Grammere = Grammar? Syntaxe = Syntax? Early Modern English = Present-day English?

Dolores González-Álvarez and Javier Pérez-Guerra
University of Vigo

Abstract

The aim of this paper is to explore grammatical variation between early Modern and Present-day English by means of computational devices. To that end, we compare the automatic output which the English Constraint Grammar Parser offers of an updated corpus of Renaissance texts and its corresponding modern version. In the first half of the paper we give information about the technical process; in particular, we focus on the description of the parser. The software parses every constituent and associates it with a tag which provides morphological information and dependency links (head-modifier/complement syntactic relations). It is also equipped with a disambiguation tool which reduces the number of the alternative morphosyntactic analyses of each lexical entry. The second half of the paper is devoted to the evaluation of the results obtained after the application of the parser to the Renaissance and the contemporary passages. Since the parser's lexicon is designed to cope with only contemporary English, orthographic, lexical and morphological pre-edition has been necessary so that the parser can deal with (an adaptation of) the Renaissance source. By examining the instances exhibiting either unjustified ambiguity or parsing failure we determine to what extent the morphosyntactic rules designed for Present-day English can be suitably applied to earlier stages of the language.

1 Introduction[1]

The aim of this paper is to determine on objective grounds to what extent the grammar of a Renaissance text differs from the grammar of contemporary English, where 'grammar' refers to the rules that govern the overt design of grammatical sentences. This approach takes for granted that such rules can be described in a computational way – we shall come back to this issue in Section 2. We assume that the computer-based analysis of the surface structure of both

early Modern English (eModE) and Present-day English (PDE) linguistic productions is revealing as regards the determination of the factors that merit attention from the point of view of linguistic explanation. If a computational grammar parser which is trained to cope with PDE also deals correctly with eModE, then one may hypothesise that there are no significant differences between the grammar (or, more precisely, syntax) of eModE and PDE. If, by contrast, such PDE-based parser fails when it is required to handle older texts, then the conclusion is that the grammars are considerably different.

What follows is organized into five sections. In Section 2 we outline the methodological issues and assumptions resorted to in the investigation of the textual material in the ensuing sections. Section 3 gives information on the corpus material. Section 4, which constitutes the backbone of this pilot study, deals with the examination of the output of the computational process which has been applied to the textual material. Finally, Section 5 puts forward the conclusions warranted by the analysis of the data in Section 4.

2 Methodology

A consequence of the assumption that un-/grammaticality[2] in speech production is governed by context-dependent rules is that (at least part of) the grammar of a given language can be thought of as a language-particular computational system. To the end of assessing the degree of similarity between the rules operating in eModE and in PDE, we have made use of the automatic parser 'Connexor Machinese Syntax' (CMS), based on a Functional Dependency Grammar (FDP) (see Järvinen and Tapanainen 1997 for the description of the parser and for a guide to Dependency Grammar), and its associated analyser ENGCG (see Voutilainen and Heikkilä 1994 or Tapanainen 1996 for the technical description of ENGCG) for PDE, both developed in Finland by Connexor.[3]

The grammar of this computational framework is derived from a constraint-based grammar. This means that the CMS parser uses constraint methodology which is mainly based on the surface analysis of the utterances and the distributional properties of the constituents, not on local statistical generalisations obtained through the exploration of large manually-tagged corpora (see Voutilainen 1994a: Sections 3.2 and 3.3 in this respect). A constraint grammar assumes that by observing exclusively both the surface structure of an utterance and the core features of its various constituents, the computational system can implement a closed list of alternative analyses of a given word, one of which is correct. The absence of reference to extralinguistic factors[4] – in, at least, the first theoretical stage –, on the one hand, and of abstract underlying syntactic repre-

sentations, on the other, have been decisive for the selection of a constraint-based framework in this study.

The key-concept in this parsing technology is thus ambiguity, which refers to the existence of multiple output analyses associated with the same utterance. As it will be shown in Section 4 when we deal with the actual texts, the CMS parser gives several solutions on many occasions, which must be understood as a consequence of the parsing process itself. Alternatively put, unless the number of ad-hoc constraints is increased, the parser will not be able to select the correct output in every case and will thus offer a number of possible analyses.[5] In a constraint grammar disambiguation is resolved by removing among the alternative analyses those which are not likely to be correct by means of constraints or negative rules.[6]

To give an example, (1) reflects the shallow parsing of the PDE sentence *He told me how Furbusher dealt with him, very headily sure* as given by the CMS parser:

(1)

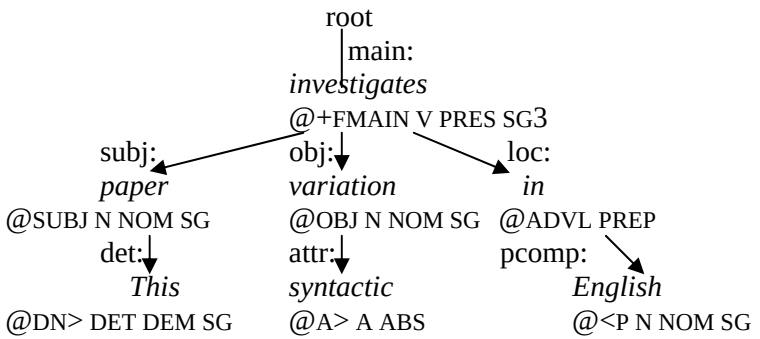
Text	Baseform	Syntactic relation	Syntax and morphology
1 He	he	subj:>2	@SUBJ %NH PRON PERS NOM SG3
2 told	tell	main:>0	@+FMAINV %VA V PAST
3 me	i	dat:>2	@I-OBJ %NH PRON PERS ACC SG1
4 how	how	man:>6	@ADVL %EH ADV WH
5 Furbusher	furbusher	subj:>6	@SUBJ %NH <?> N NOM SG
6 dealt	deal	obj:>2	@+FMAINV %VA V PAST
7 with	with	phr:>6	@ADVL %EH PREP
8 him	he	pcomp:>7	@<P %NH PRON PERS ACC SG3
9,	,		
10 very	very	ad:>11	@AD-A> %E> ADV
11 headily	headily	ad:>12	@AD-A> %E> ADV
12 sure	sure		@ADVL %EH ADV
			@<P %NH A ABS
			@OBJ %NH A ABS
			@APP %NH A ABS

The on-line parser has not been able to disambiguate among the tags assigned to the word *sure* in (1), namely, adverb (ADV) or adjective (A) – complement of a preposition (<P), of an object (OBJ) or an apposition (APP).[7]

In this study ambiguity is not considered disadvantageous at all since the shallower the parsing process and the lower the number of ad-hoc constraints at work, the better. Put differently, we are interested in the application of a computational technique which is not based on non-systematic language-particular rules; otherwise we will not be able to discern whether the different outputs obtained for eModE and for PDE are due to language-internal, i.e. systematic, factors or whether they are reflections of the capricious historical behaviour of the language. In an attempt to eliminate non-systematic ad-hoc machinery we have opted for the rather shallow level of morphosyntactic analysis offered by the on-line CMS parser (www.connexor.com), which excludes the application of powerful disambiguation.

Apart from assigning morphosyntactic tags with information about lexical and phrasal categorisation to the lexical material, the CMS parser shows relations between words, as marshalled by (2) [our graphical adaptation] (see Voutilainen 1994b for examples of syntactic analyses):

(2) This paper investigates syntactic variation in English.



In the graphical output, the arrows indicate syntactic relations of modification between heads and, say, satellites. In a dependency grammar, "every element of the dependency tree has a unique head [and] the verb serves as the head of the clause" (Tapanainen and Järvinen 1997: 65). That stated, the representation in (2) must be interpreted as follows: the main element – the verbal form *investigates* (finite main predicator, present, third-person singular) – is modified by three functional constituents, namely, the subject, the object and a locative segment, each of them consisting of functional heads – *paper* (noun, nominative, singular), *variation* (noun, nominative, singular) and *in* (adverbial, preposition),

respectively – and their corresponding modifiers – *this* (premodifying determiner, demonstrative, singular), *syntactic* (premodifying adjective, absolutive or uninflected for comparison) and *English* (post-head complement of a preposition, noun, nominative, singular).

So far we have offered details about the computational process of shallow parsing, as offered by the CMS parser. The system is capable of discriminating among alternative surface analyses of lexical items and idiomatic expressions with the assistance of a language-particular lexicon or dictionary,[8] a basic morphosyntactic analyser and a robust constraint grammar containing negative rules or constraints whose goal is to eliminate superfluous and incorrect analyses. It seems in order here to stress that the whole process is independent of artificial theoretical rules which are disconnected from the actual surface of the material acting as the input.

3 The corpus

In this pilot study we utilise the CMS system so as to examine the results of its application to two versions of the same text, namely, pages 79.27 to 89.9 and 128.21 to 145.20 of *An Elizabethan in 1582: The Diary of Richard Madox, Fellow of All Souls*, totalling approximately 5,000 running words,[9] an electronic version of which can be extracted from the *Helsinki Corpus of English Texts* (see Kytö 1996). Further investigation into a larger corpus will serve for the purposes of both the corroboration of the results in this research project (see Section 4 in this respect) and the provision of more grammatical information about the intricacies of the grammar of eModE.

The selection of *The Diary of Richard Madox* was not a random but a meditated choice. On the one hand, we wanted a text in prose with no literary aims and not greatly affected by generic conventions in an attempt to avoid some of the problems caused by distance in time between eModE and PDE. On the other hand, investigating linguistic issues in *The Diary of Richard Madox* would allow us to trace parallelism between this study and González-Álvarez and Pérez-Guerra (2004).

The goal of this pilot study is to check whether a parser which (i) has been fully trained for the analysis of PDE texts, and (ii) is based on surface rules and not on abstract axioms will do successfully when applied to older texts or not. As already mentioned in Section 1, if the result is positive then one might argue that there are no significant differences between the grammar of eModE and PDE. If, by contrast, the degree of success is considerably lower with the eModE texts, one should conclude that the whole computational system should

be rebuilt, which indicates that the grammars of the two periods under examination are different.

To our knowledge, the only studies which have applied the CMS parser to, respectively, early Modern English historic texts and late Modern English letters are Kytö and Voutilainen (1995) and (1998). The perspective adopted in their investigation is somewhat different from ours. Kytö and Voutilainen's goal is twofold: on the one hand, they want to adapt (in their terminology, 'teach') the parser so that it can handle older textual productions. To that end, they both increased the lexicon and built a specialised grammar on top of the basic analyser containing new constraints, which discards most of the ambiguity. On the other hand, they wonder "to what extent does Present-day English differ from early English and to what extent is it possible to formalize this difference for the purposes of the parser" (1998: 149). In this paper we will be referring exclusively to their second objective, namely, the determination of the systematic differences between a computational grammar which copes successfully with PDE texts and the potential constraints which should be operative in a grammar for eModE productions.

Our starting point is thus an eModE text, which must serve as the input material for the computational process. Even though the parser is claimed to be robust, that is, capable of handling unedited text, we have edited the input sample by simply updating the spelling and adapting to PDE the inflectional endings and the medieval lexicon which is no longer used in English so that the parser will not fail in the analysis of the text due to the impossibility of interpreting the material. Thus, to give a few examples, *supt* in the Renaissance manuscript has had to be translated by the corresponding PDE term *drank*. Likewise, verbal forms such as *hath* or *beginneth* were rendered as *has* and *begins*. An example of the adaptation is shown as follows: whereas (3) contains a passage from page 140 of the Renaissance sample, which will not be able to undergo automatic parsing at all, (4) offers the updated version MADOX1 with which the parser will be confronted in this research project:

(3) M. Walker and I went thither purposing to have walked
only, but M. leiftenent which was now come from Sir
Fraunces Drake at Bucland had us to M. Whoodes howse
wher we supt with M. Whyticars hath maried M.
Hawkins syster, and after we returned to the Edward
wher we discoursed with the viceadmirall of many mens
maners and many matters, advising how love myght best
be maynteyned and good order kept, but wher overweening

pevishnes is once planted, and myxed with a kynd of creeping dissimulation, yt is hard ther to setle the seeds of any good advice, for now beginneth the hydden poyson to breth owt.

(4) Mr. Walker and I went thither purposing to have walked only, but Mr. lieutenant which was now come from Sir Francis Drake at Bucland had us to Mr. Whoodes house where we drank with Mr. Whyticars has married Mr. Hawkins' sister, and after we returned to the Edward where we discoursed with the vice-admiral of many men's manners and many matters, advising how love might best be maintained and good order kept, but where overweening peevishness is once planted, and mixed with a kind of creeping dissimulation, it is hard there to settle the seeds of any good advice, for now begins the hidden poison to breath out.

The second textual source which will serve as the basis of comparison and contrast will be a literal modern correct/acceptable version of the adapted material in (4), here illustrated by way of (5):

(5) Mr. Walker and I went thither purposing to have walked alone, but Mr. lieutenant, who had now come from Sir Francis Drake at Bucland, led us to Mr. Whood's house where we drank with Mr. Whyticars, who has married Mr. Hawkins' sister, and later we returned to the Edward where we discoursed with the vice-admiral about many men's manners and many matters, advising how love might best be maintained and good order kept, but where overweening peevishness is once planted, and mixed with a kind of creeping dissimulation, it is hard to settle there the seeds of any good advice, for now the hidden poison begins to breath out.

If we compare (4) and (5, or MADOX2), we will observe, for example, that *alone* and *who* in the PDE version substitute for, respectively, *only* and *which* in the updated text, that a relative proform *who* had to be added in *where we drank with Mr. Whyticars, who has married Mr. Hawkins' sister*, that punctuation marks have had to be incorporated in *Mr. lieutenant, who had now come from Sir Francis Drake at Bucland, led us to Mr. Whood's house*, or that word order

has had to be adapted to contemporary standards (*now the hidden poison begins to breath out* for *now begins the hidden poison to breath out*).

In this section we have given some details about the corpus and the edition of the text, necessary for the subsequent computational treatment by the CMS parser. In the ensuing section we describe the results produced by the parser on the two versions of the original medieval source, namely, MADOX1, which coincides with the Renaissance text except for minor changes in the orthography and the lexicon, and MADOX2, which is perfectly grammatical (and mostly acceptable) in PDE.

4 Parsing of the Renaissance and the modernised versions

This section focuses on the output offered by the parser for the eModE and PDE versions of the same texts. Before discussing the differences between the resulting analyses, in Section 4.1 we concentrate on some constructions and syntactic contexts which lead to the failure of the parser both with the older and with the modernised textual input. In Section 4.2 we hypothesise on the reasons that led the parser to offer different analyses for, respectively, the eModE and the PDE (or modernised) texts, and will argue that such differences constitute the bases of a (contrastive) grammar of the older period, with consequences for all the levels of syntactic constituency (word, phrasal and constructional levels). Finally, Section 5 contains some final remarks which round off our discussion on the applications of parsing software designed for contemporary English to older textual material.

4.1 Failures of the parser in MADOX1 and MADOX2

Although the purpose of this paper is to ascertain to what extent Renaissance English differs from PDE by analysing the capacity of a PDE-based parser when it is required to cope with eModE data, we shall first consider a number of performance failures of the software, which suggest that some internal rules of the parser need further elaboration. It is needless to say that the degree of success of the parser is very high despite these failures. In what follows, we shall pay attention to parsing difficulties in the treatment of syntactic relations of modification, in the determination of the syntactic functions of some prepositional phrases, in the analysis of particles in phrasal verb groups, *make*-sentences, relativisers (*that* in particular), in contexts of coordination, etc.

First of all, the parser has serious problems when trying to identify the head of modifiers. In (6), for example, it is unable to identify the head of the two

prepositional phrases introduced by *with*, while in (7) the relative clause is analysed as a modifier of *Bel* and not of *Mr Creswels*. Another case in point is the analysis of infinitival clauses depending on other constituents; to give an example, in (8), the infinitive is said to modify solely the immediately preceding word, that is, *carpenters*, not the main object:

(6) Mr Banester hunting for the votes of the most vain masses **with dinner expenses and gifts** [...] had drawn out a sheet of paper for to be set on the main mast with prayers for morning and evening [MADOX1]

(7) We dined and lay at Mr. Creswels of the Bel **who made unto us many a substantial lie** [MADOX1]

(8) and did press a tinker and two carpenters **to go** with us [MADOX1]

The parser also has difficulties with the analysis of *in* and *to*. In the former case it often fails to distinguish between the locative function and other functions of the preposition *in*, as evinced by (9), in which *in* is wrongly labelled as a locative. Examples (10) to (12) illustrate different faulty analyses of *to*. In (10) *to us* is analysed as a dative, in (11) *to Newport* is interpreted as a modifier of *Tobias*, while in (12) the parser treats *to* as a preposition:

(9) he would put them **in** fear of the frenzy [MADOX1]

(10) Mr. Brown and Mr. Baker [...] came **to** us [MADOX1]

(11) I [...] went with Mr. Walker, Mr Lewis and Mr Tobias **to** Newport [MADOX1]

(12) The wind began **to** fresh up [MADOX1]

Another category which often triggers wrong analyses is the particle in a phrasal verb, sometimes analysed as a preposition, sometimes as an adverb. Thus, *over* in (13) is wrongly interpreted as a manner adverb, while *off* in (14) is incorrectly analysed as a preposition governing the NP *a piece*:

(13) whom the master combed **over** for losing his sounding lead [MADOX1]

(14) the Elisabeth being behind shot **off** a piece and struck sail [MADOX1]

The analysis of *make* and its complements is another source of incorrect interpretations. In examples like (15) *make* is repeatedly analysed as a complex transitive verb:

(15) where Banester with his Robin-hood rhymes **made** us good sport [MADOX1]

The analysis of the multifunctional word *that* also causes difficulties to the parser, which is often unable to disambiguate between the relativiser, the complementiser and the demonstrative uses of this item. Thus, in both (16) and (17) *that* is incorrectly interpreted as a subordinating conjunction introducing a noun clause:

(16) When I see **that** I plead as in Arte Poetica [MADOX1]

(17) He told me of many **that** he had occupied [MADOX1]

The parser is unable to cope with sentential relative clauses, whose relativisers are invariably analysed as modifiers of the immediately preceding word (example (18)), and with inverted conditionals (as in (19)). These latter lead to the total collapse of the parser when it tries to account for the grammatical status of the constituents involved in the construction. Just as a token, *that* in (19) is analysed as a pronoun functioning as the subject of *had*, *he* is interpreted as the subject of *supposed*, while *supposed* is analysed as a finite main verb to which the parser is unable to assign any dependency relationship:

(18) When we were come to Hurst Castle the Elisabeth being behind shot off a piece and struck sail **which put us in a doubtful marvel**, [MADOX1]

(19) The master told me that **had he supposed the voyage would have turned to pilfering** [...] he would not have undertaken it [MADOX1]

Finally, coordination is one of the main triggers of faulty analyses, particularly when the distance between the coordinators is large. Thus, in (20) *signs* is said to be coordinated with *evening* and not with *prayers*. Similarly, in (21) the parser understands that *and* coordinates *5* and *the Moluccas* and is thus unable to assign any function to *were cast*. (22) is an illustration of a linguistic context in which the parser cannot cope with coordinates belonging to different categories, in this case a prepositional phrase and a clause. Finally, example (23) illustrates its inability to handle ellipsis:

(20) Mr Banester [...] had drawn out a sheet of paper for to be set on the main mast with prayers for morning and evening **and** signs to know when they should be sick [MADOX1]

(21) He told that the King of Spain had sent 8 ships to the Moluccas **and** 5 were cast away on the coast of Barbary. [MADOX1]

(22) We did also sharply rebuke Muns the master for his disloyal pride **and** because he went about to discourage some of our men for the voyage. [MADOX1]

(23) Wednesday morning we found ourselves in front of Lyme **and** the next tide in front of Exmouth. [MADOX1]

4.2 Parsing differences between MADOX1 (eModE) and MADOX2 (modernised version)

The differences between the analyses produced by the parser affect the whole scale of linguistic categorisation, namely, the lexical, phrasal and sentential levels. The description of the linguistic phenomena triggering different analyses will thus be organised according to the syntactic layers (subphrasal in Section 4.2.1, phrasal in Section 4.2.2 and supraphrasal in Section 4.2.3) on which the parser has acted distinctly. In Section 4.2.4 we will finally include a separate section devoted to punctuation, which constitutes one of the main triggers of mistaken outputs in MADOX1.

4.2.1 Subphrasal level of analysis

At this level, we have come across different types of disparities between the analysis of the eModE and PDE texts: lexical issues related to the lexicon operative in both stages of the history of the English language (Section 4.2.1.1), variation in the categorisation of lexical items and expressions (Section 4.2.1.2), verbal subcategorisation (Section 4.2.1.3) and punctual changes in the paradigm of grammatical classes (Section 4.2.1.4).

4.2.1.1 Lexical issues

At the lexical level we have found some mistakes in the analyses due to differences in the inventory of idiomatic collocations in eModE and PDE. An example of this is the different behaviour of the parser with the collocation *at the least* and its updated version *at least*. The parser is unable to cope with *at the least* in (24), whereas it correctly interprets *at least* in MADOX2 as an idiomatic adverbial. It goes without saying that lexical issues illustrated by collocations like these have no consequences for the grammatical system the parser is built upon, whose analysis constitutes the focus of this study:

(24) My lord Foster being a little drunk went up to the main top to fetch down a rebel and 20 **at the least** after him [MADOX1]

4.2.1.2 Grammatical categorisation

The results of the parser in those cases in which the category of a linguistic item in a construction changes across time provide an interesting insight into the nature of the process of linguistic classification in the periods under investigation (see, in this respect, González-Álvarez 2002: 181). In what follows we shall discuss examples of pairs of categories which undergo modification from eModE to PDE.

First, examples such as (25) and (26) illustrate the use of morphologically unmarked forms as adverbs. The parser tackles the morphological analysis of these examples in a correct way but it fails in their syntactic parsing, which is based on rules valid for PDE. *Marvellous* is analysed as an attributive adjective qualifying *negligent*, not as an intensifier, in (25), whereas *fair* in (26) is analysed as an adjective functioning as a predicative complement, not as a manner adverbial:

(25) but is **marvellous** negligent and bold [MADOX1]

(26) although he speak me **fair** yet [...] [MADOX1]

Second, example (27) illustrates the use of *after* as a temporal adverb in eModE, which is replaced by the adverb *afterwards* in the modernised version. Only in MADOX2 does the parser ascribe the adverb to the correct category, that is, time adverbial:

(27) **After** were we so encumbered with shore-haunters that [...] [MADOX1]

A third instance of wrong categorial ascription concerns *beside*, an item which can function as a conjunction in eModE, as (28) illustrates:

(28) Mr. Banester [...] had drawn out a sheet of paper for to be set on the main mast with prayers for morning and evening and signs to know when they should be sick which **beside** it was immeasurably beyond all modesty, the conceit was also so gross that [...] [MADOX1]

The parser is only trained to interpret *beside* as a preposition, as it actually does in this case, and thus fails to analyse it as a conjunction. Its replacement with *although* in MADOX2 yields the expected results.

4.2.1.3 Verbal subcategorisation

Some verbs have undergone diachronic changes in their subcategorisation or argument frame. The parser is, in principle, able to deal with subcategorisation differences because it does not rely on a strict lexicon of selectional restrictions

valid uniquely for PDE. However, it fails in a number of instances, admittedly very few, as illustrated by (29):

(29) Mr. Colman who was Mr. Wolley's man came with a broad seal to **stay** Mr. Boze [MADOX1]

In this example the parser is unable to provide a transitive interpretation for *stay* and treats it as an intensive verb, thus assigning the function of subject complement/predicative to *Mr Boze*. The replacement of *stay* with a transitive verb like *fetch* warrants a correct analysis.

4.2.1.4 Other changes affecting the paradigms

A paradigm which has undergone considerable changes in the course of the eModE period is that of relative and *wh*-forms (Barber 1997: 209–216; Rissanen 1999: 293–299). Example (30) illustrates the use of *which* with a human antecedent. Such use leads to the total collapse of the parser, which is not able to assign *which* a correct label and parses it as a determiner without any associated function. The parser also fails to analyse *that* in syntactic terms and is incapable of depicting the syntactic dependency holding between *go* and *concluded* and between the *which*-clause and the main clause. Once *which* has been replaced by *who* in the modern version, the syntactic analysis becomes quite felicitous:

(30) Captain Parker concluded **that** he **which** could endure the Irish service and please my Lord of Aburgeny might go for a soldier and a servingman in any place of England [MADOX1]

It is fair to point out here that the parser's grammar is able to cope with other uses which are no longer possible in PDE, such as the use of the relativiser *that* in non-defining relative clauses, illustrated by (31a). The parser's outputs for (31a) and (31b) are identical:

(31) a. At supper we talked of tattlers and counted Hearle, **that** betrayed Madder but a knave as is Nichols [MADOX1]

b. At supper we talked of tattlers and counted Hearle, **who** betrayed Madder, a knave, as is Nichols [MADOX2]

4.2.2 Phrasal level of analysis

Here the main difference concerns the structure of negative verb groups. The PDE-based constraints ruling the syntactic structure of negative verbal groups (see, among others, Ellegård 1953: 161–162; Tieken-Boon van Ostade 1987: 228; Barber 1997: 193–196 or Rissanen 1999: 239–248, 269–277) are not in keeping with eModE examples such as (32a), (33a) and (34a), with negative ver-

bal constructions with no auxiliaries. In (32a) the parser is incapable of assigning a function to *whom*, while the modified version offers the correct analysis:

(32) a. Here lost we again our tinker and a carpenter and I **know not** whom else [MADOX1]

b. Here we lost again our tinker and a carpenter and I **do not know** whom else [MADOX2]

In (33a), unlike (33b), the parser does not recognise *supped* as the main verb of the clause:

(33) a. Mr. Captain Ward **supped not** with us [MADOX1]

b. Mr. Captain Ward **did not sup** with us [MADOX2]

Finally, in (34a) *not* is analysed as a negative particle without any associated syntactic function, whereas in the modernised version *not* is plausibly taken as the negator modifying the auxiliary *do*:

(34) a. I know he loves me **not** [MADOX1]

b. I know he does **not** love me [MADOX2]

4.2.3 Supraphrasal level of analysis

At the clause/sentence level we shall focus on infinitive clauses (Section 4.2.3.1) and on word-order issues (Section 4.2.3.2).

4.2.3.1 For to *as an infinitive marker*

Infinitive clauses are no longer introduced by the preposition *for*, which was a possibility in older stages of the English language, mainly (though not exclusively) in clauses functioning as purpose adverbials (Rissanen 1999: 309). This is the reason why the parser is not able to decide on the status of *for* and the subsequent infinitive clause in examples like (35):

(35) Mr. Banester [...] had drawn out a sheet of paper **for** to be set on the main mast [MADOX1]

Once the preposition has been discarded in the modernised version, the assignment of functional labels is done correctly since *set* is correctly interpreted as a modifier of *sheet of paper.*

4.2.3.2 Issues of word order

That the surface word order of the sentence strongly conditions the output of the parser in many respects is shown by the ensuing facts. First, we will consider the

consequences which the discontinuity of constituents has for the analysis of MADOX1. To give an example, in (36), in which the place adjunct *in the Frances* plus the relative clauses initiated by *where* are split by the occurrence of the prepositional phrase *with Captain Drake*, the parser wrongly treats *Drake* as the antecedent of *where*:

(36) We dined in the Frances with Captain Drake where we had good cheer and good friendly welcome [MADOX1]

Likewise, in (37a) the two clauses are analysed as coordinated, since the parser's grammar is unable to cope with relative clauses which are not immediately preceded by their antecedents. The modernised version in (37b) yields the correct analysis:

(37) a. We supped in the Elisabeth with the vice-admiral also, where Captain Skevington made us good cheer [MADOX1]

b. We supped also with the vice-admiral in the Elisabeth, where Captain [...] [MADOX2]

A second word-order issue which affects the parser's results is its preference for postverbal adverbials. Thus, when an adverbial occurs before the verb in a position which is not canonical in PDE, the parser produces chaotic analyses of the ensuing constituents, as in (38a), in which the initial placement of the locative adverbial *there* triggers an existential interpretation. In MADOX2, where *there* has been moved to post-verbal position, *there* is correctly interpreted as a locative adverbial:

(38) a. when the ebb came we fell down to Yermouth and **there** anchored [MADOX1]

b. when the ebb came we fell down to Yermouth and anchored **there** [MADOX2]

A third major difference between the two versions as far as word order is concerned is the placement of adverbials between the verb and the object. As is well accredited in the literature (Quirk *et al*. 1985: §8.22, 498–500; Biber *et al*. 1999: 771; Huddleston *et al*. 2002: 780), the placement of (non-parenthetical) adverbials between verbs and objects is prohibited in PDE. The parser's contemporary grammar is thus induced to offer a mistaken analysis in examples such as (39a), in which the parser fails to analyse *coming* as the complement of the preposition *at* and *volley* as the object of *gave*. Desirably, the analysis of (39b), with no intervening adverbial, is correct:

(39) a so that she gave at her coming a gallant volley of shot for an homage [MADOX1]

b. so that at her coming she gave a gallant volley of shot for an homage [MADOX2]

Fourth, a further case of wrong analysis due to diachronic changes in word order is subject-verb inversion after initial adverbials:

(40) a. After were we so encumbered with shore-haunters that [...] [MADOX1]

b. Afterwards we were so encumbered with shore-haunters that [...] [MADOX2]

Thus, in (40a), unlike (40b), the parser is unable to identify *were* as the main verb of the clause. This is, however, not always the case. To give an example, in (32a) above subject-verb inversion does not cause any problems for the parser.

Finally, another word-order feature which brings about differences between the eModE and the modernised version is the relative order of direct objects and periphrastic indirect objects. The PDE-based constraints ruling the relative order of direct and indirect objects are incompatible with arrangements such as those in (41a), which were found well into the eModE period (Rissanen 1999: 268) since, as is well known, when the analytic dative first develops, almost all orders are possible (Fischer 1992: 381–382):

(41) a. Our general gave **to all the ships** very necessary instructions for the voyage [MADOX1]

b. Our general gave very necessary instructions for the voyage **to all the ships** [MADOX2]

In (41a) *to all the ships* is assigned the interpretation of heuristic adverbial. The analysis of (41b), in which the direct object precedes the prepositional dative, is correct.

4.2.4 Punctuation

The major differences between the outputs obtained for the eModE text and its modernised version are indeed due to punctuation. Differences in the conventions for punctuation in eModE and in PDE (see, among others, Salmon 1999 or Görlach 2001: Chapter 3) lead to the collapse of the parser when it has to deal with examples such as (42) to (47), discussed below.

In (42a), the parser treats the time clause as a modifier of *set*, analyses *cried* as the complement of *was* and does not know which constituent is being coordinated by *but*. The insertion of commas before and after the temporal clause, as illustrated in (42b), avoids such misinterpretation:

(42) a. wherefore he was set on shore in the Wight and when he was there he cried unto the boat gang to take pity on him and to take him back without his chest but they refused. [MADOX1]

b. wherefore he was set on shore in the Wight and, when he was there, he cried unto the boat gang to take pity on him and to take him back without his chest, but they refused. [MADOX2]

In (43a) *Brown* and *preachers* are analysed as coordinated subjects and both *Mr* and *Baker* are analysed as modifiers of *preachers*; MADOX2 yields the correct analysis, in which *Brown* and *Baker* are coordinated and *preachers* is an apposition to both:[10]

(43) a. Mr. Brown and Mr Baker preachers with the bailies of Newport came to us [MADOX1]

b. Mr. Brown and Mr Baker, preachers with the bailies of Newport, came to us [MADOX2]

In (44a), the parser does not analyse the relative clause as a modifier of *ship*. As already suggested in the previous example, the addition of commas before and after the relative clause, here in (44b), yields the expected results:

(44) a. the king of Portugal's ship **which lay at Meedhole** was likely to be stolen away by the knaves [MADOX1]

b. the king of Portugal's ship, **which lay at Meedhole**, was likely to be stolen away by the knaves [MADOX2]

In (45a), the parser is unable to identify the main verb, while in (45b), the parser correctly analyses *came* as 'MAIN>0', that is, as the matrix predicator:

(45) a. Mr. Hawkins of Plymouth **riding to London** came to us. [MADOX1]

b. Mr. Hawkins of Plymouth, **riding to London**, came to us. [MADOX2]

In (46a), *up* is wrongly treated as a preposition and *caused* as a prepositional complement. With the insertion of a comma after *up* in the modern version, the assignment of functional labels is done correctly:

(46) a. The wind began to fresh **up** which caused us to weigh upon the ebb [MADOX1]

b. The wind began to fresh **up**, which caused us to weigh upon the ebb [MADOX2]

Finally, the parser is unable to cope with vocatives such as the one in (47a) when no commas are provided; (47b) offers the correct interpretation:

(47) a. but now **sir** he has no skill in medicine [MADOX1]

b. but now, **sir**, he has no skill in medicine [MADOX2]

It is nonetheless worth mentioning that adequate punctuation does not always bring about a better analysis. In (48a) the parser correctly interprets the *having*-clause as a modifier of *Edward Horsey*, while in (48b) it is unable to establish any syntactic dependency at all between the constituents just mentioned:

(48) a. Sir Edward Horsey **having complained to our general that the king of Portugal's ship** [...] [MADOX1]

b. Sir Edward Horsey, **having complained to our general that the king of Portugal's ship** [...] [MADOX2]

Similarly, in (49b), the parser fails to assign the appropriate function to *one of the Azores*, in spite of the comma:

(49) a. We heard that Captain Laundrey and the French had taken St. Michaels, **one of the Azores in behalf** [...] [MADOX1]

b. We heard that Captain Laundrey and the French had taken St. Michaels, **one of the Azores in behalf** [...] [MADOX2]

Finally, the parser offers a correct analysis of the relativiser in (50a), whereas in (50b), in which the relative clause is between commas, it is unable to identify the antecedent of the relativiser, and is incapable of depicting the syntactic dependency holding between *which* and *saw*:

(50) a. There was a small comet **which** I saw 8 days ago in the breast of Erychtonius [MADOX1]

b. There was a small comet, **which** I saw 8 days ago in the breast of Erychtonius [MADOX2]

5 Final remarks

Our aim in this paper was to ascertain the extent and the nature of grammatical variation between eModE and PDE and not so much how the Constraint Grammar Parser should be 'taught' so as to cope with the language of earlier periods.

Though, admittedly, the parser was able to handle a large proportion of the Renaissance text, much more than in the case of late Middle English, as investigated in González-Álvarez and Pérez-Guerra (2004), the degree of success achieved is lower than with the PDE version, thus indicating that the grammars of the two periods are still significantly different at some points.

The differences found have been shown to affect all the levels of linguistic categorisation: lexical, phrasal and sentential. As regards the subphrasal level, the main divergences between eModE and PDE analyses derive from (i) the recategorisation of several lexical items, (ii) the alterations in verbal subcategorisation, and (iii) the changes in paradigm membership. The differences at the phrasal level are, however, not so striking as one might expect if one takes into account the significant changes that have affected the verb group in eModE. Such changes have few consequences for the behaviour of the parser because it simply analyses the surface structure of the utterances. Negative verb phrases stand out among the changes that cause diverging analyses at this level. As for the clause- or sentence-level, the evidence drawn from the parser's outputs suggests that the most extensive changes at this level concern the organisation of infinitival clauses and word-order strategies. Finally, differences in the conventions for punctuation in the two periods proved to be one of the main triggers of faulty analyses in the eModE text.

To conclude, we regard the results of this pilot study as highly encouraging. The vast amount of linguistic data which we have obtained out of a limited number of words opens the use of the parser to methodological applications. One such application may well be the teaching of early stages of the language. As a teaching tool, the outputs offered by the Constraint Grammar Parser on historical data may be used not only as a resource for information but also as a medium for student-centred discussion.

Notes

1. The research reported has been funded by the Spanish Ministry of Education and Science, grant number HUM2005–02351/FILO, which is hereby gratefully acknowledged.
2. 'Grammaticality' is here understood as a deductive rather than as a inductive (or 'generative'; see Järvinen and Tapanainen 1997: 5) characteristic of

linguistic explanation. Alternatively put, the grammatical component of the language will assume that the linguistic productions which act as the input of the grammatical parser are grammatical and, in consequence, the grammar of the language will have to develop a set of rules with which the utterances are conformant.

3. More information on the CMS parser can be obtained at http://www.connexor.com. ENGCG has been used for the morphosyntactic annotation of several millions of words of the *Bank of English*. In Voutilainen and Heikkilä (1994), the ENGCG is used for the analysis of 12,548 words from *The Independent* (1990).
4. It must be remarked that the system is not absolutely context-dependent since it houses heuristics which analyse by prediction the unanalysed words (see Voutilainen 1994b: Chapter 6 in this respect).
5. The proportion of ambiguous morphosyntactic interpretation once the disambiguator has been applied to the preliminary output, as reported by Tapanainen and Järvinen (1997: 70), is not alarming at all – 3.2 per cent –, whilst the percentage of the presence of the correct morphosyntactic analysis in the output of ENGCG and the FDP parser (the so-called 'success rate') is 97 per cent.
6. An illustration of a basic syntactic (negative) constraint is given in, for example, Tapanainen and Järvinen (1997: 65): REMOVE (@I-OBJ) IF (*–1C VFIN BARRIER SVOO LINK NOT 0 SVOO) – the declaration is simpler in the FDP model. The goal of this rule is to discard (REMOVE) the analysis of a given constituent as the indirect object (@I-OBJ) when, first, there are no ditransitive verbs (SVOO) between the first (1) constituent to the left (–) which is clearly (C) a finite verb (VFIN) and the (supposed) indirect object, and, second, the very same verb under examination does not subcategorise for indirect objects (LINK NOT 0 SVOO). Not every constraint belongs to the REMOVE-type. An example of a, say, 'positive' morphological constraint or rule is: "<past>" =! (DET) (–1C DET) (1C NUM) (2C NPHEAD), which implies that *past* is a determiner when: (i) the first word (1) to the left (–) is unambiguously (C) a determiner (DET), (ii) the first word to the right is a numeral (NUM), or (iii) when the second word (2) to the right is a noun or pronoun (NPHEAD).

   In view of these constraints, one realises that the only linguistic operations at work are, on the one hand, positional syntax (distributional circumstances) and rough verbal feature-checking, on the other. In this connection, it must be mentioned that feature-checking is, to a certain extent, incompatible with the theoretical basis of our investigation, since it draws on the existence of some level of abstract description in which we are not inter-

ested in this study. To the end of comparing surface (eModE and PDE) grammars, the ideal parser would be one which does not rely on the subcategorisation of the constituents at all.

7. The tags and their grammatical justification are based on Quirk *et al.*'s (1985) grammar of English.
8. The lexicon of the ENGCG system, so-called ENGTWOL, is based on corpora (*Brown*, *LOB*) and dictionaries (*Longman Dictionary of Contemporary English*, *Collins COBUILD Language Dictionary*).
9. Kytö and Voutilainen investigate 32,700 and 37,000 running words in their (1995) and (1998) papers, respectively.
10. It is fair to point out here that the parser is not always flawed when it tries to identify appositions not separated by commas, even though failure is the usual tendency. To give an example, in *Thomas Cley the carpenter, the carpenter* would be correctly analysed as an apposition.

References

Barber, Charles. 1997. *Early Modern English.* 2nd edition. Edinburgh: Edinburgh University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson Education.

Ellegård, Alvar. 1953. *The auxiliary* do*: The establishment and regulation of its use in English*. Stockholm: Almqvist & Wiksell.

Fischer, Olga C. M. 1992. Syntax. In N. F. Blake (ed.). *The Cambridge history of the English language. Volume II: 1066–1476*, 207–408. Cambridge: Cambridge University Press.

González-Álvarez, Dolores. 2002. *Disjunct adverbs in Early Modern English: A corpus-based study.* Vigo: University of Vigo (Servicio de Publicacións).

González-Álvarez, Dolores and Javier Pérez-Guerra. 2004. Profaning Margery Kempe's tomb or the application of a constraint-grammar parser to a Late Middle English text. *International Journal of Corpus Linguistics* 9(2): 225–251.

Görlach, Manfred. 2001. *Eighteenth-century English*. Heidelberg: Winter.

Huddleston, Rodney, Geoffrey K. Pullum and Laurie Bauer. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Järvinen, Timo and Pasi Tapanainen. 1997. *A dependency parser for English*. Helsinki: University of Helsinki (Department of General Linguistic, technical report TR-1).

Kytö, Merja (comp.). 1996. *Manual to the diachronic part of the Helsinki Corpus of English Texts. Coding conventions and lists of source texts*. Helsinki: University of Helsinki (Department of English).

Kytö, Merja and Atro Voutilainen. 1995. Applying the constraint grammar parser of English to the Helsinki Corpus. *ICAME Journal* 19: 23–48.

Kytö, Merja and Atro Voutilainen. 1998. Backdating the English Constraint Grammar Parser for the analysis of English historical tracts. In R. M. Hogg and L. van Bergen (eds.). *Historical linguistics 1995. Selected papers from the 12th International Conference on Historical Linguistics, Manchester, August 1995*, 149–166. Amsterdam: John Benjamins.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Harlow: Longman.

Rissanen, Matti. 1999. Syntax. In R. Lass (ed.). *The Cambridge history of the English language. Volume III: 1476–1776*, 187–331. Cambridge: Cambridge University Press.

Salmon, Vivian. 1999. Orthography and punctuation. In R. Lass (ed.). *The Cambridge history of the English language. Volume III: 1476–1776*, 13–55. Cambridge: Cambridge University Press.

Tapanainen, Pasi. 1996. *The constraint grammar parser CG–2*. Helsinki: University of Helsinki (Department of General Linguistics, publication 27).

Tapanainen, Pasi and Timo Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, 64–71. Washington, DC.: Association for Computational Linguistics.

Tieken-Boon van Ostade, Ingrid. 1987. *The auxiliary* do *in eighteenth-century English. A sociohistorical linguistic approach*. Dordrecht: Foris.

Voutilainen, Atro. 1994a. *Three studies of grammar-based surface parsing of unrestricted English texts*. Helsinki: University of Helsinki (Department of General Linguistics, publication 24).

Voutilainen, Atro. 1994b. *Designing a parsing grammar*. Helsinki: University of Helsinki (Department of General Linguistics, publication 22).

Voutilainen, Atro and Juha 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 14th International Conference on English Language Research on Computerized Corpora, Zürich 1993*, 189–199. Amsterdam: Rodopi.