A Bilingual Treebank for the FraCaS Test Suite

Peter Ljunglöf and Magdalena Siverbo

Department of Computer Science and Engineering
University of Gothenburg and Chalmers University of Technology
peter.ljunglof@gu.se

Abstract

We have created an open-source bilingual treebank for 99% of the sentences in the FraCaS test suite (Cooper et al., 1996). The treebank was built in conjunction with associated English and Swedish lexica written in the Grammatical Framework Resource Grammar (Ranta, 2009). The original FraCaS sentences are English, and we have tested the multilinguality of the Resource Grammar by analysing the grammaticality and naturalness of the Swedish translations. 86% of the sentences are grammatically and semantically correct and sound natural. About 10% can probably be fixed by adding new lexical items or grammatical rules, and only a small amount are considered to be difficult to cure.

1. The FraCaS corpus

The FraCaS textual inference problem set (Cooper et al., 1996) was built in the mid 1990's by the FraCaS project, a large collaboration aimed at developing resources and theories for computational semantics. The test set was later modified and converted to a corpus in XML format,[1] and it is this modified version that has been used in this project. The corpus consists of 346 problems each containing one or more statements and one yes/no-question. The total number of unique sentences in the corpus is 874.

The FraCaS problems are divided into 9 broad categories which cover many aspects of semantic inference. The categories are called *quantifiers*, *plurals*, *anaphora*, *ellipsis*, *adjectives*, *comparatives*, *temporal reference*, *verbs*, and *attitudes*, and they are also sub-categorised and sub-sub-categorised in an hierarchy of semantic phenomena. Each problem starts with one or more premises, and a question that can be answered with *yes*, *no* or *unknown*. Here is an example from the *ellipsis* category, with two different answers depending on whether the pronoun "*one*" refers to the "*red car*" or just the "*car*":

P: John owns a red car.
P: Bill owns a fast one.
Q: Does Bill own a fast red car?
A: Yes / Unknown.

2. Grammatical Framework

Grammatical Framework (GF) (Ranta, 2011) is a grammar formalism based on type theory. The main feature is the separation of abstract and concrete syntax. The abstract syntax of a grammar defines a set of abstract syntactic structures, called abstract terms or trees; and the concrete syntax defines a relation between abstract structures and concrete structures. The concrete syntax is expressive enough to describe language-specific linguistic features such as word order, gender and case inflection, and discontinuous phrases.

GF has a rich module system where the abstract syntax of one grammar can be used as a concrete syntax of another grammar. This makes it possible to implement grammar resources to be used in several different application domains. These points are exploited in the GF Resource Grammar Library (Ranta, 2009), which is a multilingual GF grammar with a common abstract syntax for 25 languages, including Finnish, Persian, Japanese and Urdu. The main purpose of the Grammar Library is as a resource for writing domain-specific grammars.

[1] http://www-nlp.stanford.edu/∼wcmac/downloads/fracas.xml

3. The English Grammar and Treebank

To be able to construct a GF treebank we need a grammar and a lexicon that can describe every sentence in the corpus. We have used the GF Resource Grammar as underlying grammar, and added lexical items that capture the FraCaS domain. On top of the resource grammar we have added a few new grammatical constructions, as well as functions for handling elliptic phrases.

In total, we used 107 grammatical functions out of the 189 that are defined in the resource grammar. In addition we added four new grammatical constructions that were lacking, and grammar rules for different elliptic phrases.

The lexicon has in total 531 entries, divided into 63 adjectives, 77 adverbials, 20 conjunctions/subjunctions, 34 determiners, 142 nouns, 19 numerals, 40 proper nouns, 15 prepositions, 12 pronouns, and 109 verbs.

3.1 Additions to the grammar

Four different grammatical constructions were added to the grammar. They consist of natural extensions to and slight modifications of existing grammar rules. An example of a grammar extension is the idiom "*so do I*" / "*so did she*".

The resource grammar cannot handle all kinds of conjunctions and elliptical phrases. In the FraCaS corpus there are 35 sentences with more advanced elliptical constructions. Examples include "*Bill did [...] too*", and "*Smith saw Jones sign the contract and [...] his secretary make a copy*". Our solution was to introduce elliptic phrases in the grammar, one for each grammatical category. E.g., the first example contains an elliptic verb phrase, and the second an elliptic ditransitive verb. To reduce ambiguity, each elliptic phrase is explicitly linearized into the string "*[...]*".

3.2 Coverage

Of the 874 unique sentences, 812 could be parsed directly with the Resource Grammar and the implemented lexicon,

	Total	% of sentences
Unique sentences	874	100%
Accepted by the RG	812	92.9%
- with grammar extensions	826	94.5%
- with elliptic phrases	860	98.4%
- with minor reformulation	866	99.1%
Unable to parse	8	0.9%

Table 1: Coverage of the English FraCaS grammar

No. parse trees	No. sentences	
1 – 9	598	69.1%
10 – 99	203	23.4%
100 – 999	49	5.7%
$\geq$ 1000	16	1.8%

Table 2: Ambiguity of the FraCaS treebank

as shown in table 1. With the three additional grammatical constructions 14 more sentences were parsed. The addition of elliptical phrases increased the number of sentences by another 34. Of the 14 remaining sentences, we could parse 6 more by doing some minor reformulations, such as moving a comma or adding a preposition.

All trees in the FraCaS treebank are implemented in the GF grammar described above. This grammar can be used by itself for parsing and analysing similar sentences. We parsed the 866 sentences covered by the grammar and counted the number of trees for each sentence. Table 2 shows that the grammar is moderately ambiguous, where almost 70% of the sentences have less than 10 different parse trees, and over 90% have less than 100 trees. The median is for a sentence to have 5 parse trees, and the largest number of trees for a sentence is 33,048.

Note that the number of parse trees are misleading for the 34 sentences with elliptic phrases, since ellipsis is linearised as "*[...]*" in the FraCaS grammar. If we had made the elliptic phrases invisible, the number of parse trees would increase dramatically.

4. The Swedish Corpus

As a first step towards making the treebank multilingual, we created Swedish translations of the sentences, by writing a new Swedish lexicon. Then we evaluated the translations and iteratively made changes to the trees to make the translations better. Note that since we use exactly the same syntax trees for the Swedish and English sentences, we had to make sure that the original English sentences were not changed when we modified the trees.

This means that we did not translate the English sentences manually, but instead we translated the lexicon and let the Swedish Resource Grammar take care of linearizing the treebank into Swedish. Currently, out of the 866 trees in the treebank, 748 are linearized into grammatically correct and comprehensible Swedish sentences.

4.1 Coverage

Table 3 gives an overview of the coverage of the Swedish lexicon and grammar. Of the 866 unique trees in the treebank, we consider 748 to have good Swedish translations. The remaining 118 sentences had some problems which we divided into five different classes – idioms, agreement, future tense, elliptical phrases, and more difficult errors. Of these 118 problematic Swedish sentences we believe that more than two thirds should be possible to add to the treebank without too much trouble.

	Total	% of sentences
Sentences in treebank	866	100%
Correct Swedish translation	748	86.4%
Problematic sentences	118	13.6%
– idioms	31	3.6%
– agreement	24	2.8%
– future tense	12	1.4%
– elliptical	19	2.2%
– uncomprehensible	32	3.7%

Table 3: Coverage of the Swedish FraCaS grammar

5. Conclusion

The FraCaS treebank was created in 2011 as a small project financed by the Centre for Language Technology (CLT) at the University of Gothenburg. The project used less than three person months to create a treebank for the FraCaS test suite, together with a bilingual GF grammar for the trees. The coverage of the English grammar is 95–99%, depending on whether you include elliptic phrases or not. The Swedish grammar has a coverage of 86%.

The making of this treebank has been a stress test, both for GF and for the resource grammar. The main work in this project has been performed by a person who is an experienced computational linguist, but had never used GF before. This means that the project has been a test of how easy it is to learn and start using GF and its resource grammar. Furthermore, it was a test of the coverage of the existing grammatical constructions in the resource grammar.

The treebank is released under an open-source license, and can be downloaded as a part of the Gothenburg CLT Toolkit.[2] There is also a technical report describing the treebank in more detail (Ljunglöf and Siverbo, 2011).

6. References

Robin Cooper, Dick Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jaspars Jan, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, Steve Pulman, Ted Briscoe, Holger Maier, and Karsten Konrad. 1996. Using the framework. Deliverable D16, FraCaS Project.

Peter Ljunglöf and Magdalena Siverbo. 2011. A bilingual treebank for the FraCaS test suite. CLT project report, University of Gothenburg.

Aarne Ranta. 2009. The GF resource grammar library. *Linguistic Issues in Language Technology*, 2.

Aarne Ranta. 2011. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford.

[2] Available from URL http://www.clt.gu.se/clt-toolkit