Evaluation of the Syntactic Parsing Performed by the ENGCG Parser

Klas Prytz

Department of Linguistics
Uppsala University
Box 513, S-751 20 Uppsala, Sweden
Klas.Prytz@ling.uu.se

1. Introduction

This paper presents an evaluation of the syntactic parsing performed by the ENGlish Constraint Grammar parser (ENGCG). The parser is described in *'Constraint Grammar; A Language-Independent System for Parsing Unrestricted Text'* (Karlsson, Voutilainen, Heikkilä, Anttila, 1995) and a version of it is available on Internet. Although ENGCG is both a morphological and a syntactic parser, this paper deals exclusively with the syntactic component. A small corpus has been compiled from five short texts from different text categories and has been sent to the parser. The returned analysis has been checked and the performance of the parser has then been evaluated.

The following section provides an introductory presentation of the ENGCG parser, the method used in parsing and the representation of the analysis. The third section describes the method used in the evaluation while the fourth presents the result of the evaluation. The fifth section contains some comments about errors found in the analysis, the sixth points to some factors that may affect the performance and the last section consists of some conclusions.

2. English constraint grammar

Constraint Grammar (CG) is a language independent framework for morphological and syntactic parsing of natural language(s) (Karlsson, 1995, p. v). Any language should be possible to parse within this framework provided there is a correct description of that language. ENGCG is an implementation of CG for the parsing of English. Words are assigned possible analyses by tag assigning modules and the output from these modules are then disambiguated by the application of constraining rules. A word may be assigned several morphological readings as well as syntactic analyses. The ambiguous output from the tag assignment may look as follows:

```
"<tended>"
    "tend" <SV> <SVO>  V PAST VFIN @+FMAINV
    "tend" <SV> <SVO>  PCP2  @APP @-FMAINV
```

The first line contains the word form found in the text. The following lines contain two different morphological analyses. The word surrounded by quotation marks is the lexical form of the word. Then follows a set of tags expressing the analyses. The tags surrounded by angular brackets, here, indicate that both readings of the word are either intransitive or monotransitive.

The rest of the morphological and syntactic analyses is represented by the tags that follow. V PAST VFIN indicates that the word is a finite verb in past tense, while PCP2 on the third line indicates that the word is a participle. As far as the morphological analysis is concerned, one line expresses one unambiguous analysis. A single line may, however, contain more than one syntactic analysis. Syntactic analyses are represented by tags preceded by @. @FMAINV indicates that the word is a finite main verb. In the example above, the second reading is ambiguous between an analysis as an apposition and a non-finite main verb.

In order to solve ambiguities Constraint Grammar rules are applied to this analysis. All Constraint Grammar rules have the same basic construction. A rule consists of a target description that singles out a particular analysis, a set of context conditions that has to be satisfied for the rule to apply, and an action that is undertaken if the rule is applied. As a result of this action the current analysis may be discarded or singled out as the correct one and all other analyses discarded. The morphological constraints discard complete lines together with syntactic tags. The remaining syntactic tags are then subject to the application of syntactic constraints. In this way a run through the parser may reduce the number of analyses. Consecutive runs through the parser may reduce this number further since discarding certain analyses may create context that allows the application of other rules. The last reading is always retained. In this way the output from the parser will always contain at least one analysis for each word.

There are two different set of rules used by the parser. The first set consists of 'safe' rules, that are supposed to always yield a correct result, while the second set consists of heuristic rules that may result in erroneous analyses. The user has the option of choosing the application of both sets of rules or the application of only the non-heuristic set.

3. Method

For the evaluation of the syntactic parsing performed by the ENGCG parser a small corpus was compiled consisting of five texts from different categories, all in all 2628 words, representing different genres (see table below).

Category	Tokens	Average sentence length
Learned	560	43.1
Fiction	488	40.7
Biography	514	39.5
Governmental	562	62.4*
Press	504	14.5
TOTAL	2628	32.0

Table 1. Division of text sample

- In order to obtain text pieces of reasonable length the last 355 words in this text sample have been divided into four clauses using semicolon as a delimiter.

The texts were divided into 'chunks' of about three hundred words that were sent to the on-line ENGlish Constraint Grammar parser accessible on Internet (http://www.lingsoft.fi/cgi-pub/engcg). The parser gives the option of analysing the text with or without the use of heuristics. Both these options have been used for this study. The same text sample has, thus, been analysed twice, once with the use of the heuristic set of rules in addition to the ordinary set and once without.

The output from the parser has been manually checked. The number of words in each sentence has been counted as well as the number of syntactic units. *Words* here refers to word tokens and letters but not to punctuation marks. Syntactic units are single words or groups of words entered on single lines (see example below).

```
(1). a."<press_conference>"
     b. "<as=if>"
```

Both these sequences of words are fitted together on a single line by the Constraint Grammar Preprocessor and are considered as one unit but two words. The '_' character is used for compounds while phrasal idioms and multi-word prepositions are fitted together with the character '='.

Different morphological analyses are represented as separate lines; readings, along with syntactic analyses (see example below).

```
(2). "<*you>"
      "you" <*> <NonMod> PRON PERS ACC SG2/PL2  @OBJ
      "you" <*> <NonMod> PRON PERS NOM SG2/PL2  @SUBJ
```

In this case the word *you* has two morphological readings with one syntactic analysis each.

When checking the analysis, tags that are deemed to be incorrect have been marked and correct ones, if missing, have been added to the analysis. All these tags have been counted yielding figures for returned tags, correct tags, and intended correct tags for each sentence.

In some cases a word or unit has been marked with more than one tag that was deemed to be correct.

```
(3).     "<the>"
                "the" <Def> DET CENTRAL ART SG/PL @DN>
         "<*m25>"
                "m25" <*> ABBR NOM SG  @<P
         "<in>"
                 "in" PREP  @<NOM @ADVL
         "<*kent>"
                "kent" <*> <Proper> N NOM SG  @<P
         "<$.>"
```

In the example above, the preposition *in* is tagged both with an @ADVL tag, marking it as a independent adverbial, and a @<NOM tag, marking it as a post-modifying preposition. Both these analyses are possible and this unit will for this reason yield two correct tags and two intended correct tags. In other cases different morphological readings may yield the same correct syntactic analysis.

```
(4). "<more=than>"
        "more=than" ADV  @ADVL
        "more=than" <CompPP> PREP  @ADVL
```

In this example the ADV reading is incorrect and the PREP reading correct. Both readings are, however, marked with the correct syntactic tag ((@ADVL). In this case both analyses are deemed correct and this unit is, thus, assigned two correct tags. This may be considered a kind of overgeneration of syntactic tags. It does not, however, affect the result in a negative way. In most cases there is only one intended correct tag for each syntactic unit.

4. Result

Since the analysis may yield more than one tag, correct as well as incorrect, for one syntactic unit, it is reasonable to divide the result into two parts: recall and precision. Recall is calculated by the formula: returned correct tags divided by intended correct tags, while precision is calculated by the formula: returned correct tags divided by all returned tags. For example: the sentence *'All in all, Berlin and Kay appear to have dealt a severe blow to the notion of linguistic relativism.'* consists of 19 words. *All in all* is considered an idiom and all three words are fitted together on a single line. The sentence, thus, has 17 units. No unit is morphologically ambiguous so the number of readings is also 17. The preposition *to* before the last noun phrase is syntactically ambiguous between @<NOM @ADVL yielding a number of 18 syntactic tags in total. The @<NOM analysis for this reading is considered incorrect. This gives a number of 17 correct tags equalling the number of intended correct tags. The recall for this sentence will be 100% (17 correct tags divided by 17 intended correct tags. The precision is about 94.4% (17 correct tags divided by 18 returned tags)

The recall and precision for each test text is presented in table 2 below.

		Learned	Fiction	Biogr.	Govern.	Press	Total
No Heur.	Recall	93.9	97.4	96.0	96.4	99.4	96.5
	Precision	70.1	65.7	71.3	68.4	78.1	70.5
Heur.	Recall	93.4	96.4	95.7	96.4	98.8	96.0
	Precision	74.2	67.1	73.3	68.5	78.1	72.0

Table 2. Recall and precision for the text sample (%)

The recall is higher for the non-heuristic parsing than for the heuristic parsing. This is true in all categories except Governmental. In that category, recall shows identical figures for the heuristic and the non-heuristic analyses. The precision is lower for the non-heuristic analysis than for the heuristic. This is true in all categories except press. The immediate conclusion to be drawn from this is that the optional addition of heuristics lowers the recall but increases the precision. The heuristic rules discard many erroneous tags along with some of the correct ones.

5. Comments

Verb sequences are often correctly analysed and there is only a small amount of overgeneration.

```
(5).    "<will>"
                "will" V AUXMOD VFIN @+FAUXV
        "<be>"
                "be" <SV> <SVC/N> <SVC/A> V INF  @-FAUXV
        "<returned>"
                "return" <SVOC/A> <SV> <SVO> PCP2  @-FMAINV
```

Adverbial phrases are often correctly tagged. Modifiers in noun phrases, such as determiners, pre- and post-modifying nouns and adverbials mostly get correct analyses.

```
(6).    "<his>"
                "he" PRON PERS MASC GEN SG3  @GN>
        "<late>"
                "late" A ABS  @AN>
        "<years>"
                "year" N NOM PL  @<P
```

Nominal heads cause a considerable amount of overgeneration. The parser is often unable to relate heads of nominal phrases to the rest of the sentence.

```
(7).    "<term>"
                "term" N NOM SG  @SUBJ @OBJ @<P
```

6.The Effect of Sentence Length and Morphological Analysis on the Performance of the Parser

Recall and precision do not seem to be affected at all by sentence length when single sentences are compared. Even the extremely long sentences in the test text have quite high recall and precision. If, however, the average recall and precision for each text category are compared to the average sentence length of corresponding category, sentence length affect both recall and precision. Text types with longer average sentence length have lower recall and precision than text types with shorter. This may, however, be caused by the difference in text types rather than by sentence length itself. A text type with longer sentences may be more complicated and more difficult to analyse. To determine to what extent this is the case is, unfortunately, outside the scope of this study.

The number of morphological analyses also affects recall and precision. If the number of morphological readings per unit is compared to the precision, there is a visible correlation in that a high number of morphological readings per unit gives a low figure for precision. Overgeneration of morphological analyses seems to feed overgeneration of syntactic labels. The correlation between the number of morphological readings per word and recall seems to be reversed. A high number of morphological readings gives a high recall, indicating that a high number of morphological readings increases the chances of correct syntactic analyses.

7. Conclusions

This study has shown that the recall for the syntactic parsing seems to lie around 93-99%, higher for the non-heuristic parsing and lower for the heuristic. The precision is considerably lower and falls between 66-78%, higher for the heuristic parsing and lower for the non-heuristic. There seems to be a balance between recall and precision. The addition of heuristics in the parsing lowers the recall but makes the precision higher as the use of heuristic rules discards both erroneous and correct analyses. This connection between recall and precision seems to be quite constant throughout all text categories.

There seems to be no connection between sentence length and the performance of the parser in terms of recall and precision. Text types with longer average sentence length, however, have both lower recall and precision than text types with shorter average sentence length.

Morphological ambiguities seem to feed syntactic ambiguity. This is not at all surprising since every reading of a single syntactic unit will receive at least one syntactic label. Overgeneration of morphological readings feeds overgeneration of syntactic analyses since every morphological reading more than necessary generates at least one syntactic tag more than necessary. The overgeneration of syntactic tags does, however, not always affect recall and precision in a negative way. The overgeneration of syntactic tags seems to be the reason why the connection between recall and morphological overgeneration seems to be reversed. If a word has many syntactic tags attached to it, the chance of finding the correct one among them is greater than if only one syntactic tag is present.

The test text has been compiled from following sources:

Learned
Geoffrey Sampson, Schools of Linguistics, 1980, Hutchinson, London, Melbourne, Sydney, Auckland, Johannesburg, pp. 98-100.

Fiction
Margaret Drabble, The Needle's Eye, 1972, Penguin Books, pp. 190-191.

Biography
T. C. Duncan Eaves and Ben D. Kimpel, Samuel Richardson, A Biography, 1971, pp 11-12.

Governmental
Document authorised by The Council of The Stock Exchange under section 154 (1) (b) of the Financial Service Act 1986, Terms and Conditions of Application.

Press
John, Steele, Friday, May 24, 1996, The Daily Telegraph, pp. 1-2.

References

Karlsson, Fred; Voutilainen, Atro; Heikkilä, Juha; Anttila, Arto. eds. (1995). *Constraint Grammar; A Language-Independent System for Parsing Unrestricted Text.* Mouton de Gruyter, Berlin, New York.