

Comparing test-suite based evaluation and corpus-based evaluation of a wide-coverage grammar for English

Rashmi Prasad and Anoop Sarkar

Institute for Research in Cognitive Science
University of Pennsylvania
3401 Walnut St., Ste 400A,
Philadelphia, PA 19104 USA
{rjprasad,anoop}@linc.cis.upenn.edu

Abstract

In this paper we present our experiences in the evaluation of a wide-coverage grammar of English: the XTAG English grammar. We give a brief history of previous evaluations done using the XTAG grammar and then describe a pair of new evaluations done on a corpus of weather reports and the CSLI LKB test suite. Based on these experiments, we discuss the differing merits of naturally occurring corpora and test suites in the evaluation of wide-coverage grammars.

1. Introduction

The term *wide-coverage* when applied to natural language parsing has two meanings: one is that of grammatical coverage: how many linguistic phenomena does the grammar handle; the second is that of parsing coverage: for how many sentences from a naturally occurring corpora of text does a parser produce an appropriate derivation. Evaluation methods such as test-suite based evaluation handle the former notion of coverage, while corpus-based metrics (Harrison et al., 1991) attempt to handle the latter notion of coverage.

In this paper, we report on our experience in conducting an evaluation along these two dimensions on XTAG: a wide-coverage grammar of English. We give a brief history of previous evaluations and describe some new evaluations of the grammar and parsing system. We compare the changes proposed to the XTAG English grammar as a result of these evaluations. Based on this comparison we argue for a method for evaluating the coverage of a grammar/parser that tries to combine both of these notions as an attempt to reach an evaluation method that is better suited to improve grammar development for

wide-coverage grammars.

2. The XTAG English grammar

The XTAG English grammar (XTAG-Group, 1998) is a wide-coverage English grammar based on the lexicalized Tree Adjoining Grammar (LTAG) formalism. LTAG is a lexicalized mildly-context sensitive tree rewriting system (Joshi et al., 1975). It is capable of supporting structural descriptions not supported by context-free grammars. In LTAG, each piece of syntactic structure is encoded in an object called an *elementary tree*, which is lexicalized. Parsing is accomplished by combining these elementary trees selected by the words in the input sentence. Currently the XTAG parser has access to approximately 1.8 million lexicalized trees while parsing, with an average of 44.5 trees per word. Only a small subset of these trees is picked out by the words in a given sentence. Lexicalization is determined using on-line lexical resources and default rules based on morphological analysis. Grammar development is done at a more general level than individual lexicalized trees. The grammar is organized into several collections of tree templates and tree template families related by their predicate-argument structure. A complete description of the linguistic details in the XTAG grammar is available on the project web-page: <http://www.cis.upenn.edu/~xtag>.

We would like to thank Aravind Joshi and Fei Xia for their help and suggestions. This work was partially supported by an NSF Grant, SBR 8920230.

3. A Combined Evaluation Metric

A subset of the approaches to evaluating parsing systems can be grouped as being *intrinsic evaluation* methods (Srinivas et al., 1998), as they measure the performance of a parsing system in the context of the framework in which it was developed. This kind of evaluation helps system developers and maintainers to measure the performance of successive generations of a system. In particular, for grammar-based systems such as XTAG, it helps identify the shortcomings and weaknesses in the grammar, and provides a direction for the productive development of the grammar.

Approaches to intrinsic evaluation can be divided into *test suite-based* and *corpus-based* methods. In the test suite-based method, a list of sentences for each syntactic construction that is covered and not covered by the grammar is maintained as a database. This test suite is used to track improvements and verify consistency between successive generations of grammar development in a system. The corpus-based methods are further divided into those which use annotated data and those which use unannotated data. In this paper, we will be concerned only with those methods that use unannotated data. Unannotated corpus-based evaluation methods use unrestricted texts as corpora for evaluating parsing systems.¹ One example of this method is measuring *coverage* which is a measure of the percentage of sentences in the corpus that can be assigned one or more parses by a parsing system (Briscoe and Carroll, 1995).

The advantage of the test suite-based evaluation is that it is relatively straightforward and the information provides a direction for improving the system. However, the disadvantage is that it does not quantify how the performance of a parsing system would scale up when parsing unrestricted text data. For the evaluation of a wide-coverage system, however, the corpus-based evaluation makes up for this disadvantage.

In the rest of the paper, we describe previous and current evaluations of XTAG, both test suite-based and corpus-based, and show that combining the evaluation methods has been very produc-

tive for our grammar development efforts.

3.1. Previous Evaluations

In the evaluation of the grammar presented in (Doran et al., 1994), a subset of the Wall Street Journal and the Brown Corpus was parsed with the grammar and then subjected to a detailed error analysis. The results of the evaluation are shown in Table 1. Based on this evaluation the grammar was updated to handle errors caused due to #1, #2, #3, #7, #12, #13 and #14.

Rank	No.errors	Category of error
#1	11	Parentheticals/appositives
#2	8	Time NP
#3	8	Missing subcat
#4	7	Multi-word construction
#5	6	Ellipsis
#6	6	Not sentences
#7	3	Gapless Relative clause
#8	2	Funny coordination
#9	2	VP coordination
#10	2	Inverted predication
#11	2	Who knows
#12	1	Missing entry
#13	1	Comparative
#14	1	Bare infinitive

Table 1: Results of Corpus-Based (WSJ and Brown) Error Analysis

In addition to the corpus-based evaluation, the sentences (and phrases) of the TSNLP (Test Suites for Natural Language Processing) English Corpus (Lehmann et al., 1996) were also parsed using the XTAG grammar (Doran et al., 1997; Srinivas et al., 1998). The corpus contains 1409 grammatical sentences and phrases and 3036 ungrammatical ones. 61.4% of the grammatical examples (1367) were reported as parsed by the system.² Detailed results of the error analysis are given in Table 2.

We have presented these earlier evaluations of the XTAG grammar in order to emphasize the changes occurring the grammar due to continual evaluations. We now compare these earlier evaluations with a new corpus-based evaluation and test-suite based evaluation in Section 3.2..

¹In unrestricted texts, the sentences are not annotated with any linguistic information.

²42 of the 1409 sentences were judged as ungrammatical and removed from the test corpus.

Error Class	%
POS Tag	19.7%
Missing item in lexicon	43.3%
Missing tree	21.2%
Feature clashes	3%
Rest	12.8%

Table 2: Breakdown of TSNLP Test Suite errors

3.2. Current Evaluation

For the current evaluation of the XTAG grammar, we parsed a corpus of weather reports and the 1348 sentences of the CSLI LKB (Linguistic Knowledge Building) test suite (Copestake, 1999), and compared the results of the two evaluations. The weather reports were provided to us by CoGenTex.³ The sentences in this kind of corpus tend to be quite long (an average of 20 tokens/sentence) and complex. The examples given in the Appendix are illustrative of the type of sentences and terminology in this domain.

3.2.1. The Weather Corpus

(Doran et al., 1997) parsed the corpus of weather reports and found that it contained several relative clauses that caused problems for the XTAG grammar and also that the parser was unable to handle some of the more complex longer sentences. The problematic cases involved two kinds of relative clauses (roughly 40%) which were not accounted for by the grammar developer at the time. The first kind contained examples like *A frontal system approaching from the west* which included an *-ing* form in the relative clause predicate, and the other kind contained examples like *The disturbance south of Nova Scotia early this morning* which had a directional noun phrase as the predicate. As a result of these shortcomings and also due to the long and complex sentences in this test set, we could parse only about 20% of the test set (10 out of 48 sentences). After a recent overhaul of the relative clause analysis in the XTAG grammar we wanted to test the updated coverage of the grammar on the same test set and evaluate the degree of improvement in performance.

³Thanks to the Contrastive Syntax Project, Linguistics Department of the University of Montreal, for the use of their weather synopsis corpus.

The test set of sentences from the weather corpus in the current evaluation is the same as in the previous study (Doran et al., 1997). There are 48 sentences in the test set (the size of the test set was purposefully kept small since the output had to be checked by hand to check parser accuracy).

Before parsing the corpus, we preprocessed the corpus with a noun-phrase chunker (Ramshaw and Marcus, 1995)⁴ and hand-corrected the few errors it made. There were a total of 536 noun phrase chunks and 4% (21/536) of these chunks were incorrect. We then parsed the (corrected) NP-chunked text with the current XTAG parser. Parsing the entire test set took about 175 minutes. The parse times for each sentence in the test set is shown in Figure 1.

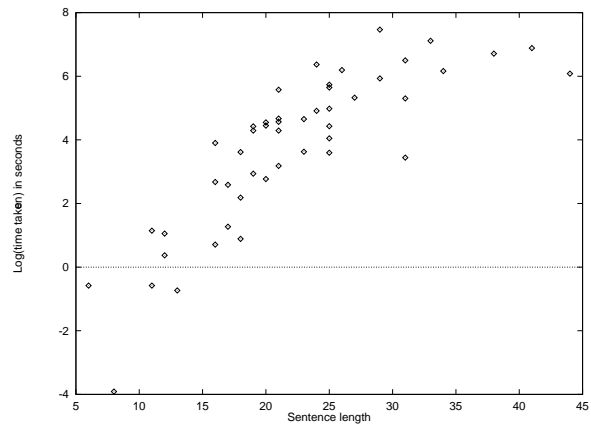


Figure 1: Parse times plotted against sentence length for the test set from the weather reports corpus.

With increased grammar coverage due to previous corpus evaluations we have obtained a much better performance of the XTAG parser on the sentences in the weather domain. Of the 48 sentences in the corpus, we were able to parse 43 sentences (89.6%). A comprehensive breakdown of the errors is described in Table 3. The errors were due to the reasons listed below. In all the examples, the problematic word or phrase appears in bold.

⁴We used Adwait Ratnaparkhi’s part-of-speech tagger (Ratnaparkhi, 1996) to tag the input sentences before sending them to the chunker. The POS tags were not used as input to the parser – all POS ambiguity was retained while parsing.

R1 (1 error): Lexical item does not get the right part of speech and therefore does not select the correct trees. *Example:* [The air] is [mainly dry both east and west] of [the ridge] over [the valley] with [a tendency] **toward** [stronger southeasterly winds] in [milder air] west of [the ridge line]. *toward* is not assigned a “Preposition” POS and therefore does not select the post-NP modifier tree.

R2 (1 error): Lexical item does not select the necessary tree or family. *Example:* [A ridge] of [high pressure] **moving in** from Ontario will give [a sunny day] to [all three provinces] on Saturday. *moving in* is not analyzed as an intransitive verb-particle.

R3 (3 errors): Missing analysis for VP coordination in the current grammar. *Example:* [This disturbance] **will slowly move eastward today and should lie to [the east] of [our regions] on Wednesday.**

Error Class	No.	%
POS Tag (R1)	1	2.08%
Missing Tree (R2)	1	2.08%
No VP coordination (R3)	3	6.25%
Total	5	10.4%

Table 3: Results of Corpus-Based Error Analysis

3.2.2. The CSLI LKB Test Suite

As a test-suite that had not been parsed before, we used the CSLI LKB test suite. The parser took 41.5 minutes to parse the grammatical sentences and around 13 minutes to parse the ungrammatical sentences. The times taken for parsing the grammatical sentences are shown in Figure 2.

In the LKB test suite, of the 966 grammatical sentences we were unable to parse 26 (2.7%). The errors were caused by the following problems (see also Table 4):

E1 (1 error): Missing entry for intransitive *interview*. *Example:* Abrams will **interview** forever.

E2 (3 errors): Missing entry for transitive *evaluated*. *Example:* Chiang evaluated Abrams.

E3 (1 error): Missing tree for negation having scope over determiners. *Example:* **Not many programmers** were interviewed by Browne.

E4 (3 errors): Missing tree for adverbial *then*. *Example:* If Devito hired Browne, **then** the project

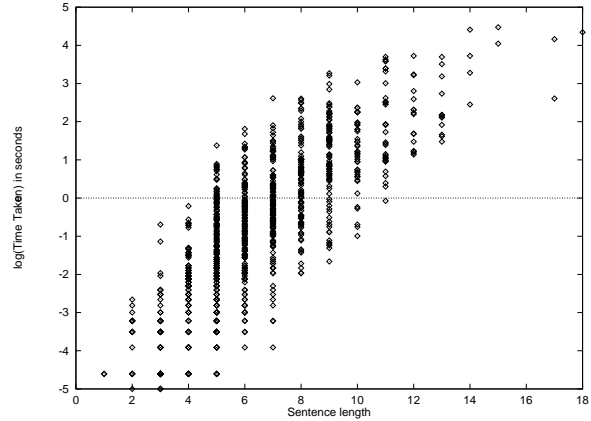


Figure 2: Parse times plotted against sentence length for the LKB test-suite.

had three programmers.

E5 (2 error): Missing analysis for inverse predication. *Example:* **In the office** is the bookcase.

E6 (18 errors): Missing analysis for ellipsis. *Example:* If managers are not, a consultant is interviewing programmers.

E7 (1 error): Proper nouns don’t get the predicative tree by default. *Example:* There is Abrams, Browne, and Chiang.

Error Class	No.	%
Missing Entry (E1-E2)	4	0.4%
Missing Lexicalized Tree (E3-E4)	4	0.4%
No Inverse Predications (E5)	2	0.2%
No analysis for Ellipsis (E6)	18	1.8%
Default entry error (E7)	1	0.1%
Total	26	2.7%

Table 4: Results of CSLI Test Suite-based Evaluation

Errors R3, E5, and E6 were noticed before (Srinivas et al., 1998) as lacking in our grammar, but the analysis for the three kinds of constructions (VP coordination, ellipsis, inverse predication) has not been added yet. Of the 387 ungrammatical sentences, 189 did not get any valid parses after feature unification. We did not examine the parses obtained for the remaining sentences to see if they contained a legitimate ambiguity or an error in the grammar.

3.2.3. Comparison of Test-Suite and Corpus-Based Evaluation Methods

Looking at the history of evaluations presented in this paper we saw that the initial evaluations of the XTAG English grammar on corpora like the Brown corpus and the Wall Street Journal corpus identified several instances of linguistic phenomena which were missing from the grammar. A subsequent analysis of the TSNLP test-suite was useful in finding cases where lexical entries did not select the appropriate trees. It is important to note here that there was very little overlap in the kinds of changes identified in the two kinds of evaluation. Compare the errors found after parsing corpora in Table 1 and the errors found after parsing a test-suite in Table 2. We get a similar variation in the evaluations conducted in this paper. Different kinds of errors were detected while parsing the weather corpus (see Table 3) when compared to the errors detected after parsing the CSLI LKB test-suite (see Table 4).

By parsing the weather reports corpus, a domain that was novel to the XTAG grammar, we found that two particular types of reduced relatives (see Section 3.2.1.) were not getting the right analysis. This caused dismal parsing performance on that corpus: only 20% of the test set was parsed. After a recent overhaul of the XTAG relative clause analysis, we re-parsed the same test set with greater success, and found that now we were able to parse 89.6% of the corpus. Parsing the weather corpus has thus enabled us to recognize the predominance of these particular kinds of reduced relative clauses and the corresponding shortcoming in our grammar with respect to the analysis of these constructions. More importantly, our results indicate that this improvement has been possible only as a result of the corpus-based evaluation.

We believe that using test-suites as the only evaluation metric for a wide-coverage grammar faces the following disadvantages compared to an approach that also includes a corpus-based evaluation. Firstly, test-suites are constructed by hand to allow for systematic variation only within a particular range of grammatical phenomena and, in contrast to the sentences found in naturally-occurring corpora (see above), are unlikely to point us towards an increasing set of novel constructions. Secondly, in a test-suite, the same lex-

ical items are used very often, causing a distribution of words that is unlike naturally occurring text (see (Doran et al., 1997) for further discussions about the use of test-suites for evaluation). Finally, each sentence in a test-suite usually handles a single grammatical phenomena and, so, interactions between different phenomena are seldom explored. Corpus sentences, on the other hand, typically contain several grammatical phenomena within them, and are thus more likely to reflect the real-world complexity in parsing using the wide-coverage grammar. Comparing Figure 1 and Figure 2 we can see that parsing the LKB test-suite took far less time (41.5 minutes) when compared to time taken to parse the weather corpus (175 minutes) even though the test-suite was roughly 20 times larger.

Test-suites, however, are useful to maintain the consistency of a wide-coverage grammar. When new additions are made to the grammar, a test-suite can systematically check that earlier analyses have not broken as a result of the additions. However, for this purpose it is better to have a test-suite tailored to a particular grammar, and to this end, the XTAG grammar has an internal test-suite that contains all the example sentences (grammatical and ungrammatical) from the continually updated documentation of the grammar. This internal test-suite is used as an internal check so that new additions to the grammar are not inconsistent. Test-suites can also be used as a coarse metric to find which grammatical phenomena are missing from a particular wide-coverage grammar, thus enabling comparison with other wide-coverage grammars. Finally, test-suites are also useful for accounting for certain rare phenomena that do not occur commonly in corpora. For example, certain examples of ellipsis such as *If managers are not, a consultant is interviewing programmers* may never be detected using a corpus-based evaluation.

Based on the experiences of evaluations detailed above, we argue for a combined methodology for the evaluation of wide-coverage grammars that includes test-suite based evaluation and corpus-based parsing coverage-based evaluation. As shown in this paper, this approach has served us well in the evaluation and the continuing development of the XTAG English grammar.

4. Conclusion

In this paper we argued for a combined methodology for the evaluation of wide-coverage grammars that includes test-suite based evaluation and corpus-based parsing coverage-based evaluation. We presented our experiences in evaluating the wide-coverage XTAG English grammar using these methods. We analyzed the output of parsing a corpus of weather sentences as well as parsing the CSLI LKB test-suite and compared the benefits of conducting both of these experiments to the grammar development process. Based on these results, we concluded that both corpus-based and test-suite based evaluations are important for grammar development in the XTAG English grammar.

5. References

- T. Briscoe and J. Carroll. 1995. Developing and Evaluating a Probabilistic LR parser of Part-of-Speech and Punctuation Labels. In *Proc. of the Fourth International Workshop on Parsing Technologies*, Prague, Czech Republic.
- A. Copestake, 1999. *The (new) LKB System*. CSLI, Stanford University.
- C. Doran, D. Egedi, B. A. Hockey, B. Srinivas, and M. Zaidel. 1994. XTAG System - A Wide Coverage Grammar for English. In *Proc. of the 17th International Conference on Computational Linguistics (COLING '94)*, Kyoto, Japan, August.
- C. Doran, B. Hockey, P. Hopely, J. Rosenzweig, A. Sarkar, B. Srinivas, F. Xia, A. Nasr, and O. Rambow. 1997. Maintaining the Forest and Burning out the Underbrush in XTAG. In *Workshop on Computational Environments for Grammar Development and Language Engineering (ENVGRAM)*, Madrid, Spain, July.
- P. Harrison, S. Abney, D. Flickinger, C. Gdaniec, R. Grishman, D. Hindle, B. Ingria, M. Marcus, B. Santorini, and T. Strzalkowski. 1991. Evaluating syntax performance of parser/grammars of english. In *Proc. of the Workshop on Evaluating Natural Language Processing Systems, ACL, 1991*.
- A. J. Joshi, L. S. Levy, and M. Takahashi. 1975. Tree adjunct grammars. In *Journal of Computer and System Sciences*.
- S. Lehmann, S. Oepen, S. Regnier-Prost, K. Netter, V. Lux, J. Klein, K. Falkedal, F. Fouvry, D. Estival, E. Dauphin, H. Compagnion, J. Baur, L. Balkan, and D. Arnold. 1996. TSNLP — Test Suites for Natural Language Processing. In *Proc. of COLING 1996*, Copenhagen.
- L. A. Ramshaw and M. P. Marcus. 1995. Text chunking using transformation-based learning. In *ACL 3rd Workshop on Very Large Corpora*, pages 82–94.
- A. Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proc. of EMNLP-96*, May 17-18.
- The XTAG-Group. 1998. A Lexicalized Tree Adjoining Grammar for English. Technical Report IRCS 98-18, University of Pennsylvania.
- B. Srinivas, A. Sarkar, C. Doran, and B. A. Hockey. 1998. Grammar and Parser Evaluation in the XTAG Project. In *Workshop on Evaluation of Parsing Systems*, Granada, Spain, 26 May.

Appendix

1. [The warm air mass] affecting [southwestern Quebec] for [the past few days] is still moving slowly eastwards and will make room for [a much colder air mass] overnight and Wednesday.
2. [A low] lying near [Nova Scotia] [this morning] is inducing [a northeast flow] over [the St Lawrence].
3. [This disturbance] will slowly move eastward today and should lie to [the east] of [our regions] on Wednesday.
4. Cloud will move into [the western regions] in [the wake] of [the high pressure area] [Saturday afternoon].
5. Behind [this area] [a moderate flow] will cause [an inflow] of [milder air] in [southwestern Quebec] producing temperatures slightly above normal on Sunday.
6. West of [the ridge line] [a gradual increase] in cloudiness is being experienced today and as [the cloud] thickens [this evening] showers will commence and continue in [the western half] of Quebec tomorrow.
7. [The disturbance] south of [Nova Scotia] early [this morning] will slowly pull away but nevertheless give [that province] [cloudy skies] and [isolated shower activity] today.
8. [The cloud] will gradually clear as [the system] pulls away from [the district] although [Cape Breton] should receive [a few showers] from it.
9. [Very mild air] preceding [a disturbance] in [north Dakota] will gradually push into part of [St Lawrence River valley] today and Monday.
10. [A mass] of [cool air] is presently pushing across [the Great Lakes] behind [the disturbance] affecting Quebec today.