A typical example from the corpus:

Not sure if you ever got Linux installed dbessell, but this brings up a good point.

Dictionary definitions: GCIDE

In a study with researchers at NTT on the feasibility of extracting ontology relationships from dictionary definitions (Nichols, Bond, & Flickinger, 2005) using the ERG, a treebank was constructed with 10,000 English definition sentences from the GNU Contemporary International Dictionary of English (GCIDE). The data includes a very high frequency of relatively short fragments, but also a perhaps surprising wealth of linguistic phenomena. A typical example:

Form: to shape, mold, or fashion into a certain state or condition;

Essay: "The Cathedral and the Bazaar"

The ERG is just one of many grammars under development within a common implementation framework provided by researchers working in the international collaboration called DELPH-IN (`www.delph-in.net`). To further the study of cross-linguistic comparisons among these grammars, and in particular the semantic representations they compose, the consortium resolved to construct treebanks for each grammar of translations of the essay "The Cathedral and the Bazaar" by Eric Raymond. The average length and the linguistic complexity of these sentences is markedly higher than the other treebanked corpora. A typical example:

> *One key to understanding is to realize exactly why it is that the kind of bug report non-source-aware users normally turn in tends not to be very useful.*

Chemistry papers: SciBorg

In the context of a joint project with researchers at University of Cambridge on an eScience project called SciBorg (Rupp et al., 2008), focused on knowledge extraction from a large collection of chemistry research papers, we collaborated with a domain expert to construct a treebank of papers from this collection. The average length of the sentences in this corpus is considerably greater than in the other treebanked corpora, and consists almost entirely of full declarative sentences. The text is preprocessed with a set of chemistry-specific rules to deal with chemistry compound names, formulae, etc. A typical example after this preprocessing:

> *By taking advantage of the growth steering properties of the OSCARCOMPOUND film we were able to prepare nearly perfectly ordered hexagonal arrays of OSCARCOMPOUND clusters with a uniform distance of 4.5 nm between the particles.*

Technical manuals: CheckPoint

The German Artificial Intelligence Research Center (DFKI) conducted an investigation into the use of deep grammars like the ERG and its German counterpart in a hybrid system for grammar-checking for technical manuals (Crysmann, Bartomeu, Adolphs, Flickinger, & Klüwer, 2008). Using anonymized real-world data provided by the Berlin-based software company Acrolinx GmbH, we built a treebank of 4000 sentences (many containing errors), to train a genre-specific statistical model. For this task, the ERG was extended with a small set of robustness rules to explicitly license some mild but frequent instances of mismatches with the standard register defined in the ERG, such as

omitted determiners or the null objects also found in recipes (Culy, 1996). The relative noisiness of this data is reflected in the lower survival rate of the resulting treebank. A typical example:

> *Park tractor on flat level surface, shut engine off and place transmission in park.*

References

Adolphs, P., Oepen, S., Callmeier, U., Crysmann, B., Flickinger, D., & Kiefer, B. (2008). Some fine points of hybrid natural language parsing. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (pp. 1380–1387). Marrakech, Morocco.

Baldwin, T., Beavers, J., Bender, E. M., Flickinger, D., Kim, A., & Oepen, S. (2005). Beauty and the beast: What running a broad-coverage precision grammar over the BNC taught us about the grammar — and the corpus. In S. Kepser & M. Reis (Eds.), *Linguistic evidence: Empirical, theoretical, and computational perspectives* (pp. 49–70). Berlin: Mouton de Gruyter.

Bender, E. M., Flickinger, D., & Oepen, S. (this volume). Grammar engineering and linguistic hypothesis testing: Computational support for complexity in syntactic analysis. In E. M. Bender & J. Arnold (Eds.), *Readings in cognitive science: Papers in honor of tom wasow.* Stanford: CSLI.

Bos, J., et al. (Eds.). (2008). *Coling 2008: Proceedings of the workshop on cross-framework and cross-domain parser evaluation.* Manchester, UK: Coling 2008 Organizing Committee.

Brants, T. (2000). TnT - A statistical part-of-speech tagger. In *Proceedings of the 6th ACL Conference on Applied Natural Language Processing* (pp. 224–231). Seattle, WA.

Bresnan, J. (1973). Syntax of the comparative clause construction in English. *Linguistic Inquiry*, *4* , 275 – 343.

Bunt, H. C. (1985). *Mass terms and model theoretic semantics.* Cambridge University Press.

Butt, M., Dyvik, H., King, T. H., Masuichi, H., & Rohrer, C. (2002). The parallel grammar project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation* (pp. 1 – 7). Morristown, NJ.

Callmeier, U. (2000). PET — A platform for experimentation with efficient HPSG processing techniques. *Natural Language Engineering (Special Issue on Efficient Processing with HPSG)*, *6*(1), 99 – 108.

Carter, D. (1997). The TreeBanker. A tool for supervised training of parsed corpora. In *Proceedings of the Workshop on Computational Environments for Grammar Development and Linguistic Engineering* (pp. 9–15). Madrid, Spain.

Copestake, A. (2002). *Implementing typed feature structure grammars*. Stanford, CA: CSLI Publications.

Copestake, A., & Flickinger, D. (2000). An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second Linguistic Resources and Evaluation Conference* (pp. 591 – 600). Athens, Greece.

Copestake, A., Flickinger, D., Pollard, C., & Sag, I. A. (2005). Minimal Recursion Semantics: An introduction. *Journal of Research on Language and Computation*, *3*(4), 281 – 332.

Crysmann, B., Bartomeu, N., Adolphs, P., Flickinger, D., & Klüwer, T. (2008). Hybrid processing for grammar and style checking. In *Proceedings of the 22nd International Conference on Computational Linguistics* (pp. 153–160). Manchester, England.

Culy, C. (1996). Null objects in english recipes. *Language Variation and Change*, *8*, 91 – 124.

Culy, C. (1998). Statistical distribution and the grammatical/ungrammatical distinction. *Grammars*, *1*(1), 1 – 13.

Flickinger, D. (2000). On building a more efficient grammar by exploiting types. *Natural Language Engineering (Special Issue on Efficient Processing with HPSG)*, *6*(1), 15 – 28.

Flickinger, D., Copestake, A., & Sag, I. A. (2000). HPSG analysis of English. In W. Wahlster (Ed.), *Verbmobil: Foundations of speech-to-speech translation* (pp. 321 – 330). Berlin, Germany: Springer.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press.

Lønning, J. T., Oepen, S., Beermann, D., Hellan, L., Carroll, J., Dyvik, H., et al. (2004). LOGON. A Norwegian MT effort. In *Proceedings of the Workshop in Recent Advances in Scandinavian Machine Translation* (pp. 1–6). Uppsala, Sweden.

Miller, G. A., Leacock, C., Tengi, R., & Bunker, R. T. (1993). A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology* (pp. 303 – 308). Plainsboro, NJ.

Nichols, E., Bond, F., & Flickinger, D. (2005). Robust ontology acquisition from machine-readable dictionaries. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (pp. 1111 – 1116). Edinburgh.

van Noord, G. (2004). Error mining for wide-coverage grammar engineering. In *Proceedings of the 42nd Meeting of the Association*

for Computational Linguistics (pp. 446–453). Barcelona, Spain.

Nunberg, G., Wasow, T., & Sag, I. A. (1994). Idioms. *Language*, *70*, 491 – 538.

Oepen, S., & Carroll, J. (2000). Performance profiling for parser engineering. *Natural Language Engineering (Special Issue on Efficient Processing with HPSG)*, *6*(1), 81 – 97.

Oepen, S., Flickinger, D., Toutanova, K., & Manning, C. D. (2004). LinGO Redwoods. A rich and dynamic treebank for HPSG. *Journal of Research on Language and Computation*, *2*(4), 575 – 596.

Oepen, S., Toutanova, K., Shieber, S., Manning, C., Flickinger, D., & Brants, T. (2002). The LinGO Redwoods treebank: Motivation and preliminary applications. In *Proceedings of the 19th International Conference on Computational Linguistics* (pp. 1 – 5). Taipei, Taiwan.

Pelletier, F. J. (1975). Non-singular reference: Some preliminaries. *Philosophia*, *5*, 451–465.

Pollard, C., & Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. Chicago, IL, and Stanford, CA: The University of Chicago Press and CSLI Publications.

Riehemann, S. (2001). *A constructional approach to idioms and word formation*. Unpublished doctoral dissertation, Stanford University, Department of Linguistics.

Rupp, C., Copestake, A., Corbett, P., Murray-Rust, P., Siddharthan, A., Teufel, S., et al. (2008). Language resources and chemical informatics. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (pp. 2196 – 2200). Marrakech, Morocco.

Schäfer, U. (2007). *Integrating deep and shallow natural language processing components - representations and hybrid architectures*. Doctoral dissertation, Saarland University, Saarbrücken, Germany.

Toutanova, K., & Manning, C. D. (2002). Feature selection for a rich HPSG grammar using decision trees. In *Proceedings of the 6th Conference on Natural Language Learning* (pp. 1–7). Taipei, Taiwan.

Toutanova, K., Manning, C. D., Shieber, S. M., Flickinger, D., & Oepen, S. (2002). Parse disambiguation for a rich HPSG grammar. In *Proceedings of the 1st Workshop on Treebanks and Linguistic Theories* (pp. 253–263). Sozopol, Bulgaria.

Velldal, E. (2008). *Empirical realization ranking*. Unpublished doctoral dissertation, University of Oslo, Department of Informatics.

Wahlster, W. (Ed.). (2000). *Verbmobil. Foundations of speech-to-speech translation*. Berlin, Germany: Springer.

Ytrestøl, G., Flickinger, D., & Oepen, S. (2009). Extracting and annotating wikipedia sub-domains: Towards a new escience community resource. In *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theory* (pp. 185 – 197). Groningen.

Zhang, Y. (2007). *Robust deep linguistic processing*. Doctoral dissertation, Saarland University, Saarbrücken, Germany.

Zhang, Y., & Kordoni, V. (2006). Automated deep lexical acquisition for robust open texts processing. In (pp. 275–280). Genoa, Italy.

Zhang, Y., & Kordoni, V. (2008). Robust parsing with a large HPSG grammar. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (pp. 1888–1893). Marrakech, Morocco.

Zipf, G. (1949). *Human behavior and the principle of least-effort*. Addison-Wesley.