Applying the Constraint Grammar Parser of English to the Helsinki Corpus[1]

Merja Kytö and Atro Voutilainen
University of Helsinki

1. Introduction

The international break-through of the ENGCG Parser, or the Constraint Grammar Parser of English (Karlsson *et al*. 1995), as a system suitable for analysing Present-day English, opened the field for applications capable of dealing with regional, diachronic and other varieties of English. Using the parser on historical data is an obvious extension. To what extent should one "teach" the parser to cope with language from earlier periods? To what extent does Present-day English differ from early English and to what extent is it possible to formalize this difference for the parser?

In this pilot study we will report some recent findings, after applying the ENGCG parser to a number of texts from the Early Modern English section of the Helsinki Corpus of English Texts. We first explain how the parser works and its relevance to our study. We then introduce our data, explain how we carried out the ENGCG analysis, and discuss the parser's difficulties in dealing with early English. Finally, we evaluate our experiment and the ENGCG parser's potential in diachronic multi-purpose corpora such as the Helsinki Corpus.

2. The English Constraint Grammar Parser ENGCG

In this section we outline the English Constraint Grammar Parser (ENGCG) and present some performance figures to illustrate the capacity of the programme applied to Present-day English data.

2.1 Background

The first version of the ENGCG description (preprocessor, lexicon and grammar) was developed in 1989–1992 by Atro Voutilainen, Juha Heikkilä and Arto Anttila (Voutilainen *et al.* 1992, Karlsson *et al*. 1995) as part of an ESPRIT II project (SIMPR, Project 2083), based on the Constraint Grammar theory of Fred Karlsson (Karlsson 1990, 1995), who also wrote the first version of the parser-interpreter in LISP. Timo Järvinen has since developed the ENGCG syntactic rules, and Pasi Tapanainen has developed a fast new implementation of the parsing program in C.

ENGCG is presently available for academic and industrial uses.[2] To date, the most important application of ENGCG is the morphosyntactic annotation of the 200-million-word Bank of English Corpus compiled by Harper/Collins Publishers (Glasgow) and the COBUILD team (Birmingham, England). The annotation has been done at the Research Unit for Computational Linguistics at the University of Helsinki; the whole corpus will have been syntactically analysed by the end of 1994.

2.2 The system in outline

We now examine the parsing modules in ENGCG and give some facts about the descriptive components. The ENGCG parser consists of the following sequentially applied modules:

INPUT: 7-bit ASCII text

- Preprocessing
- Morphological analysis:
  - Lexical component
  - Heuristic component
- Morphological disambiguation
- Syntactic analysis:
  - Introduction of syntactic ambiguities
  - Resolution of syntactic ambiguities

OUTPUT: Morphosyntactically analysed sentences

Next, we outline these components.

2.2.1 Preprocessing and morphological analysis

The first stage in parsing is preprocessing, i.e. identification of punctuation marks (incl. sentence boundaries); identification of fixed syntagms such as certain compounds, multiword prepositions and other idiomatic constructions; and normalisation of certain orthographical conventions. This module is implemented as a set of some 7000 rewrite rules, most for recognising fixed syntagms in the BETA programming language (cf. Brodda 1990).

The morphological analyser's main component is a morphosyntactic lexicon designed according to Koskenniemi's famous Two-level Model (Koskenniemi 1983). At present, the English lexicon ENGTWOL, which contains about 84,000 lexical entries, represents the core vocabulary of Present-day English, including all inflected and central derived English word-forms. Prefixes and endings are represented in separate 'minilexicons' accessible from the 'stem' lexicon.

ENGTWOL also employs a feature system largely based on Quirk *et al*. (1985), incorporating 139 morphosyntactic tags, some for parts of speech, others for minor categories – number, case, mood, etc. – and for essentially syntactic properties like verb valency (Heikkilä 1995).

The ENGTWOL lexicon is run by *twol*, the two-level program. The ENGTWOL analyser recognizes about 95–99% of all running-text word-form tokens, depending on text type. For each recognized word-form token, one or more morphological analyses are given. Here is the ENGTWOL analysis of the sentence *That round table might collapse*:[3]

```
"<*that>"
    "that" <*> <**CLB> CS @CS
    "that" <*> DET CENTRAL DEM SG @DN>
    "that" <*> ADV AD-A> @AD-A>
    "that" <*> PRON DEM SG
    "that" <*> <NonMod> <**CLB> <Rel> PRON SG/PL
"<round>"
    "round" <SVO> <SV> V SUBJUNCTIVE VFIN @+FMAINV
    "round" <SVO> <SV> V IMP VFIN @+FMAINV
    "round" <SVO> <SV> V INF
    "round" <SVO> <SV> V PRES -SG3 VFIN @+FMAINV
    "round" PREP
    "round" N NOM SG
    "round" A ABS
    "round" ADV ADVL @ADVL
"<table>"
    "table" N NOM SG
    "table" <SVO> V SUBJUNCTIVE VFIN @+FMAINV
    "table" <SVO> V IMP VFIN @+FMAINV
    "table" <SVO> V INF
    "table" <SVO> V PRES -SG3 VFIN @+FMAINV
"<might>"
    "might" <-Indef> N NOM SG
    "might" V AUXMOD VFIN @+FAUXV
```

```
"<collapse>"
     "collapse"  N  NOM  SG
     "collapse"  <SV>  <SVO>  V  SUBJUNCTIVE  VFIN  @+FMAINV
     "collapse"  <SV>  <SVO>  V  IMP  VFIN  @+FMAINV
     "collapse"  <SV>  <SVO>  V  INF
     "collapse"  <SV>  <SVO>  V  PRES  -SG3  VFIN  @+FMAINV
"<$.>"
```

This sentence is heavily ambiguous. On average, however, each word receives 1.7–2.2 alternative morphological analyses from ENGTWOL.

As noted above, the ENGTWOL analyser leaves some 1–5% of all word-form tokens unanalysed. There are two main solutions for processing this residue. One option is updating the ENGTWOL lexicon itself by employing a semi-automatic routine to identify the unrecognized words and convert them into lexical entries. This option proved highly profitable in analysing diachronic texts (see below). The other option is using a heuristic rule-based module, ‘Morphological heuristics’, which automatically assigns one or more ENGTWOL-style analyses to unanalysed words by predicting from certain patterns in the words, e.g. endings. If no patterns match, the word is analysed as a nominal by default.

ENGTWOL and Morphological heuristics together provide each word with one or more analyses. About 99.9% of all word-form tokens in running text usually receive a contextually appropriate analysis from the morphological analyser (Voutilainen and Heikkilä 1995).

2.2.2 Morphological disambiguation

The morphological analyser provides several alternative analyses for 35–50% of all words in the input sentences, but usually only one analysis ‘fits’ in context. The morphological disambiguator identifies the correct alternative by discarding as many contextually illegitimate alternatives as possible.[4]

Optimally, the disambiguator outputs unambiguous and correctly tagged sentences. In practice, this goal is very difficult, perhaps impossible, to achieve in the analysis of unconstrained text. ENGCG rejects only those alternatives involving a very small risk of error. Because the few hardest cases are left pending, the output of the tagger is somewhat ambiguous.[5]

Most morphological or part-of-speech disambiguators (Church 1988, Leech *et al*. 1994, de Marcken 1990) employ co-occurrence-based and lexical statistics usually derived from (manually) tagged corpora. ENGCG, by contrast, employs only hand-written linguistic rules, or constraints expressing restrictions on the linear order of words and tags. Usually these constraints are very partial and roundabout expressions of essentially syntactic statements (Voutilainen 1994). Generally they take the form ‘discard reading X if all context conditions are satisfied; otherwise leave X intact’. The context conditions can refer to fixed word positions (e.g. ‘the second word to the left contains the tag X’) or to unbounded context within the sentence (e.g. ‘to the right, there is no X’). The details of the rule formalism need not concern us here; we only paraphrase some simple constraints to give the general idea:

(i) "Discard all finite verb readings if the preceding word is an unambiguous determiner".

(ii) "Discard all subjunctive readings unless the left-hand context contains *that* or *lest* as a subordinating conjunction".

(iii) "Discard all finite verb readings if the preceding word is *to*".

The present grammar contains two sections, the ‘grammar-based’ section and the optionally applicable ‘heuristic’ section, the former presently containing about 1,150 constraints, making 93–97% of all words unambiguous, with at least 99.7% retaining the correct morphological analysis. Applying the 200-odd heuristic constraints also, 96–98% of all words become unambiguous, but at this stage only about 99.5% retain the correct morphological analysis.[6]

2.2.3 Syntactic analysis

Syntactic analysis in ENGCG is based on the use of shallow dependency-oriented functional tags attached to words, much like morphological analysis. The present version uses some 30 syntactic tags flanked with the ‘@’-sign, e.g. @SUBJ, @OBJ, @I-OBJ, @<P (preposition complement), @AN> (premodifying adjective), @+FMAINV (finite main verb) (Anttila 1995). This kind of syntactic tagging does not directly indicate e.g. the phrase structure of a sentence. This shallow annotation could undoubtedly identify some low-level phrases, but recovering the full phrase structure from ENGCG analysis of a typical text sentence would require additional linguistic knowledge.

Syntactic analysis in Constraint Grammar proceeds much like the previous stages. Syntactic descriptors are first introduced as alternatives with simple mapping rules. For instance, a default rule for nouns introduces about ten syntactic tags as a list of alternatives for each noun reading, e.g:

```
"<table>"
      "table" N NOM SG @NPHR @SUBJ @OBJ @I-OBJ @PCOMPL-S
      @PCOMPL-O ...
```

This mapping produces considerable ambiguity, though some is avoidable even at this stage by using context-sensitive mapping rules before context-free default rules.

Syntactic disambiguation is carried out with constraints formally similar to constraints for morphological disambiguation. The only difference is that the discarded unit is a syntactic tag, not a morphological reading. The present syntactic grammar contains about 550 ‘grammar-based’ and 250 heuristic constraints, rendering about 85% of all words syntactically unambiguous. About 98% of all words retain the correct syntactic function tag at this stage (Järvinen 1994).

Finally, here is our sample sentence after full ENGCG parsing:

```
"<*that>"
      "that" <*> DET CENTRAL DEM SG @DN> ;; determiner
"<round>"
      "round" A ABS @AN>  ;; premodifying adjective
"<table>"
      "table" N NOM SG @SUBJ ;; subject
"<might>"
      "might" V AUXMOD VFIN @+FAUXV ;; finite auxiliary
"<collapse>"
      "collapse" <SV> <SVO> V INF @-FMAINV ;; nonfinite main verb
"<$.>"
```

The parser analysed this simple sentence perfectly.

2.3 Technical information

The parser is available for such computers as personal computers and Unix workstations. The main parts of the system have been implemented in C. On a Sun SPARCStation 10, Model 30, ENGCG parses about 400 words per second, from preprocessing through syntax.

3. ENGCG analysis of historical texts: the data

The Helsinki Corpus covers roughly a millennium, from the eighth century to the early 1700s.[7] Anticipating problems caused by distance in time between early English and the language accounted for in the ENGCG description, we decided to concentrate on the more recent Early Modern English period (1500–1710). To help the parser further, the study was restricted to the same text type, history writing, prose with similar kinds of subject matter and less esoteric generic conventions. The six texts considered total 32,700 running words (see Kytö 1993).

While the orthographic conventions and morphology of many texts from the third subperiod of Early Modern English in the Helsinki Corpus (1640–1710) approach Present-day English, great orthographic and morphological variation is still apparent in texts from the two first subperiods in this section (1500–1570 and 1570–1640). Two major remedies helped the parser deal with this variation. The first and perhaps most pressing need was to update (or, rather in our case, "back-date") the lexicon; the second was the need for possible modifications in the disambiguation grammar. Interest in seeing how well the parser would deal with early English data with minimal changes initially, led us to concentrate on supplementing the lexicons for the time being and postponing modifying other grammar modules until we had learnt from this first experiment. Results presented in this study are thus based on the use of ENGTWOL, that is, the (updated) lexicon and the morphological disambiguator; no use has been made of the syntax modules as yet.

The following tables indicate the work needed to update the lexicon of the six texts compared with that available in the Present-day ENGTWOL lexicon of the ENGCG parser, containing some 84,000 words at this point of the present study (for references to source texts, see Kytö 1993):[8]

Table 1. Token ratios (* = percentage of all tokens).

AUTHOR	DATE	ALL TOKENS	TOKENS NOT FOUND IN ENGTWOL	
Sir Thomas More	1514–18	5660	1680	*(30%)
Robert Fabyan	1516	5510	2140	(39%)
John Stow	1580	4840	1110	(23%)
Sir John Hayward	1627	5330	930	(17%)
John Milton	1670	5860	580	(10%)
Gilbert Burnet	*ante*1703	5870	40	(0.7%)

Table 2. Type ratios (* = percentage of all types).

AUTHOR	DATE	ALL TOKENS	TOKENS NOT FOUND IN ENGTWOL	
Sir Thomas More	1514–18	1580	950	*(60%)
Robert Fabyan	1516	1380	1000	(72%)
John Stow	1580	1500	720	(48%)
Sir John Hayward	1627	1620	560	(35%)
John Milton	1670	1670	360	(22%)
Gilbert Burnet	*ante*1703	1320	30	(2%)

In the two texts from the first subperiod, 30% to 40% of the tokens do not appear in the Present-day lexicon used by ENGCG (orthographic variants being counted as individual tokens). The percentages drop across the centuries so that in the extracts from the early 18th century work, *Burnet's History*, practically all forms can be found in the ENGTWOL lexicon. The same holds for the percentages obtained for different types, but more so: in Fabyan's *New Chronicles* from the early 1500s, over 70% of the types are new to ENGCG; in *Burnet's History*, only 2%.

4. Updating the lexicon

We opted to deal with the differences between the lexicons of our early English data and the ENGTWOL file through a data-based strategy and compiled a full word list from the six texts. This allowed us to build up a repertory of inflectional endings, with all orthographic variants. To facilitate subsequent analysis of other texts, we added further likely orthographic variants in the lists when convenient. To illustrate, we cite the beginning of one category in our list for verb stems (to help the morphological analyser, a default feature, the possibility of a transitive use, was added to each form in angle brackets):

```
LEXICON EMOD-Vroot
abandon .V "= <SVO> ";
abat .V "= <SVO> ";
abhorr .V "= <SVO> ";
abolish .V "= <SVO> ";
abus .V "= <SVO> ";
abyd .V "= <SVO> ";
accept .V "= <SVO> ";
accompany .V "= <SVO> ";
accompt .V "= <SVO> ";
accord .V "= <SVO> ";
accoumpani .V "= <SVO> ";
account .V "= <SVO> ";
accus .V "= <SVO> ";
accustom .V "= <SVO> ";
acknowledg .V "= <SVO> ";
acquaint .V "= <SVO> ";
act .V "= <SVO> ";
add .V "= <SVO> ";
address .V "= <SVO> ";
adher .V "= <SVO> ";
adiug .V "= <SVO> ";
adjur .V "= <SVO> ";
admir .V "= <SVO> ";
admonast .V "= <SVO> ";
admonish .V "= <SVO> ";
admytt .V "= <SVO> ";
adnull .V "= <SVO> ";
aduentur .V "= <SVO> ";
aduis .V "= <SVO> ";
aduyc .V "= <SVO> ";
aduys .V "= <SVO> ";
advanc .V "= <SVO> ";
adventur .V "= <SVO> ";
advic .V "= <SVO> ";
advis .V "= <SVO> ";
affect .V "= <SVO> ";
afferm .V "= <SVO> ";
affirm .V "= <SVO> ";
```

Part of the corresponding two-level rule combining the stems with some of the endings runs as follows (comment lines are indicated by an exclamation mark):

```
LEXICON .V
s # " V PRES SG3 VFIN (@+FMAINV) ";
es # " V PRES SG3 VFIN (@+FMAINV) ";
ys # " V PRES SG3 VFIN (@+FMAINV) ";
th # " V PRES SG3 VFIN (@+FMAINV) ";
eth # " V PRES SG3 VFIN (@+FMAINV) ";
yth # " V PRES SG3 VFIN (@+FMAINV) ";
the # " V PRES SG3 VFIN (@+FMAINV) ";
ethe # " V PRES SG3 VFIN (@+FMAINV) ";
ythe # " V PRES SG3 VFIN (@+FMAINV) ";
!
st # " V PRES SG2 VFIN (@+FMAINV) ";
est # " V PRES SG2 VFIN (@+FMAINV) ";
ste # " V PRES SG2 VFIN (@+FMAINV) ";
este # " V PRES SG2 VFIN (@+FMAINV) ";
!
d # " V PAST VFIN (@+FMAINV) ";
d # " PCP2 ";
de # " V PAST VFIN (@+FMAINV) ";
de # " PCP2 ";
dde # " V PAST VFIN (@+FMAINV) ";
dde # " PCP2 ";
ed # " V PAST VFIN (@+FMAINV) ";
ed # " PCP2 ";
id # " V PAST VFIN (@+FMAINV) ";
id # " PCP2 ";
yd # " V PAST VFIN (@+FMAINV) ";
yd # " PCP2 ";
!
t # " V PAST VFIN (@+FMAINV) ";
t # " PCP2 ";
te # " V PAST VFIN (@+FMAINV) ";
te # " PCP2 ";
'd # " V PAST VFIN (@+FMAINV) ";
'd # " PCP2 ";
't # " V PAST VFIN (@+FMAINV) ";
't # " PCP2 ";
!
'de # " V PAST VFIN (@+FMAINV) ";
'de # " PCP2 ";
'te # " V PAST VFIN (@+FMAINV) ";
'te # " PCP2 ";
ede # " V PAST VFIN (@+FMAINV) ";
ede # " PCP2 ";
ide # " V PAST VFIN (@+FMAINV) ";
ide # " PCP2 ";
yde # " V PAST VFIN (@+FMAINV) ";
yde # " PCP2 ";
!
en # " PCP2 ";
!
eing # " PCP1 ";
eng # " PCP1 ";
ing # " PCP1 ";
inge # " PCP1 ";
yng # " PCP1 ";
ynge # " PCP1 ";
!
einge # " PCP1 ";
enge # " PCP1 ";
eyng # " PCP1 ";
eynge # " PCP1 ";
```

Separating endings from stems was relatively easy and partly automatic with the word-processor. In the interest of further applications, we invested some time in classifying the words into different word classes manually. This was a more time-consuming operation, using the tabulator to sort the words into different columns. The *Oxford English Dictionary* was consulted for the more problematic classifications. Some spellings were quite puzzling out of context, but even these could be easily checked out from concordances.

For some special categories it was easier to write separate word-specific entries, such as the following describing a number of personal pronouns:

```
!
thow # "thou PRON PERS NOM SG2 SUBJ ";
!
shee # "she PRON PERS FEM NOM SG3 SUBJ ";
hir # "she PRON PERS FEM ACC SG3 ";
hyr # "she PRON PERS FEM ACC SG3 ";
!
hee # "he PRON PERS MASC NOM SG3 SUBJ ";
hi~ # "he PRON PERS MASC ACC SG3 ";
hime # "he PRON PERS MASC ACC SG3 ";
hym # "he PRON PERS MASC ACC SG3 ";
hy~ # "he PRON PERS MASC ACC SG3 ";
hyme # "he PRON PERS MASC ACC SG3 ";
!
hit # "it PRON NOM SG3 SUBJ ";
hit # "it PRON ACC SG3 ";
yt # "it PRON NOM SG3 SUBJ ";
yt # "it PRON ACC SG3 ";
hyt # "it PRON NOM SG3 SUBJ ";
hyt # "it PRON ACC SG3 ";
!! him # "it PRON DAT SG3 ";
!! hym # "it PRON DAT SG3 ";
!! hime # "it PRON DAT SG3 ";
!! hyme # "it PRON DAT SG3 ";
!! hi~ # "it PRON DAT SG3 ";
!! hy~ # "it PRON DAT SG3 ";
!
wee # "we PRON PERS NOM PL1 SUBJ ";
vs # "we PRON PERS ACC PL1 ";
!
ye # "you PRON PERS NOM SG2/PL2 SUBJ ";
ye # "you PRON PERS ACC SG2/PL2 ";
yow # "you PRON PERS NOM SG2/PL2 SUBJ ";
yow # "you PRON PERS ACC SG2/PL2 ";
!
thay # "they PRON PERS NOM PL3 SUBJ ";
thei # "they PRON PERS NOM PL3 SUBJ ";
thai # "they PRON PERS NOM PL3 SUBJ ";
the~ # "they PRON PERS ACC PL3 ";
theym # "they PRON PERS ACC PL3 ";
theim # "they PRON PERS ACC PL3 ";
hem # "they PRON PERS ACC PL3 ";
```

Before the analysis, we pre-edited the texts somewhat by removing headings, textual parameter coding, page numbers, source references and text-level coding; no changes were made in the text material proper other than replacing John Fabyan’s idiosyncratic use of slashes as punctuation marks by commas and periods.

5. Results

5.1 An overall view

On the whole, the results obtained were highly promising. Some extracts representative of different periods are given below, starting with earlier texts. The following is an extract from the earliest text analysed, Sir Thomas More’s "The History of King Richard III", from the early decades of the sixteenth century (p. 42):

When the lord Cardinal & these other lordes with him, had receiued this
yong duke, thei brought him into the sterrechamber where the protectour
toke him in his armes & kissed him with these wordes:

```
"<*when>"
     "*when" <*> <**CLB ADV WH @ADVL
"<the>"
     "the" <Def> ET CENTRAL ART SG/PL @DN>
"<lord>"
     "lord" <Title> N NOM SG
"<*cardinal>"
     "*cardinal" <*> <Title> N NOM SG
"<&>"
     "&" CC @CC
"<these>"
     "this" DET CENTRAL DEM PL @DN>
"<other>"
     "other" DET POST SG/PL @DN>
"<lordes>"
     "lord" N NOM PL
"<with>"
     "with" PREP
"<him>"
     "he" <NonMod> PRON PERS MASC ACC SG3
"<$,>")
"<had>"
     "have" <SVO> <SVO/A> V PAST VFIN
"<receiued>"
     "receiu" <SVO> PCP2
"<this>"                                              !!!!!
     "this" DET CENTRAL DEM SG @DN>
     "this" PRON DEM SG
"<yong>"
     "yong" A ABS
```

```
"<duke>"
    "duke" <Title> N NOM SG
"<$,>")
"<thei>"
    "they" PRON PERS NOM PL3 SUBJ
"<brought>"
    "bring" <SVO>  <SVOO> V PAST VFIN @+FMAINV
"<him>"
    "he" <NonMod> PRON PERS MASC ACC SG3
"<into>"
    "into" PREP
"<the>"
    "the" <Def> DET CENTRAL ART SG/PL @DN>
"<sterrechamber>"
    "sterrechamber" N NOM SG
"<where>"
    "where" <**CLB> ADV WH @ADVL
"<the>"
    "the" <Def> DET CENTRAL ART SG/PL @DN>
"<protectour>"
    "protectour" N NOM SG
"<toke>"                                                   !!!!!
    "tok" <SVO> PCP2
    "tok" <SVO> V PAST VFIN @+FMAINV
    "toke" <SVO> V IMP VFIN @+FMAINV
    "toke" <SVO> V PRES -SG3 VFIN @+FMAINV
"<him>"
    "he" <NonMod> PRON PERS MASC ACC SG3
"<in>"
    "in" PREP
"<his>"
    "he" PRON PERS MASC GEN SG3
"<armes>"
    "arm" N NOM PL
"<&>"
    "&" CC @CC
"<kissed>                                                  !!!!!
    "kiss" <SVO> <SV> PCP2
    "kiss" <SVO> <SV> V PAST VFIN @+FMAINV
"<him>"
    "he" <NonMod> PRON PERS MASC ACC SG3
"<with>"
    "with" PREP
"<these>"
    "this" DET CENTRAL DEM PL @DN>
```

```
"<wordes>"
      "word" N NOM PL
"<$:>")
```

The word *this* in ‘this yong duke’ would retain two interpretations even in analysing Present-day English data, this being how the parser deals with two parallel modifiers preceding the head. Similarly, the forms *toke* and *kissed* would remain ambiguous in analysing Present-day English, it often being safer to postpone the disambiguation of past tense and past participle readings until further post-editing stages.

Moving on a century in time, Sir John Hayward’s *Annals* (1627) proves to be manageable material for the parser (p. 5):

But on the other side, because yf the Bishop of Rome had power to dispense in the first marriage of King Henry the eyght, then was the subsequent marriage with Anne Bullen voyde; --

```
"<*but>"
      "*but" <*> CC (@CC)
"<on>"
      "on" PREP
"<the>"
      "the" <Def> DET CENTRAL ART SG/PL (@DN>)
"<other>"
      "other" DET POST SG/PL (@DN>)
"<side>"
      "side" N NOM SG
"<$,>"
"<because>"
      "because" <**CLB> CS (@CS)
"<yf>"
      "if" <**CLB> CS (@CS)
"<the>"
      "the" <Def> DET CENTRAL ART SG/PL (@DN>)
"<*bishop>"
      "*bishop" <*> <Title> N NOM SG
"<of>"
      "of" PREP
"<*rome>"
      "*rome" <*> <Proper> N NOM SG
"<had>"
      "have" <SVO> <SVO/A> V PAST VFIN
"<power>"
      "power" N NOM SG
```

```
"<to>"
      "to" INFMARK (@INFMARK>)
"<dispense>"
      "dispense" <SVO> <SV> <P/with> V INF
"<in>"
      "in" PREP
"<the>"
      "the" <Def> DET CENTRAL ART SG/PL (@DN>)
"<first>"
      "first" NUM ORD
"<marriage>"
      "marriage" N NOM SG
"<of>"
      "of" PREP
"<*king>
      "*king" <*> <Title> N NOM SG
"<*henry>"
      "*henry" <*> <Proper> N NOM SG
"<the>"
      "the" <Def> DET CENTRAL ART SG/PL (@DN>)
"<eyght>"                                                   !!!!!
      "eyght" NUM CARD
"<$,>"
"<then>"                                                    !!!!!
      "than" <**CLB> CS (@CS)
      "then" ADV ADVL (@ADVL)
"<was>"
      "be" <SV> <SVC/N> <SVC/A> PAST SG1,3 VFIN
"<the>"
      "the" <Def> DET CENTRAL ART SG/PL (@DN>)
"<subsequent>"
      "subsequent" A ABS
"<marriage>"
      "marriage" N NOM SG
"<with>"
      "with" PREP
"<*anne>"
      "*anne" <*> <Proper> N NOM SG
"<*bullen>"
      "*bullen" <*> <Proper> N NOM SG
"<voyde>"
      "voyd" A ABS
"<$;>"
```

Two minor details could be pointed out. Compiling the lexicon from a mere word list makes some context-dependent errors almost unavoidable; for instance, the form *eyght* has been included in the list of cardinal numbers though this form is an ordinal here. Similarly, the use of the word *then* is of interest in early English because the variant spelling *than* adds to its ambiguity.

5.2 Differences in early and Present-day English grammars

Difficulties did however arise. Ambiguous readings are not a problem at this level of analysis while one of the alternatives satisfies the contextual conditions. A more serious error emerges when none of the alternatives offered is the one sought.

A 500-word extract was selected from each text for closer manual checking. Many ambiguities would have remained in an analysis of Present-day English extracts as well (see examples above). These ambiguities can be reduced somewhat in the future by considering the highly significant fundamental differences between the grammars of Present-day and early English.

In the interest of future development of the grammar modules, only two major error types emerged in these 500-word extracts. Both relate to differences in the grammars, one hitting the highly polyfunctional word *that*:

Whereupon sone after that is to wit, on the friday the thirtene day of Iune many Lordes assembled in the tower, and there sat in counsaile, deuising the honorable solempnite of the kinges coronacion, of which the time appointed then so nere approched, that the pageauntes and suttelties were in making day and night at westminster, **and much vitaile killed therfore, *that* afterward was cast away** (Sir Thomas More, "The History of King Richard III" (1514–1518), p. 46).

```
"<and>"
     "and" CC @CC
"<much>"
     "much" ADV ABS
     "much" <Quant> DET POST ABS SG @QN>
     "much" <NonMod> <Quant> PRON ABS SG
"<vitaile>"
     "vitail" N NOM SG
"<killed>"
     "kill" <SVO> <SV> PCP2
     "kill" <SVO> <SV> V PAST VFIN @+FMAINV
```

```
"<therfore>"
      "therfor" ADV
"<$,>")
"<that>"                                                    !!!!!
      "that" PRON DEM SG
"<afterward>"
      "afterward" ADV
"<was>"
      "be" <SV> <SVC/N> <SVC/A> V PAST SG1,3 VFIN
"<cast>"
      "cast" <SVO> <SV> <P/in> <P/on> PCP2
"<away>"
      "away" ADV ADVL @ADVL
      "away" A ABS
```

The parser fails to offer a relative pronoun reading as an alternative: in Present-day English, the relative pronoun *that* is almost never preceded by a comma. The other major example of errors (and sometimes ambiguities) differentiating early and Present-day English grammars is the use of the word *for*. In the example below three successive instances of *for* are all interpreted as prepositions by the parser. However, in the first two instances the word functions as a conjunction in sentence-initial position, as was still possible in early English, but less recommendably so in Present-day English:

> *For* the change in Religion which then insued, and had alsoe happened not long before, was easily fore-seene by men of understanding, not onely by reasone of the consciences of the Princes, formed in them by education, but alsoe out of their particular interests and endes. *For* King Henry the eighth had taken to wife Katherine of Arragon, who had beene formerly marryed to Prince Arthur his elder brother; *for* which marriage (being within the degrees expressely prohibited in Leviticus) the Bishop of Rome gave a dispensatione. (Sir John Hayward, *Annals of the First Four Years of the Reign of Queen Elizabeth* (1627), p. 4).

```
"<*for>"                                                    !!!!!
      "*for" <*> PREP
"<the>"
      "the" <Def> DET CENTRAL ART SG/PL (@DN>)
"<change>"
      "change" N NOM SG
"<in>"
      "in" PREP
```

```
"<*religion>"
      "*religion" <*> N NOM SG
.........
"<*for>"                                                    !!!!!
      "*for" <*> PREP
"<*king>"
      "*king" <*> <Title> N NOM SG
"<*henry>"
      "*henry" <*> <Proper> N NOM SG
"<the>"
      "the" <Def> DET CENTRAL ART SG/PL (@DN>)
"<eighth>"
      "eighth" NUM ORD
"<had>"
      "have" <SVO> <SVOC/A> V PAST VFIN
"<taken>"
      "take" <as/SVOC/A> <for/SVOC/A> <SVO> <SVOO> <SV> PCP2
      "taken" <SVO>  PCP2
"<to>"
      "to" PREP
"<wife>"
      "wife" N NOM SG
"<*katherine>"
      "*katherine" <*> <Proper> N NOM SG
"<of>"
      "of" PREP
"<*arragon>"
      "*arragon" <Proper> N NOM SG
"<$,>"
.........
"<$;>"
"<for>"                                                     !!!!!
      "for" PREP
"<which>"
      "which" <NonMod> <Rel> PRON WH NOM SG/PL
"<marriage>"
      "marriage" N NOM SG
..........
```

A related use is the so-called ‘pleonastic *that*’, no longer found in Present-day English. In the following example, the parser offers ‘preposition’ or ‘subordinate conjunction’ for the first component *for* and ‘demonstrative pronoun singular’ for the second component *that*. Again, the remedy would be to incorporate this feature in the ENGCG grammar.

For instance:

So it was a marveilous motive for Queen Mary to embrace and advance the authority of the Bishop of Rome, *for that* the validity of King Henryes marryage with Queene Katherine her mother, was thereupon grounded (Sir John Hayward, *Annals of the First Four Years of the Reign of Queen Elizabeth* (1627), p. 4).

```
"<*so>"
    "*so" <*> <**CLB> CS (@CS)
"<it>"
    "it" <NonMod> PRON NOM SG3 SUBJ (@SUBJ)
"<was>"
    "be" <SV> <SVC/N> <SVC/A> V PAST SG1,3 VFIN
"<a>"
    "a" <Indef> DET CENTRAL ART SG (@DN>)
"<marveilous>"
    "marveilous" A ABS
"<motive>"
    "motive" N NOM SG
"<for>"
    "for" PREP
"<*queen>"
    "*queen" <*> <Title> N NOM SG
"<*mary>"
    "*mary" <*> <Proper> N NOM SG
.........
"<$,>"
"<for>"                                                   !!!!!
    "for" PREP
    "for" <**CLB> CS (@CS)
"<that>"                                                  !!!!!
    "that" PRON DEM SG
"<the>"
    "the" <Def> DET CENTRAL ART SG/PL (@DN>)
"<validity>"
    "validity" <-Indef> N NOM SG
"<of>"
    "of" PREP
"<*king>"
    "*king" <*> <Title> N NOM SG
"<*henryes>"
    "*henry" <Proper> N GEN SG/PL
"<marryage>"
    "marryag" N NOM SG
```

```
"<with>"
      "with" PREP
"<*queene>"
      "*queen" <*> N NOM SG
"<*katherine>"
      "*katherine" <*> <Proper> N NOM SG
"<her>"
      "she" PRON PERS FEM GEN SG3
"<mother>"
      "mother" <Title> N NOM SG
"<$,>"
"<was>"
      "be" <SV> <SVC/N> <SVC/A> V PAST SG1,3 VFIN
"<thereupon>"
      "thereupon" ADV
"<grounded>"
      "ground" <SVO> <SV> PCP2
"<$:>"
```

Our final example in this section returns us to the lexicon. The current version of the ENGCG parser conveniently treats phrases such as *as-far-as* and *least-of-all* as units. Again, for the purposes of earlier English, we should be prepared to turn the clock back: in the following example from *Burnet's History*, the phrase *to wit* points to the origin of the Present-day idiomatic phrase; both components need to be analysed independently, as preposition and noun (see OED, s.v. *wit*):

He left the business of the treasury wholly in the hands of his secretary, sir Philip Warwick, who was an honest but a weak man; he understood the common road of the treasury; but, though he pretended *to wit* and politics, he was not cut out for that, and least of all for writing of history. (Gilbert Burnet, *Burnet's History of My Own Time* (a1703), vol. 1, p. I:171).

```
.........
"<but>"
      "but" CC (@CC)
"<$,>"
"<though>"
      "though" <**CLB> CS (@CS)
"<he>"
      "he" <NonMod> PRON PERS MASC NOM SG3 SUBJ (@SUBJ)
"<pretended>"
      "pretend" <SVO> <SV> V PAST VFIN (@+FMAINV)
```

```
"<to=wit>"                                                  !!!!!
      "to=wit" DV ADVL (@ADVL)
"<and>"
      "and" CC (@CC)
"<politics>"
      "politics" <-Indef> N NOM SG/PL
"<$,>"
........
```

6. Profiting from the updated lexicon

Finally, does the new updated lexicon help to deal with further texts from the Helsinki Corpus? A reasonable test case was John Taylor’s travelogue, "Pennyles Pilgrimage", from 1630. Applying the lexicon designed for history writing, 735 word types are still new to the ENGCG parser, representing 16% of all tokens and 36% of the types not found in the Present-day English ENGTWOL dictionary; these figures approximate those obtained for Hayward, the contemporaneous historian of the late 1620s. Though the percentages do not show a drastic gain in using the new updated lexicon, a closer look at the words not recognized by the parser shows that most are either place-names or orthographic forms showing variation in the use of the Ramistic letters, ‘u’ instead of the Present-day ‘v’, and ‘i’ instead of the Present-day ‘j’. The overall result of the analysis was very good:

At last I resolu’d, that the next Gentleman that I met withall, should be acquaintance whether hee would or no: **and presently fixing mine eyes vpon a Gentleman-like obiect, I looked on him, as if I would suruay something through him, and make him my perspectiue** (John Taylor, "Pennyles Pilgrimage" (1630), p. 129).

```
"<$:>"
"<and>"
      "and" CC (@CC)
"<presently>"
      "present" <DER:ly> ADV
"<fixing>"
      "fix" <SVO> <SVOO> <SV> <P/on> PCP1
"<mine>"                                                    !!!!!
      "mine" N NOM SG
      "*i" <NonMod> PRON PERS GEN SG1 INDEP
      "mine" <SVO> <SV> <P/for> V INF
```

```
"<eyes>"
      "eye" N NOM PL
      "eye" <SVO> V PRES SG3 VFIN (@+FMAINV)
"<vpon>"
      "upon" PREP
"<a>"
      "a" <Indef> DET CENTRAL ART SG (@DN>)
"<*gentleman-like>"
      "*gentleman-like" <*> <DER:like> A ABS
"<obiect>"                                        !!!!!
"<$,>"
"<*i>"
      "*i" <*> <NonMod> PRON PERS NOM SG1 SUBJ (@SUBJ)
"<looked>"
      "look" <SVC/N> <SVC/A> <SV> <SVO> V PAST VFIN (@+FMAINV)
"<on>"
      "on" PREP
"<him>"
      "he" <NonMod> PRON PERS MASC ACC SG3
"<$,>"
"<as>"
      "as" <**CLB> CS (@CS)
"<if>"
      "if" <**CLB> CS (@CS)
"<*i>"
      "*i" <*> <NonMod> PRON PERS NOM SG1 SUBJ (@SUBJ)
"<would>"
      "would" V AUXMOD VFIN (@+FAUXV)
"<suruay>"                                        !!!!!
"<something>"
      "something" <Comp-Pron> PRON NOM SG
"<through>"
      "through" PREP
"<him>"
      "he" <NonMod> PRON PERS MASC ACC SG3
"<$,>"
"<and>"
      "and" CC (@CC)
"<make>"
      "make" <SVC/A> <SVOC/N> <SVOC/A> <into/SVOC/A> <SVO>
                 <InfComp> <P/of> <P/for> V INF
"<him>"
      "he" <NonMod> PRON PERS MASC ACC SG3
"<my>"
      "*i" PRON PERS GEN SG1 (@GN>)
```

```
"<perspective>                                    !!!!!
"<$:>"
```

In this example the analysis given for *mine eyes* seems odd: the nearest correct alternative is ‘the independent use of a genitive first person singular pronoun’. However, at this time, both the determiner and the independent forms of the first-person and second-person genitive pronouns varied freely before words beginning with vowels (Barber 1976: 207–208). Again, this could be considered in future versions of the updated lexicon.

7. Concluding remarks

In this pilot study ambiguous readings were, relatively speaking, more numerous in texts from the first subperiod than in those from the second and third. In future analyses, it would seem sensible to treat the first subperiod of Early Modern English separately, while the second and the third subperiods could be grouped together. The combined use of manual and automatic analysis from the outset is also recommended. It is reasonable to check in advance whether to introduce a rule or not, for instance, if the second person pronoun is spelt *the* instead of the usual *thee*, but as this seemed extremely rare on exploratory manual checking, it would be pointless to include this additional source of ambiguity in the analysis. Both manual and automatic methods will be needed for various post-editing stages as well.

To conclude, we regard the results of our experiment as highly encouraging. Economy suggests parsing the Helsinki Corpus starting with the later texts and proceeding by grouping the texts related by subject matter and generic conventions. Rather than one large super-grammar capable of dealing with the whole corpus in one run, we envisage several mini-grammars, devised to deal with the needs of specific subperiods.

Notes

1. This study is based on a paper given at the XV ICAME Conference in Aarhus (Denmark), 18–22 May, 1994.

2. For further details, please e-mail Atro Voutilainen at Atro.Voutilainen@Helsinki.FI, or write to Atro Voutilainen, Research Unit for Multilingual Language Technology, Department of General Linguistics, P.O. Box 4, FIN-00014 University of Helsinki, Finland.

3. For documentation of ENGCG tags, see Voutilainen *et al.* (1992) or Karlsson *et al.* (1995). Compact documentation can also be acquired via e-mail by sending an empty message to engcg@ling.helsinki.fi; the documentation comes as a reply message.

4. Note that an unambiguous analysis is not discarded, i.e. every input receives an analysis.

5. However, this remaining ambiguity can be resolved by postprocessing, either manually or by a (statistical) guess, if necessary.

6. For details about the accuracy of the ENGCG tagger, as compared to that of other state-of-the-art systems, see Voutilainen (1994) and Voutilainen and Heikkilä (1994).

7. For principles of compilation and the structure of the Helsinki Corpus of English Texts, see Rissanen, Kytö and Palander-Collin (eds.) (1993).

8. Owing to the pre-editorial changes made in texts before carrying out the ENGCG analysis, the figures for the numbers of words given here differ to some extent from those given in Kytö (1993).

References

Anttila, A. 1995. How to recognise subjects in English. In Karlsson *et al.* (1995: 315–358).

Barber, C. 1976. *Early Modern English*. London: André Deutch.

Brodda, B. 1990. A PC-oriented tool for corpus work. In Karlgren, H. (ed.) *COLING-90. Papers presented to the 13th International Conference on Computational Linguistics,* Vol. 3. Helsinki, Finland, 405–409.

Church, K. W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. *Proceedings of the Second Conference on Applied Natural Language Processing*. Austin, Texas. 136–143.

de Marcken, C. 1990. Parsing the LOB Corpus. In *Proceedings of the 28th Annual Meeting of the ACL*, 243–251.

Heikkilä, J. 1995. A TWOL-based lexicon and feature system for English. In Karlsson *et al*. (1995: 103–131).

Järvinen, T. 1994. Annotating 200 million words: the Bank of English project. In *Proceedings of COLING-94*. Vol. 1, Kyoto, 565–568.

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. 1995. Designing a parser for unrestricted text. In Karlsson *et al*. (1995: 1–40).

Karlsson, F., A. Voutilainen, J. Heikkilä, and A. Anttila (eds.). 1995. *Constraint Grammar. A language-independent system for parsing unrestricted text*. Berlin and New-York: Mouton de Gruyter.

Koskenniemi, K. 1983. *Two-level morphology: a general computational model for word-form recognition and production*. (Publications 11). Department of General Linguistics, University of Helsinki.

Kytö, M. 1993. *Manual to the Helsinki Corpus of English Texts. Coding conventions and lists of source texts*. Second Edition. Department of English, University of Helsinki.

Leech, G., R. Garside, and M. Bryant. 1994. CLAWS4: The tagging of the British National Corpus. In *Proceedings of COLING-94*. Vol. 1, Kyoto, 622–628.

OED = *The Oxford English Dictionary*. 1989. Second edition, eds. J. A. Simpson and E. S. C. Weiner. Oxford: Clarendon Press. [First edited by James A. H. Murray, Henry Bradley, W. A. Craigie, and C. T. Onions, 1888–1933.]

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Rissanen, M., M. Kytö, and M. Palander-Collin (eds.). 1993. *Early English in the computer age: explorations through the Helsinki Corpus* (Topics in English Linguistics 11). Berlin and New York: Mouton de Gruyter.

Voutilainen, A. 1994. *Three studies of grammar-based surface parsing of unrestricted English text* (Publications 24). Department of General Linguistics, University of Helsinki.

Voutilainen, A., J. Heikkilä, and A. Anttila. 1992. *Constraint grammar of English. A performance-oriented introduction* (Publications 21). Department of General Linguistics, University of Helsinki.

Voutilainen, A. and J. Heikkilä. 1994. An English constraint grammar (ENGCG): a surface-syntactic parser of English. In U. Fries, G. Tottie, and P. Schneider (eds.) *Creating and using English language corpora. Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zurich 1993*. Rodopi: Amsterdam and Atlanta, 189–199.

Voutilainen, A. and J. Heikkilä. 1995. Compiling and testing the lexicon. In Karlsson *et al.* (1995: 89–101).