Practical Experience with Grammar Sharing in Multilingual NLP

Michael Gamon, Carmen Lozano, Jessie Pinkham, Tom Reutter
Microsoft Research
One Microsoft Way
Redmond WA 98052
USA
{mgamon,clozano,jessiep,treutter}@microsoft.com

Abstract

In the Microsoft Natural Language Processing System (MSNLP), grammar sharing between English, French, Spanish, and German has been an important means for speeding up the development time for the latter grammars.

Despite significant typological differences between these languages, a mature English grammar was taken as the starting point for each of the other three grammars. In each case, through a combination of adding and deleting a modest number of grammar rules, and modifying the conditions on many others, a broad-coverage target grammar emerged.

Tests indicate that this approach has been successful in achieving a high degree of coverage in a relatively short period of time.[1]

1 Grammar Sharing

A broad-coverage multilingual NLP system such as the one currently being developed at Microsoft Research faces the challenge of parallel grammar development in multiple languages (currently English, French, Spanish, German, Chinese, Japanese, and Korean). This development is by nature a very complex and time-consuming task. In addition, the design of the overall NLP system has to be well suited to be readily portable to languages other than the one the development started with (English in our case). For these reasons, few groups have succeeded at the challenge of multilingual NLP.

One approach to multilingual development is to rely on theoretical concepts such as Universal Grammar. The goal is to create a grammar that can easily be parameterized to handle many languages. Wu (1994) describes an effort aimed at accounting for word order variation, but his focus is on the demonstration of a theoretical concept. Kameyama (1988) describes a prototype shared grammar for the syntax of simple nominal expressions for five languages, but the focus of the effort is only on the noun phrase, which makes the approach not applicable to a large-scale effort. Principle-based parsers are also designed with universal grammar in mind (Lin 1994), but have yet to demonstrate large-scale coverage in several languages. Other efforts have been presented in the literature, with a focus on generation (Bateman et al. 1991.) An effort to port a grammar of English to French and Spanish is also underway at SRI (Rayner et al. 1996.)

The approach taken in the MSNLP project focused from the beginning on possibilities for grammar sharing between languages to facilitate grammar development and reduce the development time. We want to stress that our use of the term "grammar sharing" is not to be confused with "code sharing." Grammar sharing, in our use of the term, simply means that the existing grammar for one language can be used totally or in part to serve as the development basis for a second language.

In this paper we want to demonstrate that the jumpstart through grammar sharing considerably accelerated grammar development in French, Spanish, and German. We will present test and progress data from all languages to support our claim.

2 The Microsoft NLP System

The English grammar that we used as our starting point, as well as the target-language grammars that were spawned from it, are *sketch* grammars. Sketch grammars use a computational dictionary

[1] This work has benefited from comments and suggestions from other members of the Natural Language Processing group at Microsoft Research. Particular thanks go to Simon Corston, Bill Dolan, Ken Felder, Karen Jensen, Martine Pettenaro, Hisami Suzuki, and Lucy Vanderwende.

containing part-of-speech, morphological, and subcategorization information to yield an initial syntactic analysis (the *sketch*). The rules used in *sketch* have no access to any semantic information that would allow the assignment of semantic structure such as case frames or thematic roles.

Further analysis proceeds through a stage of reattachment of phrases using both semantic and syntactic information to produce the *portrait*, then to a first representation of some aspects of meaning, the *logical form*, and to word sense-disambiguation and higher representations of meaning. In this paper, however, we will restrict our attention to the sketch grammars.

A bottom-up parallel parsing algorithm is applied to the sketch grammar rules, resulting in one or more analyses for each input string, and defaulting in cases (such as PP attachment) where semantic information is needed at a later stage of the processing (*portrait*) to give the correct result. Context-sensitive binary rules are used because they have been found necessary for the successful analysis of natural languages (Jensen et al. 1993, pp. 33-35; Jensen 1987, pp. 65-86).[2] Figure 1 gives a template for the rule formalism for a binary rule, in this case a rule that combines a verb phrase with a prepositional phrase to its right.

Each sentence parse contains syntactic and functional role information. This information is carried through the system in the form of arbitrarily complex attribute-value pairs. The *sketch* always produces at least one constituent analysis, even for syntactically invalid input, and displays its analyses as parse trees. FITTED parses are obtained when an input string cannot be parsed up to a sentence node (possibly because it is a noun phrase, a sentence fragment, or otherwise deficient). FITTED parses contain as much constituent structure as the grammar could assign to the input string.

```
VPwPPr:
VP ( Condition 1 &
     Condition 2 & ..... )
PP ( Condition 1 &
     Condition 2 & ..... )
--> VP {action 1;
        action 2; ....}
```

Figure 1. Outline of the binary rule combining a VP with a PP to its right (VPwPPr)

Two types of trees are available (Figure 2). One strictly follows the derivational history of the parse, and is therefore binary-branching. In the binary tree the names of the rules that have produced a node are displayed to the right of that node. The second (which is used in later processing because it accords better with our intuitive understanding of many structures) is n-ary branching, or "flattened," and is computed from a small set of syntactic attributes of the analysis record. The * indicates the head of the phrase.

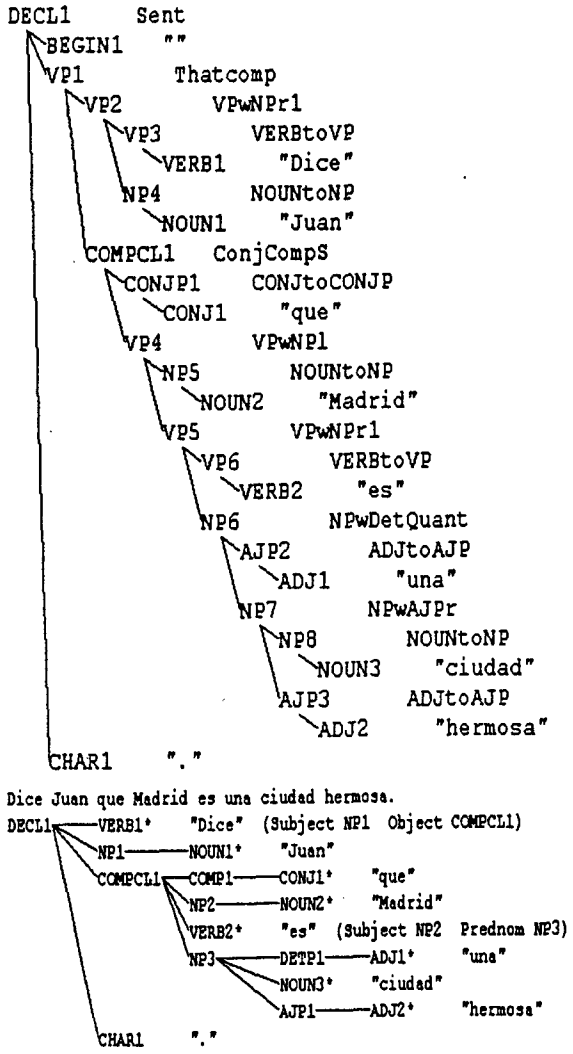


Figure 2: A derivational tree and a "flattened" tree for the sentence "Dice Juan que Madrid es una ciudad hermosa" ("John says that Madrid is a beautiful city")

The sketch grammar is written in G, a Microsoft-internal programming language that has been specially designed for NLP. The English

[2] Binary rules deal with the problem of free constituent order, which is significant even in a largely configurational language such as English. A case of free word order in English is the position of adverbials and prepositional phrases.

grammar contains 129 rules, which vary in size from two to 600 lines of code, according to the number of conditions that they contain. The coverage of English is broad. Processing time is rapid: a typical 20-word sentence requires about an eighth of a second on a Pentium 200 MHz machine.

The goal of all Natural Language Research and Development at Microsoft Research is to produce a broad coverage multilingual NLP system that is not tailored to any specific application, but has the potential to be used by any of them. To date, the English system is the foundation of the grammar checker in Word 97. We expect our multilingual technology to be used in as wide a spectrum of applications as possible.

3 The Development of the French, Spanish and German Grammars

In this section we briefly explain the common strategy of grammatical development in the MSNLP system and we give the current status of development for each of these three languages.

For each of the three languages under consideration, the development team consists of a lexicographer/morphologist, and a grammarian. Grammar work in each language proceeds according to the same rationale: the grammarian processes sentences from diverse text sources and examines the resulting parses. He/she then determines whether the resulting parse is a desirable one. If this is the case, the sentence with the correct parse is added to a regression file. If the parse is incorrect, conditions on grammar rules are modified or added to accommodate the sentence in question and similar constructions. Regression tests are run frequently to ensure that new changes do not affect the performance of the system in any negative way. A debugging tool is available for the linguist to immediately view differences that arise in the processing of the regression file compared to an earlier run. Another important tool enables the grammarian to identify conditions in grammar rules that have been tried during a particular parse, and distinguish those that succeeded from those that failed.

3.1 French

Development of the French grammar started in 1995. French grammar work has covered most major constructs including:

- clitic pronouns
- attachment of adjectival modifiers to the right of the nominal head in NPs
- the more liberal use of infinitival complements in French than in English
- questions and other subject inversion constructions
- compound noun constructions
- floating quantifiers and negatives

The French dictionary currently consists of 68,000 words. Morphology is nearly complete, with 98.13% word recognition on a 276,000 word corpus.

3.2 Spanish

Development of the Spanish grammar began in November 1995. The initial focus of grammar work in Spanish was on the following areas:

- preverbal and postverbal clitics
- sentences with no overt subjects
- varying word order of subject noun phrases
- dislocated object noun phrases
- infinitival complements introduced by prepositions
- finite complement clauses introduced by prepositions
- handling of noun phrases that function as adverbs
- homography issues

The Spanish dictionary has 94,000 words. Morphology is almost complete with 98% word recognition on a 300,000 word corpus.

3.3 German

German grammar development started in October 1996. The focus of the grammar work in German has been on:

- verb-final and verb-second word-order
- the relative freedom of constituent-order compared to English
- VP-coordination
- agreement in noun phrases (weak and strong inflection)
- separable prefixes
- homography issues

The German dictionary has over 140,000 entries. The morphology, which includes word-breaking, is nearly complete, with 97% word recognition on a 400,000 word corpus.

Because Spanish and German share the fundamental property of freer constituent order than

English, German grammar has benefited from some of the solutions for this challenge already worked out for Spanish. Grammar sharing between Spanish and German focused mainly on adoption of Spanish code from the binary rules that combine verbs and preceding/following noun phrases.

3.4 Changes from the English Grammar to the Target Grammars

In spite of the numerous areas of divergence between the target grammars and English, we found that the fundamental organization of the grammar changed as little as 10-19% (see specifics in Table 1). The bulk of the required modifications occurred in the conditions on the rules. Since these conditions are complex, it is difficult to illustrate them fully here. To give one simple example, in French and Spanish, we found it necessary to exclude all NPs that consisted of clitic pronouns from rules that attach modifiers on NPs.

Few rules had to be added or completely removed from the grammar. For example, bootstrapping the Spanish grammar from an English grammar consisting of 129 rules required that only 13 of the original English rules (10.1%) be deleted, while 10 new rules (7.8%) were introduced.

Language	% Deleted	% Added
Spanish	10.1	7.8
German	10.7	8.6
French	7.8	2.3

Table 1: percentages of deleted/added rules with respect to the English source grammar.

The new rules were added to accommodate constructions in the target language that are (virtually) non-existent in English. Spanish, for example, added rules to handle nominalized prepositional phrases like *el de Juan* and nominal uses of infinitives: *al verlo.* French needed rules to handle present participles introduced by *en*: *en partant*, and for sentential constructions like *Heureusement qu'il est venu!* German added rules for constructions such as postposed genitive NPs (*das Buch Peters*) and participial VPs premodifying NPs: *die dem Mann gegebenen Bücher.*

4 Testing and Progress Measurement

Testing NLP systems is known to be a difficult task, and one that is hard to standardize across systems with different aims and different grammatical foundations (see e.g. the discussion in Balkan et al. n.d.). One relatively simple measurement that we found particularly useful for the beginning stages of grammar development is the percentage of non-FITTED parses on a corpus containing sentences from different types of text (news, literature, technical writing etc.).

In what follows, this corpus for each language will be referred to as a *benchmark corpus* and *coverage* refers to the percentage of non-FITTED parses for the benchmark corpus. *Sentence length* refers to the number of words in the sentence. In testing, the linguist does not examine the output parses obtained from the benchmark corpus, in order to avoid targeting modifications of the grammar towards the particular problems with FITTED parses in the benchmark file. This "blind" test yields a rough measure of the real coverage of the grammar. It should be noted that although some non-FITTED parses may not constitute the desired parse, many FITTED parses yield a largely usable parse which has only failed at the sentence level.[3] But more important at this point is the fact that our measurement against a benchmark allows us to reliably track progress over time.

Even though not all of the successfully parsed sentences are guaranteed to have received a desired parse, a stable increase in the percentage of parsed sentences during language-specific grammar work has proven to be a reliable measurement of progress. This is particularly true given that grammar work (as described above) proceeds on the basis of example sentences that come to a large extent from real-life text.

A factor that influences the coverage considerably is sentence length. In order to assess the relationship between sentence length and the percentage of parsed sentences in a corpus, we use a tool that extracts information from a parsed corpus on the ratio of successfully parsed sentences to FITTED parses depending on sentence length.

[3] Additional testing of considerable magnitude would be required to evaluate "perfect correctness". This would take us away from development, and provide slower feedback of progress.

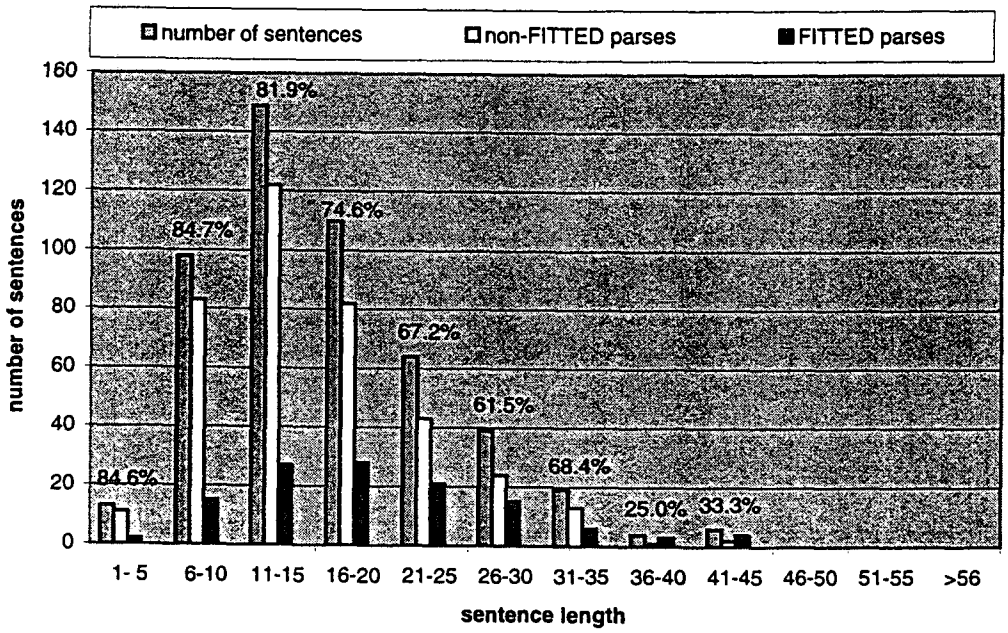
5 Results

5.1 French

The corpus gathered for French is 500 sentences long, and 16.9 words average in length. It covers a number of different text types (news, letters, online discussions, legal, literature, etc.) (see Pinkham 96 for details)[4]. The text is used 'as is' from the Web, with only a few spelling corrections. Coverage on this corpus approximately a year ago was 54%; today it is 75%.[5]

Development work for French up until now has been biased toward sentences under 20 words. On the basis of the data collected from the experiment below, we can also deduce that effort spent on sentences in the 20+ word range would produce the quickest improvement overall in the future. Figure 3 shows coverage across different sentence length intervals for French. The coverage (i.e. the percentage of parses that is non-FITTED) is shown for each category on top of the columns.

Figure 3: The number of non-FITTED versus FITTED parses in relation to sentence length for French (showing percentage of coverage)



[4] This is in contrast to the Test Suites for Natural Language Processing (TSNLP) test suite data (cf. Balkan et al. n.d.), where the grammatical sentences for French are on average 7 words long, and artificially simple in terms of lexical and grammatical complexity. On the TSNLP data, coverage of the French system is 96%.

[5] We estimate that coverage at the very beginning of French development approximately 18 months ago would have been 25% (on the basis of tests done with other text).

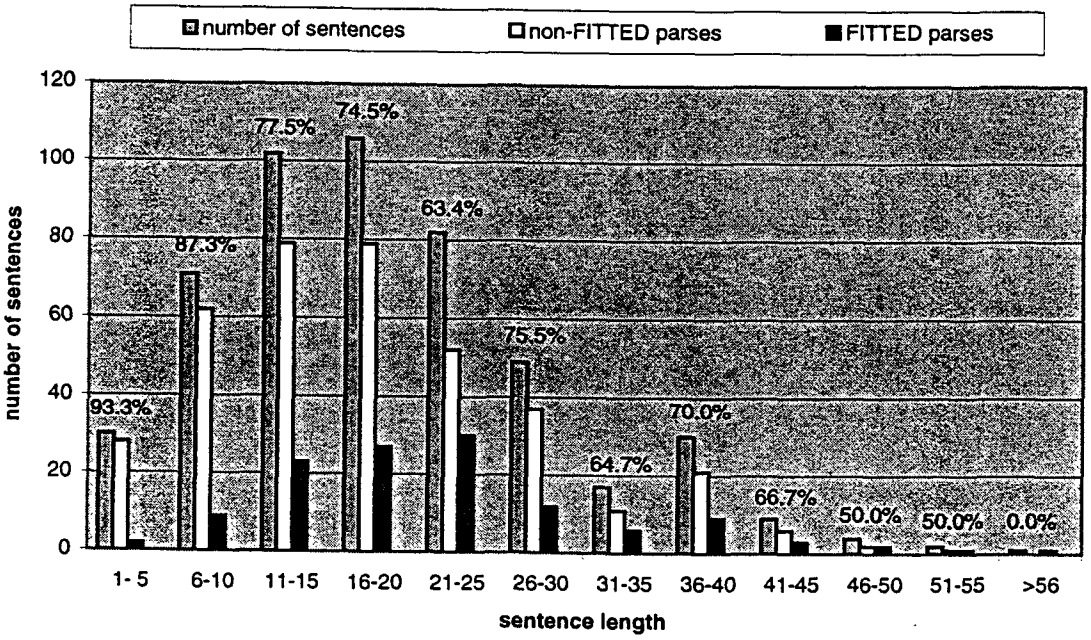
5.2 Spanish

The Spanish benchmark file contains 503 sentences from textbooks, magazines, news articles, a children's book and literary writing (novel). To control for regional variation, both Latin American and Castilian Spanish are represented (the sources are from Spain, Chile, Argentina, and Mexico). The average sentence length is 19.1 words. Current coverage on the benchmark file is 75.15%.

Because Spanish started grammar development while there was only a small prototype dictionary of about 2000 words, no coverage data were taken at the earlier stages of grammar work.

Figure 4 shows the current status of the Spanish grammar with respect to coverage across different sentence length categories in intervals of 5 words in the benchmark corpus.

Figure 4: The number of non-FITTED versus FITTED parses in relation to sentence length for Spanish (showing percentage of coverage)



5.3 German

The German benchmark corpus currently consists of 424 sentences with an average length of 15.3 words per sentence. The sentences are extracted from news articles, novels, children's books, travel guides, technical writing and interviews.

Figure 5 below illustrates the progress of coverage over time from the first steps in grammar work in October 1996 until February 1997. At that point, the coverage had reached over 56.13%. Note that the increase in coverage over time resembles the facts reported in section 5.1 for French. In November 1996, the size of the benchmark corpus was increased from 229 to 424 sentences. This addition of new sentences from new sources had very little impact on the statistics.

Figure 6 shows statistics on the make-up of the corpus and coverage across different sentence-length categories in intervals of 5 words.

Figure 5: Coverage Progress in German

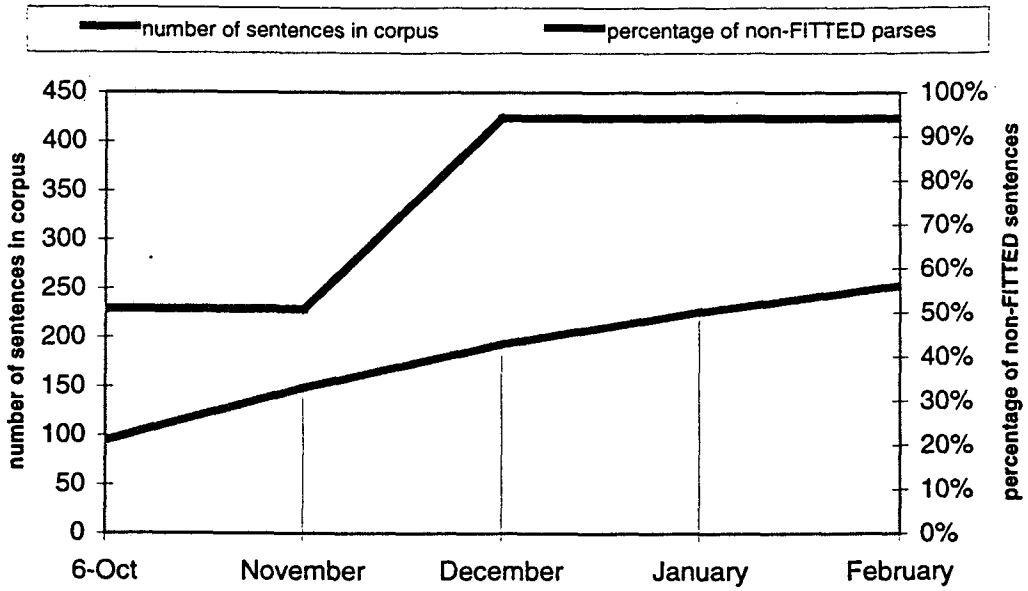
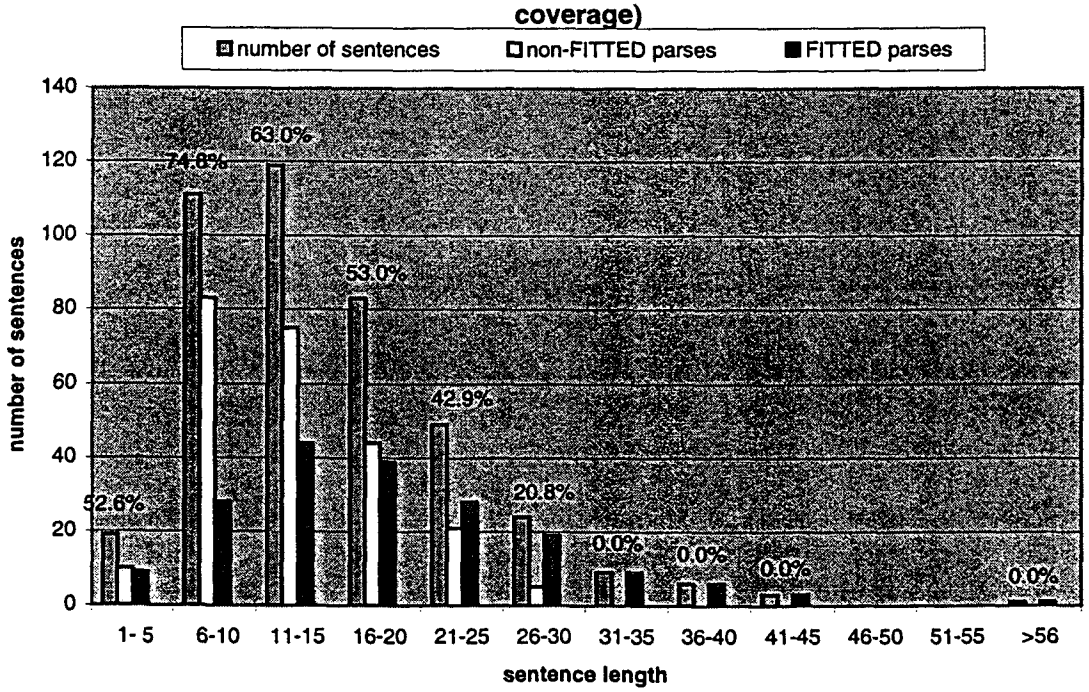


Figure 6: The number of non-FITTED parses versus FITTED parses in relation to sentence length for German (showing percentage of coverage)



6 Conclusion

The results presented here further corroborate the conclusion drawn in Pinkham 1996 that the architecture of the MSNLP system lends itself particularly well to multilingual development and grammar sharing for related languages. Of special importance are the binary rule formalism (Jensen et al. 1993) and the linguistic development tools provided by the system.

While we recognize that grammar development proceeds rapidly in the early stages and slows down with increasing coverage, we have shown that the time frame for full-scale grammars can be much shorter than the 4 years reported in Cole et al. (1997), if the system is designed in the appropriate fashion.

By keeping track of progress in a quick informal fashion, we also gather information on the time-frames required for all future shared grammar development.[6]

References

Lorna Balkan, Frederik Fouvry, Sylvie Regnier-Prost. n.d.. *Test Suites for Natural Language Processing, User Manual, Volume 1.*

John A. Bateman, Christian M.I.M. Matthiessen, Keizo Nanri, and Licheng Zeng. 1991. The re-use of linguistic resources across language in multilingual generation components. *In Proceedings of the 1991 International Joint Conference on Artificial Intelligence*, Sydney, Australia, volume 2, pages 966 - 971. Morgan Kaufmann Publishers.

Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen and Victor Zue, eds. 1997. *Survey of the State of the Art of Human Language Technology.* http://www.cse.ogi.edu/CSLU/HLTsurvey.

Karen Jensen. 1987. Binary rules and non-binary trees: Breaking down the concept of phrase structure. In *Mathematics of language*, ed. Alexis Manaster-Ramer, pages 65-86. Amsterdam: John Benjamins Publishing Company.

Karen Jensen, George Heidorn, Steve Richardson, eds. 1993. *Natural Language Processing: The PLNLP Approach,* Kluwer Academic Publishers.

Megumi Kameyama. 1988. Atomization in Grammar Sharing. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, Buffalo, New York, June 7-10, 1988.

Dekang Lin. 1994. PRINCIPAR - An Efficient, Broad-coverage, Principle-based Parser. In *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, Japan, August 5-9, 1994.

Jessie Pinkham. 1996. Grammar Sharing between English and French. In *Proceedings of the NLP-IA conference*, Moncton, Canada, June 4-6, 1996.

Manny Rayner, Pierrette Bouillon. 1996. Adapting the Core Language Engine to French and Spanish. In *Proceedings of the NLP-IA conference*, Moncton, Canada, June 4-6, 1996.

Andi Wu. 1994. *The Spell-Out Parameters: a minimalist approach to syntax.* Doctoral dissertation, University of California, Los Angeles.

[6] The grammars for Korean, Japanese and Chinese are starting from the ground up, using the same binary rule strategy, but without the benefit of bootstrapping from an existing grammar. This is inevitable since they are typologically too different from the European languages profiled here.