# Measuring Implemented Grammar Evolution

*Dan Flickinger*

Stanford University

danf@stanford.edu

ABSTRACT
Natural language grammar implementations that are constructed manually can grow over time to become large, complex resources covering an ever wider range of language phenomena. It would be useful both for the grammarian and for the users of a grammar to have a good understanding of how a later version of that grammar behaves in contrast to an earlier version, particularly in terms of the treatment of linguistic phenomena. This paper presents a case study of the evolution of an implemented grammar, comparing two versions of the same grammar by measuring several properties of the analyses recorded using that grammar in two corresponding versions of an associated dynamic treebank.

KEYWORDS: implemented grammar evolution, linguistic phenomena, dynamic treebank.

## Introduction

Manually constructed grammars of natural languages can grow over time to become large, complex resources that provide detailed linguistic analyses of a wide range of language phenomena. It would be useful both for the grammarian and for the users of a grammar to have a good understanding of how a later version of that grammar behaves in contrast to an earlier version, particularly in terms of the treatment of linguistic phenomena. For the grammarian, such a comparison would show changes in the breadth and depth of analyses of phenomena as a measure of progress, and for the user, the comparison would help set expectations about how suitable the grammar will be for a specific corpus or genre of text. Treebanks can play a central role in measuring the evolution of a grammar, by enabling fine-grained comparison of the grammar's changes in precision and recall over a constant corpus, not only for whole sentences but also for phenomena identified in the analyses of these sentences recorded in the treebanks. This paper presents a case study of grammar evolution, comparing two versions of the same grammar, the English Resource Grammar (Flickinger, 2000, 2011), by measuring several properties of the analyses recorded using that grammar in two corresponding versions of the Redwoods Treebank (Oepen et al., 2004).

## 1 Grammars and Treebanks

The English Resource Grammar (ERG) is an open-source, broad-coverage, declarative grammar implementation for English, developed within the Head-driven Phrase Structure Grammar (HPSG) framework of Pollard and Sag (1994). This linguistic framework places most of the burden of linguistic description on the lexicon, employing a relatively small number of highly schematic syntactic rules to combine words or phrases to form larger constituents. Each word or phrase (more generally, each sign) is defined in terms of feature structures where the values of attributes are governed by general principles such as the Head Feature Convention, which ensures the identity of a particular subset of attribute-value pairs on a phrasal node and on one distinguished daughter, the head of the phrase. Many of these generalizations aim to be language-independent, constraining the space of possible analyses for linguistic phenomena. Central to the HPSG framework is the inclusion in each sign of constraints on multiple dimensions of representation, including at least syntactic and semantic properties, so that rules, lexical entries, and principles determine semantic well-formedness just as they do syntax. Under continuous development at CSLI since 1994, the ERG claims to provide syntactic and semantic analyses for the large majority of common constructions in written English text.

The two versions of the grammar used for comparison here are the most recent stable version of the grammar (ERG 2018) and the previous stable version from four years earlier (ERG 1214). The 2018 version consists of a 40,000-word lexicon instantiating 1400 leaf types in a large lexical type hierarchy, as well as 110 derivational, inflectional, and punctuation rules, and 250 syntactic rules. At this level of description, the previous 1214 version is somewhat smaller, consisting of a 38,000 word lexicon, 1250 leaf types, 80 lexical rules, and 210 syntactic rules.

Associated with each stable version of the ERG is a corresponding treebank which is manually updated to validate and improve the recorded analysis of each sentence in the treebank that can be correctly parsed using the grammar. For each sentence, an annotator has used a customized tool to identify the intended analysis by making choices from among the set of binary *discriminants*, which encode contrasts among alternative analyses at lexical and phrasal levels (Carter, 1997). These sentence-specific discriminants can be automatically reapplied to a freshly produced parse forest that resulted from using a revised version of the grammar, so that the human annotator need only attend to those sentences in the treebank where the existing discriminant choices did not fully resolve the ambiguities in the new parse forest. This need for additional annotation can result from additional ambiguity introduced by the revised grammar, from a change of analysis such that the grammar no longer makes available the previously recorded analysis, or from newly available analyses for sentences which were previously not in the scope of the grammar. Where the set of available analyses has changed for a sentence in the treebanked corpus, the annotator typically only needs to make a small number of additional decisions among the newly presented discriminants, apart from previously unparsable sentences, and hence the annotation costs of keeping this dynamic treebank up to date with each new stable grammar version remains manageable.

The two versions of the ERG-parsed treebank used for this comparison are subsets of the Redwoods Ninth Growth, parsed with ERG 1214, and the emerging Tenth Growth, parsed with ERG 2018. Both treebank versions record annotations for the same set of varied text corpora, consisting of 58,000 sentences from the following sources:

- Eric Raymond essay on open-source software (769 items)
- Wikipedia – 100 articles from Wikipedia on computational linguistics (11556 items)
- SemCor – Subset of sense-annotated portion of Brown corpus (3103 items)
- DeepBank – Wall Street Journal text in the Penn Treebank, secs. 00–04 (9648 items)
- Verbmobil – Dialogues on meeting scheduling and hotel booking (12393 items)
- LOGON – Brochures and web text on Norwegian tourism (11596 items)
- Tanaka – Subset of English learner data used for Pacling 2001 (3000 items)
- E-commerce – Customer emails on product sales (5793 items)

One important difference in the two versions of the treebank is that two distinct annotation tools were used, both employing discriminant-based disambiguation, with the earlier ‘classic’ one enabling choice only among the top-ranked 500 parses for each sentence (based on a previously trained model), but the later one preserving the full (packed) parse forest. This difference has two primary effects: first, for at least some items in the corpus, the best analysis was not ranked among the top 500, and thus was not available to the annotator using the classic tool for the Ninth Growth; and second, preserving the packed full forest using the latter tool eliminated the sometimes considerable processing cost of unpacking, so that more complex sentences could be parsed within reasonable resource limits and presented to the annotator for disambiguation for the Tenth Growth. As a result of the reduction in parsing cost for the full-forest method, the raw coverage numbers for the two versions of the treebank are not directly comparable, although the differences can be mitigated by noting for each version the number of items in each corpus for which parse failure was affected by resource limitations.

## 2 Measurements of Evolution

We have already seen that the two versions of the grammar differ noticeably in the inventories of linguistic objects that comprise them, with the more recent 2018 version containing a slightly larger lexicon, and markedly more rules, both lexical and phrasal. To study the effects of these changes in the parsing of text, it should be worthwhile to quantify both coarse-grained properties such as overall coverage (parsability) and percentage of items where the recorded analysis has changed, as well as more fine-grained properties such as the usage of particular rules and lexical types, as indicators of the frequency of linguistic phenomena in the corpus that are within the scope of the grammar.

## 2.1 Corpus-level metrics

Table 1 provides a high-level view of the coverage of the 1214 version of the grammar for each component corpus, with the data taken from the file `redwoods.xls` included with the ERG source files for this earlier version. The table reports for each component the total number of items, the average number of tokens per item, the number (and percentage) of items parsed, and the items verified and thus included in the treebank.

| Corpus | Items | Tokens | Parsed | Verified |
|---|---|---|---|---|
| Essay | 769 | 22 | 711 (92%) | 604 (79%) |
| Wikipedia | 11558 | 18 | 10649 (92%) | 9237 (80%) |
| SemCor | 3103 | 18 | 2923 (94%) | 2560 (83%) |
| DeepBank | 9648 | 21 | 9255 (96%) | 8450 (88%) |
| Verbmobil | 12393 | 7 | 11949 (96%) | 11406 (92%) |
| LOGON | 11956 | 14 | 11664 (98%) | 11024 (92%) |
| Tanaka | 3000 | 12 | 2890 (96%) | 2814 (94%) |
| E-commerce | 5793 | 9 | 5627 (97%) | 5420 (95%) |

Table 1: Redwoods Ninth Growth (ERG 1214)

As noted above, the direct comparison of coverage numbers for the two versions of the Redwoods treebank is not fully informative, especially because of the differing effects of resource limits when parsing to prepare the data for annotation. However, the number of items affected by resource limits was typically different by only one or two percent for each component corpus, so that for example the Brown sub-corpus of 3103 items saw 55 unparsed items hitting resource limits in the Ninth Growth contrasted with 17 in the Tenth Growth. Hence the comparison of verified coverage annotations for the two versions of the treebank shown in Table 2 should be viewed with this potential offset in mind, varying slightly from component to component.

| Corpus | Ninth (%) | Tenth (%) |
|---|---|---|
| Essay | 79 | 93 |
| Wikipedia | 80 | 89 |
| SemCor | 83 | 91 |
| DeepBank | 88 | 93 |
| Verbmobil | 92 | 93 |
| LOGON | 92 | 96 |
| Tanaka | 94 | 96 |
| E-commerce | 95 | 98 |

Table 2: Verified coverage for Redwoods with ERG 1214 vs. ERG 2018

Setting aside these minor effects from differences in the treebanking tools, the improvements in the number of successfully annotated items when comparing the Ninth and Tenth Growths should correlate with improvements in the linguistic analyses introduced during the four years of development between the 1214 and 2018 versions of the ERG. An examination of some of these changes in the grammar's treatment of linguistic phenomena is the focus of the next section.

## 2.2 Phenomenon-based metrics

For component corpora where the average number of tokens per sentence is lower, such as for the scheduling dialogues of Verbmobil or the English learner sentences of Tanaka, the change in the number of successfully annotated items is not dramatic, suggesting that enhancements in the depth or breadth of linguistic analysis are less likely to be evidenced in these sentences. In sharp contrast, those corpora with sentences of greater average length, such as the Brown corpus of SemCor and the newspaper text of DeepBank, should help to illuminate substantive changes in the grammar from the older version to the newer one.

One example of a familiar linguistic phenomenon that might be expected to remain steady in its frequency of use within the two versions of the treebank is the passive, whose implementation in the ERG is divided into several subtypes, including not only the ordinary structure in *The cat was chased by the dog* but also the more interesting variants in *That author is looked up to by everyone* and *She can be relied on to succeed*. However, Table 3 shows several nontrivial changes from one version of the treebank to the next, including a notably higher number of uses of the ordinary passive affecting the direct object of the verb, and of the *relied on* type. In addition, the table reflects the addition in ERG 2018 of an analysis of infrequently used passives of the *looked up to* variety, lacking in the 1214 version of the grammar.

| Passives | Ninth | Tenth |
|---|---:|---:|
| All types | 10786 | 11728 |
| *was admired* | 10151 | 11039 |
| *was given (something)* | 553 | 585 |
| *was relied on* | 309 | 403 |
| *was believed that S* | 73 | 80 |
| *was looked up to* | 0 | 3 |

Table 3: Number of items with passive in two versions of Redwoods

The increase in the presence of the most frequent type of passive appears to be correlated with the higher levels of annotation success for those corpora with a greater average sentence length. For example, the 769-sentence open-source essay improved by 14% for successfully annotated sentences, and the number of annotated sentences in that essay using the ordinary passive increased from 137 to 174. Similarly, successful annotation for the 3100-sentence subset of the Brown corpus (SemCor) improved from 83% to 91%, with a corresponding increase in the use of the ordinary passive from 494 items to 583. This alignment of increased use of passives with increased verified coverage does not provide any clear indication that improvements in the grammar's analysis of passives have contributed materially to greater parsing success, since it may well be the case that other grammar improvements enabled more sentences to be parsed, and those sentences simply also make use of passives, so they appear in greater numbers in the recorded analyses.

Another construction type that appears frequently across genres is the relative clause, which is again analyzed via several subtypes in the ERG, as shown in Table 4. Here, too, the overall frequency of use of relative clauses appears to be little affected by changes in the grammar from the 1214 version to 2018, with the greater number of uses in the Tenth Growth roughly correlated with overall improvements in successful annotation. This consistency across versions holds as well for phenomena also grouped with relative clauses by (Huddleston and Pullum, 2002) in their chapter 12, including free relatives (*whatever we needed*) and *it*-cleft constructions (*it was on Tuesday that we scheduled the workshop*). Thus, while the 2018 version of the grammar draws some finer distinctions in its analysis of relative clauses than did the 1214 version, for example distinguishing PP fillers (*on whom we relied*) from NP fillers (*who we admired*), the differences are not reflected in dramatic changes in the number of analyzed sentences exhibiting this class of phenomena.

| Relative clauses | Ninth | Tenth |
|---|---|---|
| All types | 8205 | 8870 |
| *the book which we admired* | 4223 | 4722 |
| *the book admired by everyone* | 2995 | 3304 |
| *the book we admired* | 1021 | 1045 |
| *the guy to talk to* | 643 | 700 |
| *the day we arrived* | 88 | 100 |

Table 4: Number of items with relative clause in two versions of Redwoods

In contrast to these two examples of frequently occurring phenomena where inspection of the treebanks suggests that little has changed in the analyses provided by the two versions of the grammar, there are phenomena whose analysis is clearly different in the two grammar versions. One relatively frequent example is found in constrained but relatively productive noun-noun compounds where the left member is inflected for plural, as in *systems analyst* or *weapons production*, contrasted with **flowers garden* or **towels rack*. These plural compounds also include conjoined nouns as left members, as in *health and family welfare agencies*. The 1214 version of the ERG, presumably sensitive to the ungrammaticality of compounds such as **flowers garden*, did not provide a productive compounding rule for plural or conjoined left members, instead attempting to lexically list frequently found non-heads such as *systems* or *weapons*. In the 2018 version of the grammar, syntactic constructions have been added to admit plural compounds, improving overall coverage at the cost of overgenerating previously rejected compounds such as **flowers garden*. The Tenth Growth records 1055 items whose analyses use these constructions, and many if not most of those sentences would have not been included in the Ninth Growth, apart from the ones for which a use-specific lexical entry had been included in the lexicon.

Several other syntactic constructions have been added for the 2018 version of the grammar, and while they are not individually highly frequent, their successful use helps in the aggregate to account for some of the increases in annotations for the Tenth Growth compared to the Ninth. Table 5 shows the number of items analyzed in the Tenth Growth using some of these new constructions, where these items lack correct analyses in the Ninth Growth.

| Other phenomena | Example | Tenth |
|---|---|---|
| Appositive with measure-NP | *Summit hike, 20 km* | 65 |
| Indef NP as clause modifier | *A good scholar, it was likely she would thrive.* | 44 |
| Coordination of selected-for PPs | *relied on us and on you* | 26 |
| Parenthetical adjective | *the parent (subsuming) class* | 24 |

Table 5: Number of items using constructions only included in ERG 2018

## 3 Related work

Since the discriminant-based approach to treebank construction and maintenance employed for Redwoods has also been adopted by developers of the TREPIL project for treebanks recording analyses using grammars in the Lexical-Functional Grammar linguistic framework, the concept of grammar and treebank evolving in lockstep is also explored there (Rosén et al., 2005, 2016). However, work in this project has not to date focused on using distinct successive stages of the grammar/treebank pair to help to illuminate systematic changes made to the grammar from one version to the next.

A more direct connection to the present work is a hypothesis proposed in Bender et al. (2012), suggesting on p. 192 that software developed for searching treebanks could be adapted to "facilitate the exploration of the evolution of analyses either of particular examples in an implemented grammar, or of classes of examples." The authors acknowledge that such an extension to their search interface would require considerable additional effort; the present study has not aimed at designing such a user interface, instead employing direct textual search within the recorded treebank analyses for the names of specific constructions or groups of constructions associated with phenomena of interest. This direct search method demands a level of familiarity with the naming conventions used for rules and lexical types in the ERG, but these are documented at http://moin.delph-in.net/ErgTop.

It should be noted that the term *grammar evolution* is sometimes used to refer to the process of grammar change over time within a linguistic community, but this is not related to the present study, where the language being analyzed is presumed to be constant for the moment, and it is instead the grammar's *implementation* which is treated as evolving over time in order to express improved analyses of the language.

## 4 Conclusion and Outlook

Dynamic treebanks such as Redwoods or those developed in the TREPIL project have multiple uses; this study is an exploration of one additional use of such annotations, to provide insights about the often murky nature of the changes made to a manually constructed grammar over relatively long periods of time. By examining frequencies of the use of specific rules or groups of rules used in the analysis of individual linguistic phenomena, an observer can obtain a better understanding of what has changed from one version of the grammar to the next, to help in explaining more readily observed changes in coverage of the grammar when applied to a corpus. For a more complete understanding of the state of the grammar, the implementation should be accompanied by a rich inventory of linguistic phenomena that the grammarian aspires to analyze, perhaps anchored in a comprehensive pencil-and-paper grammar such as (for English) the Cambridge Grammar of the English Language (Huddleston and Pullum, 2002). Documenting this inventory and the mapping from the English Resource Grammar to such a resource should be a priority for further work on this approach.

Another clearly desirable improvement over the methods described here would be to enable searches of the treebank for specific phenomena without requiring explicit and sometimes tedious mention of each rule involved by name, though this would require the definition of a nontrivial mapping between an inventory of linguistic phenomena at varying levels of abstraction, and the specific constructs defined in the grammar. A small beginning in this direction can be found in the Redwoods Tenth Growth with analyses of almost all of the example sentences cited by (Huddleston and Pullum, 2002) in their chapter 12 on relative clauses, a resource also used by (Letcher, 2018) in a distinct and promising approach to phenomenon discovery.

## Acknowledgments

Part of this work was conducted during a fellowship at the Oslo Center for Advanced Study at the Norwegian Academy of Science and Letters, and I am grateful for the excellent research environment and support that CAS provided.

## References

Bender, E. M., Ghodke, S., Baldwin, T., and Dridan, R. (2012). From database to treebank: On enhancing Hypertext Grammars with grammar engineering and treebank search. *Language Documentation & Conservation Special Publication No. 4 Electronic Grammaticography*, pages 179–206.

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

Flickinger, D. (2000). On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(01):15–28.

Flickinger, D. (2011). Accuracy v. Robustness in grammar engineering. In Bender, E. M. and Arnold, J. E., editors, *Language from a Cognitive Perspective: Grammar, Usage and Processing*, pages 31–50. CSLI Publications, Stanford, CA, USA.

Huddleston, R. and Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

Letcher, N. (2018). *Discovering syntactic phenomena with and within precision grammars*. PhD thesis, University of Melbourne.

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

Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. The University of Chicago Press, Chicago, IL, USA.

Rosén, V., Meurer, P., and de Smedt, K. (2005). Constructing a parsed corpus with a large LFG grammar. In Butt, M. and King, T. H., editors, *Proceedings of the LFG '05 Conference, University of Bergen*, Stanford, CA, USA. CSLI Publications.

Rosén, V., Thunes, M., Haugereid, P., Losnegaard, G. S., Dyvik, H. J. J., Samdal, G. I. L., Meurer, P., and de Smedt, K. (2016). The enrichment of lexical resources through incremental parsebanking. *Language Resources and Evaluation*, 50(2):291–319.