An English Constraint Grammar (ENGCG) a surface-syntactic parser of English

Atro Voutilainen and Juha Heikkilä

Research Unit for Computational Linguistics
University of Helsinki

Published in Udo Fries, Gunnel Tottie and Peter Schneider, Eds. (1993), *Creating and using English language corpora.* Rodopi: Amsterdam and Atlanta. Pages 189-199.

1 Introduction

This paper outlines a reductionistic surface-syntactic parsing description of English that was developed 1989–1992 by Atro Voutilainen, Juha Heikkilä and Arto Anttila (Voutilainen et al. 1992; Karlsson et al. 1994) within the Constraint Grammar framework originally proposed by Fred Karlsson (1990). The parsing system employs two main components: a morphological analyser and a reductionistic parser.

The morphological analyser and description is based on Koskenniemi's Two-level Model (1983) that has become a near standard in computational morphology. It employs a morphological description and a large lexicon, and its function is to provide all conventional analyses to each recognised word. For instance, here is the morphological analysis of the sentence *Tests concluded the paper* as produced by our English analyser:

```
"<*tests>"
        "test" <*> <SVO> <SV> V PRES SG3 VFIN
        "test" <*> N NOM PL
"<concluded>"
        "conclude" <Vcog> <SVO> <SV> PCP2
        "conclude" <Vcog> <SVO> <SV> V PAST VFIN
"<the>"
        "the" <Def> DET CENTRAL ART SG/PL
"<paper>"
        "paper" <SVO> V SUBJUNCTIVE VFIN
```

```
        "paper" <SVO> V IMP VFIN
        "paper" <SVO> V INF
        "paper" <SVO> V PRES -SG3 VFIN
        "paper" N NOM SG
("<$.>")
```

Each indented line is a morphological[1] reading that consists of a base form and morphosyntactic tags indicating e.g. part of speech and inflection. Some of the words receive several alternative readings, i.e. they are morphologically ambiguous.

The syntactic part of the system is based on the Constraint Grammar framework, originally proposed by Karlsson (1990). Constraint Grammar parsing is reductionistic: the task of constraints – partial linear-precedence rules about the distribution of morphosyntactic categories – is to discard contextually illegitimate alternatives (morphological and syntactic readings). What 'survives' the reductionistic parser is the (part-of-speech or syntactic) analysis, each word optimally retaining an unambiguous correct reading. Here is our sample sentence after morphological disambiguation and syntactic analysis:

```
"<*tests>"
        "test" <*> N NOM PL @SUBJ
"<concluded>"
        "conclude" <Vcog> <SVO> <SV> V PAST VFIN @+FMAINV
"<the>"
        "the" <Def> DET CENTRAL ART SG/PL @DN>
"<paper>"
        "paper" N NOM SG @OBJ
"<$.>"
```

A shallow dependency-oriented functional description is assigned onto the sentence. Next, let us briefly present some key features of this style of reductionistic analysis.

First, the parsing scheme – the convention according to which the parser should express the grammatical relations of the words in the input sentences – is shallow: it does not introduce certain structurally unresolvable distinctions. Consider the following fragment:

```
"<fat>"
        "fat" A ABS @AN>
"<butcher's>"
        "butcher" N GEN SG @GN>
"<wife>"
        "wife" N NOM SG @SUBJ
```

[1] Our use of 'morphological' is here quite unrestrictive; it is actually interchangeable with 'lexical'.

The function tag @*AN*> indicates that the head of the word is a nominal in the right-hand context – either another premodifier or a nominal head. In other words, there is no necessity to tackle the structurally undecidable question, whether the butcher or his wife is fat. Attachment of adverbials is another hard ambiguity that our parsing schemes do not introduce, and therefore the parser has a better chance to produce correct and unambiguous analyses.

Second, the parsing description is based on linguistic rules rather than automatically generated corpus probabilities. The hand-written constraints are independent of each other – no inherent rule order is involved in the formalism – so they can easily be validated against representative corpora.

Third, the parsing grammar regards certain constructions as ill-formed; yet the parser assigns some analysis to every input. This kind of robustness is achieved by 'constraining' the application of the constraint grammar itself: in the case of an unambiguous input, the parser does not consult the grammar, so even ill-formed constructions survive as analyses. It is an empirical question, whether this manner of 'forcing' an analysis is desirable; presumably no analysis at all is better than a misleading analysis.

Last, the formalism is language-independent, and descriptions of a few languages are under development (English, Swedish, Danish, Basque). To date, the English description is the most mature one. The English Constraint Grammar Parser is currently being used for the analysis of the 200-million word Bank of English corpus in collaboration with the COBUILD team in Birmingham and the Harper/Collins Publishers in Glasgow, United Kingdom. The parsed corpus is likely to become accessible to the research community.

2 Overview of the English description

The English description was made by Voutilainen, Heikkilä and Anttila 1989–1992. Anttila wrote the syntactic mappings and syntactic constraints; Heikkilä and Voutilainen wrote the morphological description and the lexicon; Voutilainen wrote the constraints for morphological disambiguation as well as certain smaller rule components. The description is documented in Voutilainen et al. (1992) and in Karlsson et al. (1994).

Here is the ENGCG parsing architecture:

- preprocessor
- lookup of morphological readings
  - two-level analysis
  - heuristic analysis of unknown words

- parsing
  - resolution of morphological ambiguities
  - lookup of syntactic function tags
  - resolution of syntactic ambiguities

Next, the main rule components are presented in outline.

2.1 Morphological analysis

The first module in the system is the preprocessor that identifies sentence boundaries and fixed syntagms – idioms, such as multiword prepositions, and compound nouns. The current list of fixed syntagms contains some 6,000 items.

The ENGTWOL morphological analyser uses a lexicon which contains about 56,000 entries. It accounts for all inflections and central derivations, and the number of entries is adequate to cover the core vocabulary of English. Usually ENGTWOL is capable of recognising and analysing more than 95% of input word forms, and even in more technical texts the percentage is typically over 90%. The correctness of the readings produced by ENGTWOL is near 100% (see the test results reported in Section 3.1). For the analysis of those words not represented in the ENGTWOL description, there is a rule-based heuristic program that assigns ENGTWOL-style readings. Also this utility is highly reliable (see Section 3.1 and Voutilainen 1994b).

Since many English word forms can belong to more than one part of speech in isolation, the morphological analysis produces ambiguity; on an average, the ENGTWOL analyser provides each word with 1.7–2.2 alternative morphological readings. Ambiguity is, of course, highly dependent on the linguistic description adopted. The ENGTWOL description is mainly based on Quirk et al. (1985). We have systematically introduced categorial ambiguity, and we have also generously used other features to provide ample information – the total number of features used is approximately 150. Hence, ENGTWOL typically makes somewhat finer distinctions than the tag set of the Brown University Corpus. Uninflected verb forms, for example, are ambiguous over the present tense, imperative, infinitive and subjunctive readings, whereas the Brown tag set only gives one uninflected verb form reading. Also, pronoun/determiner ambiguities (e.g., *much*, *this*) have been spelt out in ENGTWOL, while the Brown tag set often subsumes these categories under the same tag. On the other hand, the Brown tag set divides *-ing*-forms into participles, nouns and adjectives – ENGTWOL subsumes these under the feature PCP1.

The ENGTWOL feature system is well specified in the sense that linguists familiar with the ENGTWOL description tend to agree in virtually 100% of the cases when they (manually) apply the description to running text (Voutilainen 1994a). Without a well-specified 'underlying' descriptive practice or parsing scheme, the development or evaluation of a parsing grammar could obviously not be satisfactory.

2.2 Morphological disambiguation

The morphological disambiguator optimally discards all and only the contextually illegitimate morphological analyses in ambiguous cohorts. In practice, this ideal is not always attained: occasionally, a legitimate reading is discarded; sometimes an illegitimate reading survives. A realistic solution optimises both recall and precision[2].

The present grammar for morphological disambiguation contains about 1,100 constraints. These constraints express distributional restrictions about ambiguity-forming morphological tags, usually in a negative fashion.

Here, we will not go into the rule formalism; a few prose paraphrases will demonstrate the nature of these constraints:

- *Discard all verb readings if to the left there is an unambiguous determiner, and between that determiner and the ambiguous word itself, there are no nominals (nouns, abbreviations etc.).*
- *Discard all finite verb readings if the immediately preceding word is* to.
- *Discard all subjunctive readings if to the left, there are no instances of the subordinating conjunction* that *or* lest.

For instance, the first constraint would discard the verb reading in the following representation:

```
"<the>"
        "the" <Def> DET CENTRAL ART SG/PL
"<tables>"
        "table" N NOM PL
        "table" <SVO> V PRES SG3 VFIN
```

Reference to words and tags is possible. Actually, about 40% of all constraints in the present grammar are expressly about word-form types rather than grammatical tags. More generally, it is possible to express distributional generalisations also about 'smaller' classes, even about particular word-form

[2]These terms are familiar from information retrieval literature. In the present context, precision is the fraction of the retrieved morphological analyses that are appropriate; recall is the fraction of the appropriate morphological analyses that are retrieved.

types. This 'lexico-grammatical' orientation makes for the exhaustiveness of the grammar, i.e. for the accuracy of the disambiguator.

New constraints are written to resolve those ambiguities that survive the grammar written up to that point in the development. In this sense, the grammar is based on phenomena in actual texts rather than on some more or less artificially predefined collection of grammatical constructs. Although large descriptive grammars such as Quirk et al. (1985) have been consulted, the present grammar certainly is more extensive on those phenomena that occur with high frequency in running text.

The constraints have been tested against large text samples mainly representing standard written English of the British and American English varieties. More than three million words of the disambiguator's output has been proofread during the development of the grammar. The grammar is less reliable when used in the analysis of fiction, spoken language, dialect or otherwise nonstandard texts.

The grammar statements underlying the presents 1,100 constraints are relatively few and, interestingly, about the form and function of syntactic categories rather than about some phenomena particular to morphology or morphological ambiguity itself. For instance, the underlying 23 grammar statements concern the form of noun phrases, agreement, coordination, and the form of the clause and the sentence. Indeed, it is argued in (Voutilainen and Tapanainen 1993; Voutilainen 1993b; Voutilainen 1994a) that morphological disambiguation could be carried out as a mere side effect of syntactic parsing in a framework where the formalism itself does not distinguish between different 'levels' of description or ambiguity.

2.3 Constraint syntax

The morphologically analysed and disambiguated text is passed on to the syntactic module of ENGCG. The purpose of syntactic analysis is to assign the correct surface-syntactic functions to input word forms.

Like morphological information, syntactic information in ENGCG is incorporated in tags that indicate the lexical heads of major clausal constituents – e.g., @SUBJ, @OBJ for subject and object, respectively – and modifier functions, whose tags indicate the type and direction of their head – e.g., @AN> for adjectival premodifiers of nominals.

The process of determining the correct syntactic functions is carried out in two consecutive stages: morphosyntactic mapping and syntactic disambiguation. The mapping module assigns all possible syntactic function tags to each morphological reading, for instance furnishing each noun reading with a disjunctive list of tags for Subject, Object, Preposition complement, and so on.

The second stage of syntactic analysis, syntactic disambiguation, seeks to discard those functions that are contextually illegitimate. The mechanism is similar to that used in morphological disambiguation: superfluous tags are pruned on the basis of contextual information. The difference is that syntactic disambiguation only affects syntactic-function tags rather than entire morphological readings. As in morphological disambiguation, only structurally resolvable ambiguities are tackled; thus, e.g., PP attachment ambiguities will be left pending.

3 Performance tests

In this section, we report on the accuracy of the present system. The present parsing system is designed to maximise recall: the correct analysis is usually in the parser's output, even though some ambiguity also survives where the parser cannot make a safe decision. According to our past experience, this strategy works quite well:

As the combined effect of the morphological analyser and morphological heuristics, very close to 100% of all words get the contextually legitimate morphological analysis.

The morphological disambiguator is very reliable. Of all words, typically at least 99.7% retain the contextually legitimate analysis after morphological disambiguation, while the rate of ambiguity is reduced considerably: on an average, each word retains only 1.05–1.08 morphological readings. In other words, more than 93% of all words become fully disambiguated.

The syntactic module is promising though not quite as mature as the other components. Anttila (1994) reports that some 96–97% of all words retain the correct syntactic function tag, while 80% or more of all words become syntactically unambiguous.

However, we are not content with previous results alone; also new test data should be used to accumulate further evidence about the performance of the system. In this section, we test the accuracy of the morphological analyser (in conjunction with the morphological heuristics module) and of the morphological disambiguator. New results about the syntactic module will become available once current work on syntax by Timo Järvinen has reached a relative stability.

3.1 Lexicon and morphological heuristics

The test text was an extract of 12,548 words from *The Independent* newspaper of 1 October 1990, taken from the Bank of English Corpus.

The test focused on correctness. An error was registered when the analyser had failed to produce the reading which was appropriate with respect to both the context and the form of the input word.

The correctness of analyses can reasonably be measured only against the descriptive policy adopted, and some analyses, although seemingly unintuitive or unorthodox, may be perfectly appropriate with respect to the adopted description. Some reference is made to individual solutions below, but a detailed discussion of the descriptive policy and its consequences is beyond the scope of this paper (see the lexicon chapters in Karlsson et al. 1994).

Sometimes it is difficult to draw a clear-cut boundary between errors and valid analyses because readings have varying degrees of badness. It is therefore illuminating to divide the errors into three classes on the basis of how serious their consequences are (see Table 1). Errors which were clearly due to deficiencies in the input text (two instances) were not counted.

Table 1: Errors in morphological analysis.

Class	ENGTWOL	Heuristics	Total
1	9	2	11
2	4	0	4
3	6	0	6
1-3	19	2	21 (0.167%)

Class 1, eleven cases, consists of indisputable errors in which all the information provided in the readings is inappropriate. This is bound to confuse both morphological and syntactic disambiguation. A prime example of such an error is a nominal which has only received verb readings, or a verb which has only received nominal readings. Except for one case, which was caused by the idiom-detection program, all the Class 1 errors were of this type. Interestingly, as many as five of the errors were caused by the word form *Aids*, which was analysed as a verb form; ENGTWOL would have, however, correctly recognised the spelling *AIDS*.

The four cases in Class 2 have correct part-of-speech information, but either wrong number or case. This is obviously less fatal, but it may lead the parser astray in, e.g., nominal chains. Three of the errors were a direct consequence of the descriptive decision not to include proper names in the lexicon if there exists a corresponding common noun. This is expedient since the distributions of these two categories are very similar, but it produces a plural reading for proper names which end in *-s*.

Class 3 (six cases) consists of word forms whose analyses are inconsistent with our descriptive policy, but do not, however, cause any problems for syntax because the correct syntactic analysis can be accessed via some other reading. For example, the word form *upset* did not carry a noun reading, but in syntax it could still get the correct function as the nominal functions will also need to be available to past participles, which *upset* can be as well.

The performance of morphological heuristics is evaluated more leniently. As it does not attempt to make any distinctions between different types of nominal parts of speech, any of the readings A, N, ABBR, PCP1 or PCP2 has been considered appropriate for nominals. Heuristics component assigned a reading to 283 word forms. Thus approximately 97.74% of the input words were recognised and analysed by the morphological analyser.

If we take a strict, lexicographical view and count all three classes of errors, 0.17% of input word-form tokens did not carry the correct reading. If, however, the criterion is whether the "error" has consequences for syntactic analysis, we should exclude at least Class 3, and perhaps even Class 2. In these cases, the error rate would be 0.12% and 0.09%, respectively.

3.2 Morphological disambiguation

As test data, three randomly chosen extracts from two newspapers and a text book on the British constitution were used, 4,850 words in all.

The input text was actually disambiguated twice: once manually, according to our parsing scheme (the human expert selected the appropriate alternative from morphologically analysed input), and once with the ENGCG disambiguator itself. The disambiguator's output was compared to the manual analysis. This method maximises the likelihood that all misanalyses become detected; the alternative way of manually (actually, ocularly) checking the disambiguator's output runs a bigger risk of ignoring some misanalysis.

Of all words, 94.5% became unambiguous, and 99.8% (4,840) of all words retained the correct morphological analysis.

3.2.1 Misanalyses

Let us look at some of the misanalyses. Two of the ten errors were due to a mistake in the ENGTWOL description itself: the word-form *dared* had not been provided with the modal auxiliary reading in the lexicon, and this resulted in two misanalyses in the following sentence (the misanalysed words are given in UPPER CASE):

```
While keeping a watchful eye on social discontent,
they are determined to invigorate the economy in ways
the Communists never DARED CONSIDER.
```

Because *dared* missed a modal auxiliary reading, *consider* lost its infinitive reading as a domino effect (and retained another verb reading).

Let us examine a couple of the eight errors that were due to the disambiguation constraints:

```
For the middle classes Britain meant business
and they wanted constitutional change to prevent
```

```
revolution not PROMOTE it.

They were forced to ponder ON whether the Reform
Bill would end the old system, or, in aiding the
middle classes, merely add to the strength and
number of working-class oppressors.
```

The infinitive reading of *promote* was discarded because the grammar did not account for this use of *not*. The preposition reading of *on* was discarded because a constraint disallowed a subordinating conjunction as the following word; in other words, the possibility that clauses starting with the subordinating conjunction *whether* can serve as a preposition complement. Note that the immediate left-hand context strongly favours the preposition reading of *on*, i.e. allowing a clause as a preposition complement would not at least in this case result in unresolvable ambiguity.

3.2.2 ENGCG in perspective

To conclude this section: How does the present disambiguator compare with other systems in terms of accuracy? There are more than ten part-of-speech disambiguators of English that are based on statistical methods (see the bibliography in Church 1992). These statistical systems have an interestingly uniform accuracy: with all ambiguity resolved, some 95–97% of all words retain the correct reading. Also, this accuracy seems to be long-lived: already CLAWS1 (Garside et al. 1987) had an accuracy of 96–97%.

Because ENGCG leaves some ambiguity in its output, comparing its accuracy to these figures is not quite straightforward. Fortunately, some reports have also been made about how statistical systems perform when they are allowed to leave some ambiguity unresolved:

- Johansson (1986) reports about CLAWS1 that only 86% of all words can be fully disambiguated if 99% of all words are to retain the correct part-of-speech tag (with no ambiguity in the output, CLAWS1 is 96–97% accurate).

- De Marcken (1990) gives several figures about his system:

tags/word	recall (%)
1	96
1.04	97.56
1.09	98.57
1.14	99.21
1.27	99.93

- Weischedel et al. (1993) report about their POST tagger that to have 99.3% of all words with the correct part-of-speech tag, each word in the output will have as many as three alternative part-of-speech tags on an average.

Compared to these figures, the accuracy of ENGCG appears to be superior.[3] For instance, while de Marcken's figures suggest that with 1.07 tags/word, the system's recall is about 98.1%; with the same amount of ambiguity in the output, ENGCG's recall is 99.8%, i.e. almost ten times better. Considering that the statistical approach to part-of-speech disambiguation has nearly become a standard in recent years, the accuracy of the present grammar-based disambiguator should appear more than a little thought-provoking. – To say the least, it seems to us that the grammar-based approach should not be ignored in the design of accurate parsing systems (for more pessimistic views on the grammar-based approach, see e.g. Sampson 1987 and Church 1992).

4 Future directions

So far we have seen a system that has a high recall. The treatment of remaining ambiguities is the main pending problem. We have reasons for optimism here; let us conclude by outlining a few strategies.

Also heuristic constraints can be written for resolving remaining ambiguities. There is actually a heuristic add-on grammar of 200 constraints that resolves about 50% of the remaining morphological ambiguities 96–97% correctly (Voutilainen 1994b).

Another very promising approach is to use the fully disambiguated part of the text itself as a 'learning corpus' and apply the generalisations – probably lexical, perhaps also structural – to the resolution of the remaining few ambiguities (for experiments in this direction, see Voutilainen 1994b).

Perhaps the most promising possibility is optimising the design of the parsing formalism and the representation of grammatical relations. Quite a few of the ambiguities that the present grammar leaves pending are due to some undesirable properties in the design of the present parsing formalism and in the parsing scheme. New design considerations and also some experiments are outlined in (Koskenniemi 1990; Koskennniemi et al. 1992; Voutilainen and Tapanainen 1993; Voutilainen 1993).

[3] *appears* because the tag sets in the taggers are somewhat different, i.e. the tagging tasks are not quite the same.

5 Technical information

There are several computer implementations of the present parsing system. The fastest, using Pasi Tapanainen's C language implementation of the parser, analyses about 700 words per second on a Sun SPARCStation 10, Model 30, from preprocessing through syntactic analysis. The system is available; for details contact Atro Voutilainen (e-mail address: Atro.Voutilainen@Helsinki.FI).

References

Anttila, A. (1994). How to recognise subjects in English. In Karlsson et al.

Church, K. (1992) Current Practice in Part of Speech Tagging and Suggestions for the Future, in Simmons (Ed.), *Abornik praci: In Honor of Henry Kucera*. Michigan Slavic Studies.

Garside, R., Leech, G. and Sampson, G. (Eds.) (1987). *The Computational Analysis of English. A Corpus-Based Approach*. London: Longman.

Johansson, Stig 1986. *The Tagged LOB Corpus: User's Manual.* Bergen: Norwegian Computing Centre for the Humanities.

Karlsson, F. (1990). Constraint Grammar as a framework for parsing running text. In Karlgren, H. (Ed.) *COLING-90. Papers presented to the 13th International Conference on Computational Linguistics*. Vol. 3. Helsinki, Finland. 168–173.

Karlsson, F., Voutilainen, A., Heikkilä, J. and Anttila, A. (Eds.) (1994). *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text*. Berlin: Mouton de Gruyter.

Koskenniemi, K. (1983). *Two-Level Morphology: a General Computational Model for Word-Form Recognition and Production*. Publication No. 11, Department of General Linguistics, University of Helsinki.

Koskenniemi, K. (1990). Finite-state parsing and disambiguation. In H. Karlgren (Ed.), *COLING-90: Papers presented to the 13th international conference on computational linguistics*. Vol. 2. Helsinki, Finland. 229-232.

Koskenniemi, K., Tapanainen, P. and Voutilainen, A. (1992). Compiling and using finite-state syntactic rules. In *Proceedings of the fifteenth International Conference on Computational Linguistics. COLING-92* Vol. I. Nantes, France. 156-162.

Marcken, C. de (1990). Parsing the LOB corpus. In *Proceedings of the 28th annual meeting of the Association of Computational Linguistics*. 6–9 June 1990. University of Pittsburgh, Pittsburgh, Pennsylvania, USA.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Sampson, G. (1987). Probabilistic Models of Analysis. In Garside et al. (Eds.).

Voutilainen, A. (1993). *Designing a parsing grammar.* Publications of the Department of General Linguistics, University of Helsinki.

Voutilainen, A.. (1994a). Morphological disambiguation. In Karlsson et al.

Voutilainen, A.. (1994b). Experiments with heuristics. In Karlsson et al.

Voutilainen, A., Heikkilä, J. and Anttila, A. (1992). "Constraint grammar of English. A Performance-Oriented introduction". Publications No. 21, Dept. of General Linguistics, University of Helsinki.

Voutilainen, A. and Tapanainen, P. (1993). Ambiguity resolution in a reductionistic parser. In *Proceedings of EACL-93.* 394-403.

Weischedel, R., Meteer, M., Schwartz, R., Ramshaw, L. and Palmucci, J. (1993). Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics.* 19:2.