Parsing early and late modern English corpora

Gerold Schneider, Hans Martin Lehmann and Peter Schneider
English Department, University of Zurich, Switzerland

Abstract

We describe, evaluate, and improve the automatic annotation of diachronic corpora at the levels of word-class, lemma, chunks, and dependency syntax. As corpora we use the ARCHER corpus (texts from 1600 to 2000) and the ZEN corpus (texts from 1660 to 1800). Performance on Modern English is considerably lower than on Present Day English (PDE). We present several methods that improve performance. First we use the spelling normalization tool VARD to map spelling variants to their PDE equivalent, which improves tagging. We investigate the tagging changes that are due to the normalization and observe improvements, deterioration, and missing mappings. We then implement an optimized version, using VARD rules and preprocessing steps to improve normalization. We evaluate the improvement on parsing performance, comparing original text, standard VARD, and our optimized version. Over 90% of the normalization changes lead to improved parsing, and 17.3% of all 422 manually annotated sentences get a net improved parse. As a next step, we adapt the parser's grammar, add a semantic expectation model and a model for prepositional phrases (PP)-attachment interaction to the parser. These extensions improve parser performance, marginally on PDE, more considerably on earlier texts—2—5% on PP-attachment relations (e.g. from 63.6 to 68.4% and from 70 to 72.9% on 17th century texts). Finally, we briefly outline linguistic applications and give two examples: gerundials and auxiliary verbs in the ZEN corpus, showing that despite high noise levels linguistic signals clearly emerge, opening new possibilities for large-scale research of gradient phenomena in language change.

Correspondence:
Gerold Schneider,
English Department,
University of Zurich,
Plattenstrasse 47,
CH - 8032 Zurich,
Switzerland.
Email:
gschneid@es.uzh.ch

1. Introduction

Over the past decade several robust broad coverage syntactic parsers have become available. They have successfully been used for the annotation of Present Day English (PDE) corpora. More recently, large, automatically annotated corpora have been investigated in areas like syntax-lexis interactions, where enormous amounts of data are necessary (e.g. Lehmann and Schneider, 2009) and manually annotated corpora are limited by their size.

Historical corpora tend to be limited in size not only by the restrictions set by extant material but also by the effort necessary to bring the data into electronic form. However, there are fairly large unannotated diachronic corpora like the ZEN corpus with 1.6 million words, the Archer corpus with 3.2 million words, and the Old Bailey corpus with 14 million words. The entire Old Bailey proceedings contain approximately 134 million words. The main goal of the present article is to explore automatic syntactic annotation of this kind of data

Digital Scholarship in the Humanities, Vol. 30, No. 3, 2015.

doi:10.1093/llc/fqu001 Advance Access published on 6 February 2014

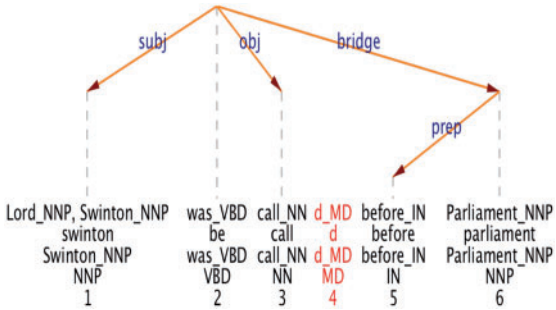

Fig. 1 Annotation problem caused by variant form *call'd*

covering a period from roughly 1650 to the present. Concerning the periodization of the English language history, we follow approaches in which the Early Modern English period (EModE) has been suggested as ranging from about 1500 to 1700 (e.g. Görlach 1991, p. 8–11, Rissanen 1999), and the Late Modern English period (LModE) from 1700 or 1800 to start of the 20th century (e.g. Tieken-Boon van Ostade 2009).

In this article we describe the automatic annotation of diachronic corpora at the levels of word-class, lemma, noun and verb chunks as well as dependency syntax. For this purpose, we adapt a framework for annotation and analysis developed for PDE (cf. Lehmann and Schneider, 2012a,b). The spelling variation found in Early and Late Modern English presents a major obstacle to automatic annotation. In section 2, we present strategies and discuss the training and adaptation of the normalization tool VARD (Baron and Rayson, 2008). Section 3 reports on the performance and adaptations made to Pro3Gres (Schneider 2008), the dependency parser we employ for the syntactic annotation. We evaluate the performance and describe the adaptations in the areas of lexical preferences and grammar rules necessary to parse the historic data as diachronic variation is potentially stronger than synchronic variation. In section 4, we explore the possibilities and limitations of the syntactically annotated diachronic corpora for historical linguistics. Specifically we discuss the problems introduced by the automatic annotation. To illustrate the new possibilities offered by the dependency annotated corpora, we present two pilot studies. We investigate diachronic change in the use of gerundials as well as the change from *be* to *have* as auxiliary in present perfect constructions.

2. Spelling Variation and Normalization

Spelling variants can cause major problems for automatic annotation. Simple variants like *call'd* for *called* typically result in wrong tagging, chunking, and parsing, as can be seen in Fig. 1. The tagger assigns the word-class general noun singular to *call* and modal to *'d*. As a consequence, the chunker fails to identify the verb group *was called*. In turn, the parser only produces two fragments and unsurprisingly fails to attach the modal *'d*.

There are two possible strategies for dealing with spelling variants. Either the annotation tool is adapted to cope with the variant directly or the spelling variants are normalized to the forms expected by the annotation tool. Our annotation framework makes use of LT-TTT2, which in turn uses the C&C tagger, the morpha lemmatizer and the LT-TTT2 chunker (Grover 2008).

Let us consider the seemingly simple problem of *hath* and *doth*. It is not enough to amend the lexicon of the tagger with forms like *hath* and *doth*. To really incorporate the variant forms, we would have to retrain the tagger with tagged text in which *hath* and *doth* actually occur. But we could not stop there because even a correctly tagged *hath* may not be recognized by the lemmatizer. And after adapting the tagger and lemmatizer we would have to change the rules of the chunker, which would otherwise not recognize *hath seen* as a verb group in the same way as *has seen*. Last but not least we would have to adapt the parser, which relies on a closed class of words that can function as auxiliaries in order to deal with auxiliaries in subject verb inversions. In our present approach we try to avoid this type of complexity by normalizing the variant forms. By simply substituting *doth* with *does,* we inherit the lexicon entry and the training data for *does* as well as the properties of *does* encoded in the lemmatizer, the chunker, and the parser, as illustrated in Fig. 2.

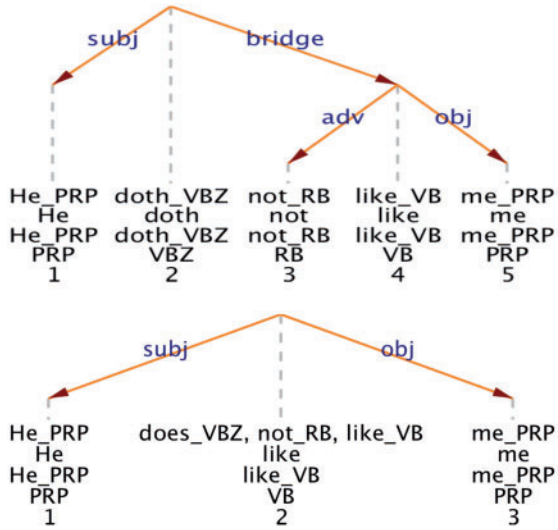

Fig. 2 Comparison of normalized and original input to the annotation chain

For automatic normalization we use VARD (Baron & Rayson, 2008). Intuitively, tagging, and consequently also chunking and parsing, improve from mapping the original spelling to the same spelling as used in the tagger and parser training resource. The statistical performance disambiguation, which uses lexical heads, should equally profit. As the normalization process also makes errors, the assumption that performance will improve cannot be taken for granted. Concerning tagging accuracy, this assumption has been tested in Rayson et al. (2007). They report an increase of about 3% (from 82 to 85% accuracy) on Shakespeare texts. As an upper bound, when texts are manually normalized, they report 89% accuracy. In the following we describe the normalization with VARD.

2.1 Using unmodified VARD for ZEN normalization

As a first step, the ZEN text was input to VARD using the default setup parameters included with version 2.4.2 of the software. The non-interactive mode of VARD compares every w-unit of the input text to a standardized PDE lexicon. If a variant does not occur in the lexicon, several algorithms are applied to find a normalized replacement, and a 'confidence score' is calculated which indicates the estimated likelihood that the replacement actually matches the original w-unit.

Using the auto-normalize function with a 50% threshold, the VARD output was analysed cursorily to get a rough idea on where it could be improved. Most of the automatic normalizations are obviously useful, such as the *-ick* and *'d* endings, and the *e->o* vowel change, while other items need a closer look (e.g. *assignees* should not be normalized to *assigns*). Table 1 shows a list of the 10 most frequent automatically suggested normalizations:

Looking at the suggested normalization in context, we found the following types of suboptimal output:

- Unnecessary Normalization
- Missing Normalizations
- Incorrect Normalizations
- Abbreviations

Since our aim was to normalize ZEN for tagging and parsing, not for lexical correctness by PDE standards, we tried to concentrate on those areas where we expected the normalization to help the part-of-speech (POS) tagger. Ideally, an optimized normalization process should observe the following maxims:

- All normalized items should retain the word class if it was correctly identifiable in the original form.
- When the tagger would not correctly identify the word class of an original item, it should be normalized to a form with the correct word class.
- Little or no effort should be made to improve the normalization of items whose original and normalized form share the same word class.

2.2 Problems and solutions for VARD processing

2.2.1 *Unnecessary normalization*

Most non-standard variants and problematic normalizations concern place names and proper names. While it may be historically interesting, the normalization of names is not really necessary in the context of part-of-speech identification since software

Table 1 Most frequent VARD normalizations of ZEN (1,586,653 tokens, of which 27,167 were automatically normalized)

Count	Original	Normalized
764	tis	it is
515	publick	public
418	publish'd	published
340	tho'	though
283	assignees	assigns
232	call'd	called
181	lett	let
180	chuse	choose
175	arriv'd	arrived
155	shew	show

for automatic tagging can identify them in the original spelling, as in example (1):

(1) This day Sir William <normalised orig="*Swann*" auto="true">*Swan*</normalised>, . . . 1671cui00013: This_DT day_NN Sir_NNP William_NNP Swann_NNP . . .

Likewise, variants of place names pose no problem, such as in (2):

(2) Letters_NNS from_IN Vienna_NNP and_CC Francfort_NNP tell_VBP us_PRP . . .

Since titles and honorifics are usually followed by one or more proper names as in *Sir John Fitz-Gerald*, VARD was instructed not to process a sequence of title variants (e.g. *Sir*, *Lord*, *Marquis*), a preposition (e.g. *de*, *of*), and one or two capitalized words. This was achieved with a set of regular expressions in the 'text_to_ignore.txt' file. Some more expressions were added to skip likely place names preceded by a set of indicators, such as *Province of* . . . , *Parish of* . . . to avoid more unnecessary normalizations.

2.2.2. *Missing normalizations*

The old verb forms *hath* and *doth* confuse the tagger. Of the 778 occurrences of *hath*, only 205 are identified as verbs, and in the case of the fifty-three instances of *doth*, only eight are seen as verbs. Since the standard lexicon contains both forms, they are not automatically normalized. This was remedied in the interactive mode of VARD by explicitly adding the normalized variants *has* and *does* to the list of mandatory replacements (variants.txt).

2.2.3 *Incorrect normalizations*

Since VARD's lexicon is derived from a word list based on most frequent items in modern corpora, many less-frequent words are missing. This means that VARD will attempt to normalize items even though they would be correctly spelled by PDE standards. Table 2 presents a list of items and their (incorrect) normalization as proposed by standard VARD.

Since ZEN has a different lexical frequency distribution compared with modern corpora, it was necessary to manually go through the most frequent variants in the VARD interactive mode, and decide if an item needs to be added to the word list ('All not variant') or to the list of mandatory replacements ('Normalize to . . .').

2.2.4 *Abbreviations*

Non-standard abbreviations occur frequently in ZEN. While abbreviated titles such as Bart (Baronet) or Esq. (Esquire) are usually non-problematic, the tagger sometimes stumbles over abbreviated first names, such as *Wm* (William) or *Edw* (Edward):

(3) 1701lgz03673: Whoever_WP secures_VBZ the_DT Mare_NNP,_, and_CC gives_VBZ Notice_NN to_TO *Edw_VB* Quane_NN . . . shall_MD have_VB 20_CD s_PRP._.Reward_NNP . . .

(4) 1701lgz03674: Whoever_WP secures_VBZ the_DT Horse_NNP . . . and_CC gives_VBZ notice_NN to_TO Wm_VB Brooke_NNP . . . shall_MD have_VB 2_CD Guineas_NNP Reward_NNP._.

Some common abbreviations were therefore added to the VARD list of items with mandatory replacements (variants.txt). In addition to first names, we included frequent items such *ult* ('last month', fifty-nine instances) and *'em* ('them', 145).

2.2.5 *Non-standard capitalization*

The tagger is sensitive to capitalization issues since capitalization is used to identify proper nouns.

Table 2 Incorrect normalizations due to lexicon limitations

ZEN original	VARD auto-normalization
Assignee	*assigns*
Patence (patentee)	*Patience*
Relict (widow)	*Relic*
Footpad (robber on foot)	*Footpath*
Porte (Ottoman Empire)	*Port*
Dom (Spanish title, or abbreviated *an[no] dom[ini]*)	*Doom*
Messuage (dwelling)	*Message*
Paul (first name)	*Pal*

Taggers do typically not identify a capitalized adjective, as in (5):

(5) 1751gat05396: . . . Prisoners_NNS in_IN the_DT Tobooth_NNP here_RB,_, were_VBD served_VBN with_IN *Criminal_NNP* Letters_NNP,_, at_IN the_DT Instance_NN of_IN his_PRP$Majesty_NNP 's_POS Advocats_NNS . . .

We do not address the problem of non-standard capitalization of nouns in ZEN in this article.

2.3 Evaluation of optimizations

2.3.1 *Summary view*

In order to evaluate the relative improvements between the original ZEN text (z0), the default VARD auto-normalized version (z1), and the optimized version (z3), the three text versions were processed by the C&C tagger. In a first attempt, individual POS tags were counted and arranged in four main groups of tags (Fig. 3). However, this evaluation only revealed a somewhat lower proportion (5%) of nouns and a very slightly higher proportion of verbs (2%) when both normalized texts z1 and z3 were compared with the original z0.

2.3.2 *A Changes-based Look at the Normalizations*

Another type of analysis was therefore necessary to reveal more relevant differences. Rather than going on counting unrelated entities, we decided to classify how normalization affected POS sequences and word+POS-tag combinations. To this end, the

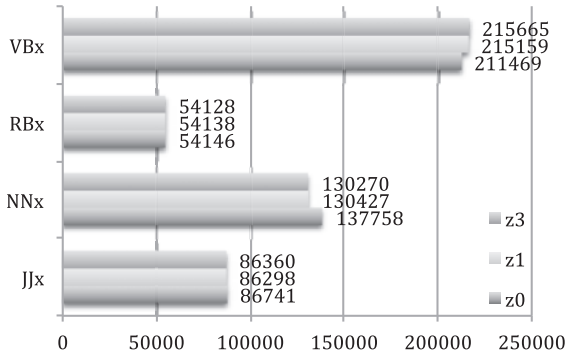

Fig. 3 Distribution of grouped POS tags (JJx: adjectives, NNx: nouns, RBx: adverbs, VBx: verbs, X axis indicates number of tags)

GNU *wdiff* tool was applied to each set z0z1, z0z3, creating a list of word+tag edits. The output annotes deleted sequences with [- and -] indicators, and corresponding replacements by {+ and +}. (6) shows the influence of normalization on POS tagging between z0 and z3:

(6) 1711evp00286: We_PRP are_VBP *[-advis_NNS 'd_VBD-] {+advised_VBN+}* that_IN Admiral_NNP Norris_NNP 's_POS Fleet_NNP met_VBD with_IN a_DT great_JJ Storm_NN in_IN the_DT *[-Gulph_NNP-] {+Gulf_NNP+}* of_IN Lions_NNPS,_, but_CC *[-suffer_VBP 'd_MD-] {+suffered_VBD+}* no_DT other_JJ Damage_NN than_IN some_DT of_IN the_DT Transports_NNS with_IN Troops_NNS on_IN Board_NNP being_VBG *[-oblig_VBN 'd_MD-] {+obliged_VBN+}* to_TO shelter_NN themselves_PRP in_IN some_DT of_IN the_DT Harbours_NNS of_IN the_DT Mediterranean_NNP._.

While the normalization of *Gulph* to *Gulf* did not prompt the tagger to analyse the item differently, the normalization of the *'d* verb forms leads to a better analysis.

For a further comparative look at the changes, regular expressions were applied to the *wdiff* output to only consider sequences where the assigned POS tags underwent a change. Table 3 lists the 10 most frequent such changes in z1:

It turns out that roughly half of the normalizations affect the tagging, as shown in Table 4. The z3

version has 15% fewer overall normalizations compared with z1, but still has a 3% higher number of tag-affecting normalizations. This is a likely result of the title sequence *ignore instructions* indicated in section 2.2.1.

Another analysis of the *wdiff* output discards other content, leaving just the POS tag intact. The results of the comparison between z1 and z3 presented in Table 5 shows that the most frequent tag sequence changes are similar, apart from the NN to VBZ transition in z3. This is the effect of the addition of *hath* to the dictionary as proposed in section 2.2.2.

Table 3 Tag-affecting changes due to normalisation ($n \geq 10$)

Count	z0	z1
287	[-publish_VB 'd_NNP-]	{+published_VBN+}
168	[-publick_NN-]	{+public_JJ+}
163	[-Tis_NNP-]	{+It_PRP is_VBZ+}
154	[-Publick_NN-]	{+Public_NNP+}
139	[-tis_VBZ-]	{+it_PRP is_VBZ+}
125	[-tho_NNS '_POS-]	{+though_IN+}
125	[-tho_NNP '_POS-]	{+though_IN+}
104	[-Republick_NN-]	{+Republic_NNP+}
93	[-tis_NNS-]	{+it_PRP is_VBZ+}
83	[-'s_POS-]	{+s_VBZ+}

Table 4 Overall and Tag-affecting normalizations in z1 and z3

Normalizations	z1	z3	Difference
Overall (o)	27,416	23,216	4,200 (−15%)
Tag-affecting (t)	13,267	13,609	342 (+3%)
Ratio o/t	2.1	1.7	

2.3.3 *Standard versus Optimized Normalization*

The same set of tools that was used to assess the differences between the original (z0) and the normalized versions (z1, z3) were also employed to evaluate the potential improvements in tagging. Similar to Table 3, tag-affecting changes between the non-optimized and the optimized normalization are summarized in Table 6. To increase legibility, we did not include changes due to differences in the form of compounds, such as the presence or absence of a hyphenation or a word space in place names with *street, lane, row* (e.g. *Fleetstreet/Fleet street,* or *Drury-Lane/Drury Lane*).

The importance of the correct identification of the verb *hath* as PDE *has* is again illustrated nicely: if *has* carries the (correct) VBZ tag, the following verb form will also be correctly identified as a past participle (VBN) instead of past tense (VBD). While most of the z1->z3 changes are welcome improvements, Table 6 shows that there are exceptions: items which were correctly normalized in z1 appear to have regressed in z3, such as the missing *'d/ed* verb ending normalization. Since the other 284 instances of *allow'd* and ninety-three instances *follow'd* are handled and normalized by VARD as expected, this is likely due to a different f-score assigned in the optimized version.

Table 5 Affected tag sequences in z1 and z3 ($n \geq 10$)

Count	z0	z1	Count	z0	Z3
879	[−NN MD−]	{+VBN+}	867	[−NN MD−]	{+VBN+}
657	[−NN−]	{+NNP+}	637	[−JJ NNP−]	{+VBN+}
647	[−JJ NNP−]	{+VBN+}	592	[−NN−]	{+NNP+}
579	[−VB NNP−]	{+VBN+}	560	[−VB NNP−]	{+VBN+}
510	[−NN−]	{+JJ+}	469	[−NN−]	{+JJ+}
425	[−NNP−]	{+NN+}	400	[−NNP NNP−]	{+VBN+}
418	[−NNP NNP−]	{+VBN+}	372	[−VB MD−]	{+VBN+}
392	[−VB MD−]	{+VBN+}	349	**[−NN−]**	**{+VBZ+}**
305	[−VBP MD−]	{+VBD+}	330	[−NNP−]	{+NN+}
193	[−NN MD−]	{+VBD+}	301	[−VBP MD−]	{+VBD+}

Table 6 Tag-affecting changes between z1 and z3 (some omitted items)

Count	z1	z3
314	[−hath_NN−]	{+has_VBZ+}
100	[−'_POS em_NN−]	{+them_PRP+}
83	[−s_VBZ−]	{+'s_POS+}
54	[−hath_NN surrendered_VBD−]	{+has_VBZ surrendered_VBN+}
35	[−hath_VBP−]	{+has_VBZ+}
26	[−'_" em_NN−]	{+them_PRP+}
19	[−allowed_VBN−]	{+allow_VB 'd_NNP+}
18	[−doth_NN−]	{+does_VBZ+}
14	[−Port_NNP−]	{+Porte_NN+}
12	[−Poultry_NN−]	{+Poultrey_NNP+}
12	[−20_CD th_NN−]	{+20th_JJ+}
11	[−tis_JJ−]	{+it_PRP is_VBZ+}
11	[−Switzers_NNS−]	{+Swiss_NNP+}
11	[−24_CD th_NN−]	{+24th_JJ+}
10	[−Tis_NNP−]	{+It_PRP is_VBZ+}
10	[−Infant_NN−]	{+Infanta_NNP+}
10	[−hath_NN sent_VBD−]	{+has_VBZ sent_VBN+}
10	[−hath_NN made_VBD−]	{+has_VBZ made_VBN+}
10	[−followed_VBN−]	{+follow_VB 'd_MD+}

3. Syntactic Parsing of Modern English Texts

Robust broad-coverage syntactic parsers, for example, Collins (1999), Nivre (2006), Schneider (2008) have now become available. Van Noord and Bouma (2009, p. 37) state that '[k]nowledge-based parsers are now accurate, fast and robust enough to be used to obtain syntactic annotations for very large corpora fully automatically'. Large corpora such as the British National Corpus (Aston & Burnard, 1998) have been made accessible in automatically parsed versions, for example, Andersen (2008) or Lehmann and Schneider (2012b), offering new perspectives for linguistic research.

A major reason for the relative accuracy and efficiency of these syntactic parsers is that they use fast finite-state technology like taggers, chunkers, and morphological analysers in the pre-processing step and that they largely rely on statistical data which minimally encodes lexical preferences. Kaplan et al. (2004) describe finite-state preprocessing as a necessary prerequisite for efficient and accurate parsing.

Concerning lexical preferences, it is important to point out that applying all grammatical rules to a sentence to be parsed massively overgenerates, i.e. often leads to hundreds of possible parses, most of which are semantically implausible. Lexical preferences are used to disambiguate and find the most likely syntactic analysis. Lexical preferences are encoded in the form of bi-lexical conditioning (e.g. Collins 1999), which means that syntactic rules in which both the governor and the dependent lexeme are likely to occur are preferred. This strategy is analogous to the dichotomy of syntax principle versus idiom principle (Sinclair 1991, Hunston and Francis, 2000) in which the application of syntactic *competence* rules is constrained and ranked by idiomatic *performance* patterns. In addition to affecting tagging performance (section 2 and 3.1), lexical statistics often fails to deliver any data (or it delivers incorrect data) if historical spelling instead of normalized spelling is used, which means that the disambiguation between various syntactically possible analyses is affected. We address this point in section 3.2.

3.1 Improvement due to normalization

The assumption that normalization improves parsing performance has first been confirmed in Schneider (2012): in a 100 sentences random

sample from the ARCHER corpus 17th century section, 131 normalizations are made (in VARD batch mode, 50% confidence level). In the normalized text, 16 of the 100 sentences receive a syntactic analysis which differs from the original. A manual inspection reveals better syntactic analysis due to VARD in twelve sentences, worse syntactic analysis due to VARD in one sentence, and improvements paralleled by new errors in three sentences.

Here we use a larger random sample from the ZEN corpus, comprising 422 sentences. We use first (section 3.3.1) a version with the standard normalization settings of VARD, then (section 3.3.2) our retrained VARD version.

3.1.1 *Standard VARD*

Of the 422 sentences, 332 obtain a different syntactic analysis when using the standard VARD settings. The results are broken down by syntactic relation in Table 7. We get an improvement of sixty-eight relations opposed to five new errors. More than 90% of the changes are improvements, and 15% of the original 422 sentences, and 19% of the sentences whose tagging was affected get a net improved parse.

An example is given in Fig. 4, where the original spelling in sentence (7) *scorbutick* is tagged as a verb (top), while the normalized *scorbutic* is tagged correctly as adjective, which leads to the correct syntactic analysis (bottom)

(7) The only short and infallible Cure for that reigning Disease the SCURVY and all scorbutick Humours, . . . (ZEN 1741CJL)

3.1.2 *Retrained VARD*

Of the 422 sentences, 132 obtain a different syntactic analysis after retraining VARD, compared to using the standard VARD. The results are broken down by syntactic relation in Table 8. We get a further improvement of eleven relations opposed to 1 new error. Of all 422 sentences, 17.3% get a net improved parse.

An example can be found in Fig. 2 in section 2. The original spelling *doth* is not normalized by VARD standard. After our retraining it is correctly normalized to *does*, which leads to the correct syntactic analysis.

Table 7 Parser improvement versus new errors with standard VARD

	Better	Worse	Equal
subj	21	2	25
obj	17		25
pobj	10	2	31
modpp	9	1	48
sentobj	11		10
$\sum$	68	5	139

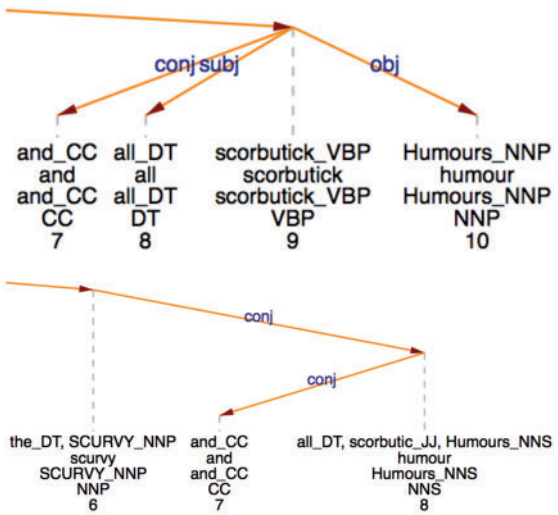

Fig. 4 Syntactic analysis with original spelling and normalized spelling

3.2 Parser adaptation

We have stated that a major reason for the relative accuracy and efficiency of syntactic parsers is that they rely on statistical data which encodes lexical preferences between governors and dependents (Collins, 1999). Lexical preference statistics are learnt from a manually annotated resource (the learning process is called *training*), typically the Penn Treebank is used (Marcus et al. 1993). While a number of parsers now reach acceptable accuracy when applied to domains that are similar to the training domain, performance drops considerably when texts from different domains are parsed (Gildea 2001). Domain adaptation is therefore a

Table 8 Parser improvement of standard VARD versus retrained VARD

	Better	Worse	Equal
subj	4		6
obj	2		8
pobj	2		11
modpp		1	18
sentobj	3		2
$\sum$	11	1	45

current research focus in broad-coverage parsing (Buchholz and Marsi, 2006; Nivre et al., 2007).

Lehmann and Schneider (2012a) have evaluated random sets from the BNC and report similar to slightly lower performance than on in-domain texts. Performance decreases increasingly with domains that differ more from the training domain, partly due to incorrect part-of-speech tagging in the preprocessing step, and partly due to inappropriate lexical preferences. There is a danger that the level of noise introduced by tagging and parsing errors will at some stage be stronger than the signal. The signal reports true quantitative differences. Schneider and Hundt (2009) evaluate parser performance on L2 varieties of English such as Indian or Fiji English. They show that for the application to regional variation the signal delivered by an automatic parser (Schneider 2008) is typically strong enough.

Even if the performance decrease for variation according to region and genre seems manageable, diachronic variation has the potential to be much stronger than synchronic variation, and not only affect lexical preferences but also the set of permissible grammar rules.

Rissanen (1999) states that from about 1700 on, the structure of PDE had largely been established.

> 'At that time [1700], the structure of the language was gradually established so that eighteenth-century standard written English closely resembles the present-day language. The language of most sixteenth-century authors still reflects the heritage of Middle English, whilst it is possible to read long passages from eighteenth century novels or essays and find only minor deviations from present-day constructions.' (Rissanen 1999, p. 187).

Denison (1998) also confirms:

> 'By 1776 the English language had already undergone most of the syntactic changes which differentiate Present-Day English (henceforth PDE) from Old English (henceforth OE)'
> (Denison 1998, p. 92).

These quotes support our initial hypothesis that except for spelling variation (which we have addressed in section 2), shifts in lexical preferences (which degrades parsing performance), and changing frequencies of certain syntactic constructions (which we hope to measure as signal with our approach) the fundamental set of grammar rules may only need large adaptations for earlier periods, in other words for the earliest texts in ARCHER and ZEN. We expect a weak decline in parser performance from the 20th century to the 18th century, and then a stronger decline for the 17th century texts.

Particularly for the LModE, it has been claimed that the differences to PDE are mainly of statistical nature. Construction types remain the same. The frequency of the types, however, may change. These changes in frequency can themselves be preparatory steps for language change.

(8) illustrates the difficulties automatic parsers face in Early Modern English. It also highlights some of the features of Early Modern English.

(8) The ship, the Amerantha, had never yett bin att sea, and therfore the more daungerous to adventure in her first voyage; butt she was well built, a fayre ship, *of a good burden*, and had *mounted in her forty pieces* of brasse cannon, two of them demy cannon, and she *was well manned*, and *of good force* and strength for warre: she was a good sayler, and would turne and tacke about well; she held 100 persons of Whitelocke's followers, and most of his baggage, *besides her own marriners, about 200*. (ARCHER 1654whit.j2b, italics added)

Processing our spelling normalised version (see section 2) of sentence (8), the parser makes a number of errors that are related to markedness. The following constructions are also possible in PDE, but highly marked.

Genitives of quality (e.g. *of a good burden* and *of good force*) are frequent in Latin or in biblical contexts but rarely used in PDE (e.g. Köstenberger and Patterson, 2011, p. 587).

X-bar violations are rare and poetic in PDE: *mounted [in her] forty pieces* is an X-bar scheme violation. In one possible syntactic interpretation of this sentence the subcategorized object *forty pieces* is further remote from the verb than the adjunct *in her* (the X-bar compatible order would be *had mounted forty pieces in her*). Notice that the non-argument is not moved outside the VP as in topicalization but may rather be a scrambling phenomenon similar to Present Day German (e.g. Grewendorf and Sternefeld, 1990).

There is also a second possible syntactic interpretation of *mounted [in her] forty pieces* in which *had* is the main verb, and *mounted in her* is a modifying participial clause. The X-bar violation then consists in having a non-subcategorized participial clause closer to the verb than the subcategorized object (the X-bar compatible order would be *had forty pieces mounted in her*). The effect of the X-bar violation here is that the chunker returns [*her forty pieces*] as a single base noun phrase.

Conjunctions are typically constrained to combine constituents that have the same word class. In *was well manned, and of good force and strength for war* an adjective and a complex prepositional phrases (PP) are in coordination. It seems that this constraint was much weaker in EModE.

It also appears that constraints on appositions were weaker: In *besides her own mariners, about 200* an apposition relation is used to convey quantity information, a use we might perhaps only find in cooking recipes in PDE.

In Modern English, particularly in EModE, sentences are considerably longer than in PDE. Fries (2010, p. 31) reports a decrease in sentence length in the ZEN corpus from forty-two words per sentence in 1661 down to twenty-nine words per sentence in 1791, while PDE figures (from the BNC) are about twenty-one words per sentence. High sentence length in itself creates considerably more scope for ambiguity. We exemplify the ambiguity for PP-attachment. The ambiguity of prepositional phrase attachment can be described by the Catalan numbers.

A sequence *verb NP n*PP* with n PPs has C_{n+1} analyses, where C_{n+1} is the $(n+1)$'th Catalan number. C_n is defined as follows:

$$C_n = \frac{1}{n+1}\binom{2n}{n} = \frac{(2n)!}{(n+1)!n!}$$

where C_n 1 . . . 12 is [1, 2, 5, 14, 42, 132, 429, 1,430, 4,862, 16,796, 58,786, 208,012]

For five PPs there are forty-two possible readings. As a crude indicator of the potential ambiguity we can compare sentence length across the centuries. Average sentence length in the ZEN corpus is 33.74 words, compared with about 21 words in the BNC. As ARCHER is not sentence-tokenized, only approximate figures can be obtained. Our own tokenization, which is very conservative, reports about 60 words per 'sentence' in the 17th century compared with about 40 words per 'sentence' in the 19th century.

In sum, we conclude that the types of parsing errors produced for both Early and Late Modern English are similar to PDE, but more frequent. This is due to increased ambiguity caused by longer sentences and marked word order. We expect more disambiguation errors, and we need more statistical data and semantic resources to improve results. In section 3.2.1, we evaluate the parser without any adaptations to ModE (but using automatically normalized spelling). In sections 3.2.2 to 3.2.4, we then address improvements and adaptations for ModE.

3.2.1 *Evaluation*

We have manually annotated 100 random sentences from each of the 17th, 18th, 19th and the 20th century from Archer corpus texts. They include the twenty-five random sentences per century which have been used in the evaluation in Schneider (2012); the current evaluation set is thus four times larger. We have used the standard VARD normalization. The evaluation results including raw frequencies are given in Table 9, in terms of precision and recall, broken down by century and syntactic relation. The F-score results, by century, are given in a bar chart in Fig. 5. The F-score is the harmonic mean of precision and recall.

Table 9 Performance of the baseline parser in absolute frequencies and in percent, on selected relations, broken down by century and syntactic relation

16xx	**is**	**should**	**%**	**17xx**	**is**	**should**	**%**
PREC.				PREC.			
subj	162	206	78.64	subj	189	226	83.63
obj	111	152	73.03	obj	106	135	78.52
pobj	77	112	68.75	pobj	90	125	72.00
modpp	82	125	65.60	modpp	91	121	75.21
sentobj	37	67	55.22	sentobj	55	80	68.75
$\sum$	469	662	**70.85**	$\sum$	531	687	**77.29**
RECALL				RECALL			
subj	161	195	82.56	subj	190	229	82.97
obj	111	138	80.43	obj	107	133	80.45
pobj	79	135	58.52	pobj	92	141	65.25
modpp	81	109	74.31	modpp	91	120	75.83
sentobj	37	81	45.68	sentobj	56	97	57.73
$\sum$	469	658	**71.28**	$\sum$	536	720	**74.44**
18xx	**is**	**should**	**%**	**19xx**	**is**	**should**	**%**
PREC.				PREC.			
subj	125	149	83.89	subj	172	197	87.31
obj	71	89	79.78	obj	93	116	80.17
pobj	77	99	77.78	pobj	93	123	75.61
modpp	52	76	68.42	modpp	72	94	76.60
sentobj	27	43	62.79	sentobj	42	70	60.00
$\sum$	352	456	**77.19**	$\sum$	472	600	**78.67**
RECALL				RECALL			
subj	124	151	82.12	Subj	173	203	85.22
obj	71	92	77.17	Obj	92	110	83.64
pobj	80	110	72.73	Pobj	93	128	72.66
modpp	52	71	73.24	Modpp	72	97	74.23
sentobj	27	65	41.54	Sentobj	42	64	65.63
$\sum$	354	489	**72.39**	$\sum$	472	602	**78.41**

subj = Subject; obj = Object; pobj = verb-attached PP; modpp = noun-attached PP; sentobj = subordinate clause.

As expected, parser performance decreases for the 18th and 19th centuries, and shows a steeper decline for the texts before 1700. There is also some fluctuation. When we inspected the errors, we noticed that the 18xx random evaluation set texts are affected by many instances of *hath*, which is not normalized by VARD standard settings, and which leads to tagging and lemmatizing errors. Our inspection of errors also revealed that some errors are relatively easy to correct, as they involve closed class words. We will briefly describe them in 3.2.2, before turning to errors that can partly be corrected by improving semantic and statistical resources in section 3.2.3.

3.2.2 *Closed class lexis extensions*

Some closed-class words, e.g. *but* as adverb in (9), are not known to the parser grammar. We have made a number of such adaptations: the conjunction *lest*, *as* in the function of a relative pronoun (which we discarded again as it led to new errors), or *gain* as a ditransitive verb.

(9) He is such an Itinerant, to speak that I have *but* little of his company. (Archer:1766aadm)

Such adaptations are straightforward and efficient. However, they only lead to small, specific improvements.

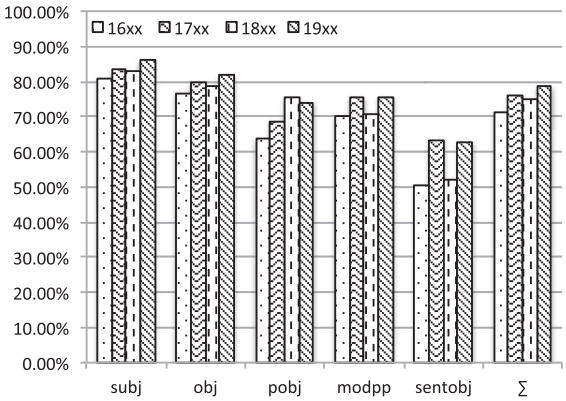

Fig. 5 F-Score performance of the baseline system, by century and syntactic relation

3.2.3 *More semantics and context*

Additional parsing errors in the texts before 1900 come from a number of sources, including the following:

- Rare 'poetic' constructions that are not licensed by the grammar. Examples include *mounted in her forty pieces* in example sentence (1), and the sentence (ARCHER 1775prie.s4b) *On what this difference depends I can not tell* where a complex PP is fronted. The grammar only licenses the fronting of simple PPs (such as *in the morning*), and relaxing this constraint generally leads to lower parsing performance.
- Lexical preferences that do not match. Examples include the genitive of quality *ship of a good burden* in example (1), and PP-attachment involving *at large* in the sentence (ARCHER 1674leew.s2b) *as I shall manifest at large in the ensuing discourse*, where *discourse* is attached to the adjective *large* instead of the verb *manifest*.
- High complexity, marked constituent order. In this category, we find parser errors that also occur in PDE, but they are more frequent in ModE; ModE is similar to PDE but harder.

Particularly the last sources of errors illustrate what could be called the ambiguity trade-off between constraining and disambiguating: if one constrains rules too much, the correct reading can often not be found, for example, if a marked constituent order is used. If one constrains too little: ambiguity explodes, the risk for incorrect disambiguation increases. Disambiguation can sometimes be improved by adding more resources. One way to help disambiguation is to include more semantics and context, as we have done in the following.

a) Semantic expectation. The original parser models probabilities using only those syntactic relations that are in competition. For example, objects (e.g. *eat pizza*) and nominal adjuncts (e.g. *eat Friday*) are modeled as being in competition, but not subjects and objects.

$$p(R,dist|a,b) = p(R|a,b) \cdot p(dist|R,a,b)$$
$$\approx \frac{f(R,a,b)}{f\left(\left(\sum R\right),a,b\right)} \cdot \frac{f(R,dist)}{f(R)}$$

We now add semantic competition as a further factor: every relation is in competition with every other relation. A sentence like the *rabbit chased the dog* now gets a lower probability than *the dog chased the rabbit* because rabbits are very unlikely to be subjects of active instances of *chase*. Our semantic world knowledge (e.g. selectional restrictions) becomes part of the model.

b) Wider context. Attachment of PP is typically the most ambiguous syntactic relation. The interaction between multiple PPs was not considered in the original statistical model of the parser. Knowledge expressed across more than one node generation was lost. We have added a model for the probability that PP2 is a dependent of PP1 (PP1 < PP2) in a verb-PP-PP sequence, given the lexical items. It is calculated as follows:

$$p(verb < (PP_1 < PP_2)) = \frac{\#(verb < (PP_1 < PP_2))}{\#(verb < (PP_1 < PP_2)) + \#((verb < PP_1) < PP_2)}$$

These two measures improve recall and precision, as can be seen in Fig. 6. The performance of the baseline parser from section 3.2.1 is shown by grey bars, and the performance of the extended parser by striped bars. Interestingly, earlier centuries profit more from the adaptation, which we believe may indicate that, due to freer word order and longer

sentences, constraints on semantics and complexity are more important.

The overall performance with the new semantic expectations and the improved PP-model are given in Table 10. As expected, we see a weak decline from the 20th century down in history to the 18th century, and then a stronger decline for the 17th century texts.

As expected, the addition of more statistical data for the highly ambiguous PP-attachment, and semantic resources modelling our expectations, improves parsing. Particularly on the historical texts, where ambiguity was found to be higher than in PDE, PP-attachment improves by 2–5%.

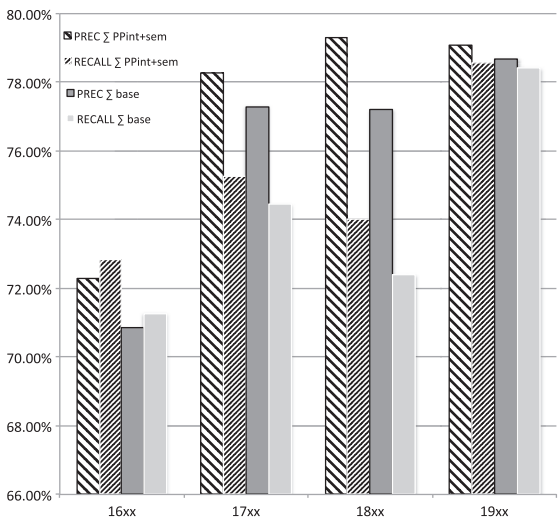

Fig. 6 Precision and recall of improved parser (striped) and baseline parser (grey)

4. Linguistic Applications

The structure of English had been established by the beginning of the 18th century (Denison 1998, Rissanen 1999); see section 3.2. López-Couso, Aarts and Méndez-Naya (2012) state that in addition to few grammatical innovations, namely, the progressive passive and the *get*-passive, the Late ModE period is marked by regulatory and statistical changes: the progressive form increases in frequency, *be* as perfect auxiliary decreases, periphrastic *do* is fully established, and non-finite complementation and relativization (Hundt, Denison, Schneider 2012a) have undergone changes.

The present progressive form, which has a relative frequency of about 40 instances per 10,000 words in ICE spoken, and about twenty in ICE written, has less than ten instances per 10,000 words in the ZEN corpus period (1661–1791). The increase of the progressive has been described in detail in Hundt (2004). When looking at *–ing* forms, we have noticed that the majority of them from the early ZEN and ARCHER texts are in fact non-finite *-ing* forms, also known as gerundials (Mair 2003), which we discuss in section 4.1. In section 4.2, we show that *be* as perfect auxiliary shows a clear decline even in the short ZEN period of 130 years.

In the following we present two pilot studies based on the new annotation. The results and distributions presented have been derived from our web-based interface developed for the dependency bank project. See Lehmann and Schneider (2012a, 2012b) for a detailed description.

Table 10 Base (base) parser compared with improved (imp) parser, F-score

F-Score	16xx		17xx		18xx		19xx	
	base	imp	base	imp	Base	imp	base	imp
subj	80.60	80.49	83.30	83.48	83.01	83.01	86.27	86.27
obj	76.73	76.49	79.48	79.78	78.47	78.47	81.90	81.90
pobj	63.63	68.38	68.62	70.73	75.25	79.51	74.13	73.78
modpp	69.96	72.85	75.52	77.89	70.83	76.28	75.41	77.32
sentobj	50.45	52.96	63.24	63.24	52.16	52.91	62.81	63.25
$\sum$	71.06	72.56	75.87	76.78	74.79	76.66	78.54	78.83

4.1 Gerundials

The progressive form, which is already found in Old English, has become an established construction in Early Modern English (Denison 1998, p. 130), and increased in frequency since, as we have just discussed. While *-ing* forms used as progressives are rare in the early ZEN and ARCHER texts, nominal *-ing* participle clauses, also known as gerundials, are quite frequent. In particular, there are surprisingly many occurrences of gerundials with subjects. In PDE English (Quirk et al. 1985, p. 1063), the functions of gerundials comprise subject, object, subject complement, appositive, adjectival, and prepositional complement. An example of prepositional complement with subject is as follows:

(10) All the Passages which were shut up on account of the **Plague** *being* at Leipzig and several places in Saxony are now again open and Trade is restored to its former Course. (ZEN 1681LGZ)

The most frequent syntactic functions by far are appositive clauses.

(11) **Some Scottish Brethren**, in the North of Ireland *finding* their wonted Practices interrupted by the late Declaration of the Lords Justices and Council against Presbyterians Anabtists. (ZEN 1661KIN)

(12) The Picaroons have not visited our Coasts these six months and indeed our Vessels so well fitted **several of them** *carrying* six eight ten and twelve Guns apiece that the small Capers which usually haunted these Coasts have no encouragement to adventure. (ZEN 1671CUI)

(13) **This Congregation** *ending* a Courier was immediately dispacht to Segnior Ravizza... (1671LGZ)

Many occurrences are also found in main clauses:

(14) Robert Pierrepoint Esq; **his Troop** *consisting* of 120 Horse whose Lieutenant is Toplady Esq; and Gregory Esq; Cornet. (ZEN 1661KIN)

(15) **Mr. Wiseman** a Mercer *accompanying* Sir George Geffryes. (ZEN 1681CUI)

Where the gerundial occurs in a main clause, such as (14) and (15), it cannot be distinguished from a present participle clause (Quirk *et al.*, 1985, p. 1,263). Present participle clauses are also known as present tense reduced relative clauses, but their state is contested (e.g. Hundt, Denison, Schneider 2012b). We have used the following syntactic search patterns: (1) subject relation, where the verb is in the progressive form and non-finite and (2) reduced relative clause where the verb is in the present. Query (1) delivers 3914 hits, (2) 1207 hits. The hits contain many appositive clauses (10–13), present tense reduced relative clauses (16–18), but also parsing mistakes and other syntactic functions.

(16) **The States of Holland** *being* now complete are resolved to dispose forthwith of the vacant Companies. (1671CUI)

(17) **Seven of the Dutch Frigates** *standing* into Margate Road cause the Lilly and another Frigate to stand for the River. (1671 CUI)

(18) Yesterday was a Council at White-Hall chiefly to hear several Appeals from out of the Island of Guernsey according to the Constitution of that place but **one of the persons** *being* dead since the Appeal was brought it could not be heard. (ZEN 1681IMP)

(19) **The Publication** of Books of Medicines and other such things *being* remote from the business of a Paper of Intelligence; This is to notify that we will not charge the Intelligence with Advertisements unless they be matter of State but that a Paper of Advertisements will be forthwith Printed apart and recommended to the Public by another hand. (ZEN 1671CUI)

Semantically, the participle often conveys an argumentative semantic function, most obviously in (12,13,18,19). In PDE this only survives in set expressions such as *this being so*. The frequency distribution delivered by the interface for query (1) is given in Table 11 and shows a clear decline, graphically rendered in Fig. 7.

The frequency of gerundials with subjects has been decreasing, and this change seems to take place early in the investigated period, between 1,661 and 1,691. The difference is very highly

Table 11 Absolute and relative frequency of gerundials with subjects in ZEN

Decade	*n*	Words	f per 10,000 wd
1661	23	4412	52.1
1671	174	43,973	39.6
1681	197	55,496	35.5
1691	209	85,185	24.5
1701	430	172,014	25
1711	255	110,055	23.2
1721	281	115,193	24.4
1731	325	132,471	24.5
1741	286	123,335	23.2
1751	495	184,524	26.8
1761	305	143,362	21.3
1771	364	187,501	19.4
1781	195	96,666	20.2
1791	375	197,339	19

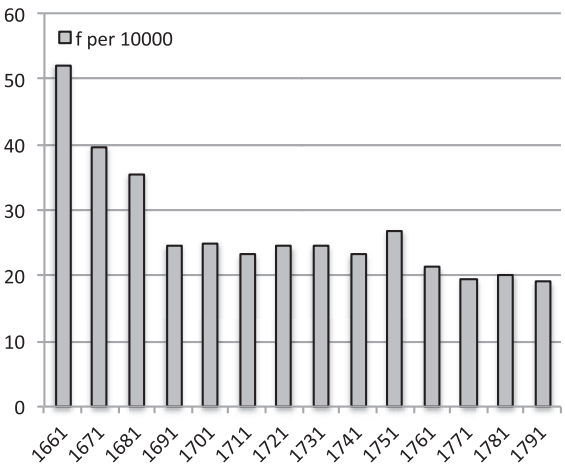

Fig. 7 Absolute and relative frequency of gerundials with subjects in ZEN

significant ($P < 2E\text{-}24$), according to Chi-Square contingency test. Even if the data from 1,661 are discarded as they may be seen as too sparse, the difference stays highly significant ($P < 5E\text{-}22$). Our findings pattern well with the larger picture drawn by López-Couso, Aarts, and Méndez-Naya (2012), who observe that:

> While we can speak of relative stability in the area of finite complementation, the realm of non-finite complementation experienced 'fundamental and rapid changes' in our period (Mair 2003, p. 329), some of them still under way.

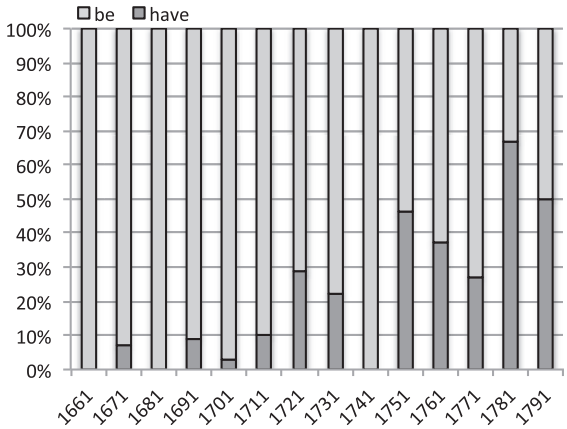

Fig. 8 Perfect auxiliary *be* and *have* with *come* in the ZEN corpus. $n = 152$

4.2 *be* or *have* as auxiliary in the perfect

Concerning the fixation of *have* as auxiliary, we have investigated auxiliary verbs in the perfect form. In some verbs, even the short ZEN period reveals clear change. While the verb *go* keeps a preference for the auxiliary *be* throughout ZEN (and *be gone* is still occasionally used in PDE), the verb *come* has shifted from the auxiliary *be* to the auxiliary *have* in the period covered by ZEN, as Fig. 8 illustrates.

As fluctuation is considerable, and as we wanted to extend to other verbs, we have also tested *go, arrive*, and *enter*, and found similar, slightly less clear trends. If all verbs with the auxiliary *be* are searched, the majority of hits are passive forms. Without manual validation of the hits, only intransitive verbs (*come, go, arrive*) or verbs that are hardly used in the passive (*enter*) can be investigated fully automatically.

5. Conclusion

We have described the automatic annotation of ModE corpora, such as ZEN and ARCHER. We have evaluated the performance of the spelling normalization tool VARD and improved its

performance on Early and Late ModE text. We have evaluated the performance of Pro3Gres, our dependency parser, and improved its performance by using statistical data and semantic resources. We have shown that these improvements can constrain the higher ambiguity observed in earlier texts. We have presented two short pilot studies illustrating applications of using automatically parsed historical corpora.

So far we have shown the potential of syntactically annotated data on the ZEN corpus. We expect the larger ARCHER corpus and Old Bailey Corpus to yield even more interesting results. In the future, application of automatic syntactic annotation to resources like the 134 million Old Bailey proceedings will open new possibilities for historical linguists that would be beyond the reach of small manually annotated corpora.

References

Andersen, Ø. E., Nioche, J., Briscoe, T., and Carroll, J. (2008). *The BNC Parsed with RASP4UIMA. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.

Aston, G. and Burnard, L. (1998). *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baron, A. and Rayson, P. (2008). *VARD 2: A Tool for Dealing with Spelling Variation in Historical Corpora. Proceedings of the Postgraduate Conference in Corpus Linguistics*. Birmingham: Aston University, 22 May 2008.

Buchholz, S. and Marsi, E. (2006). *CoNLL-X Shared Task on Multilingual Dependency Parsing. Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*. New York: Association for Computational Linguistics, pp. 149–64, June 2006.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Denison, D. (1998). Chapter 3: Syntax. In Romaine, S. (ed.), *The Cambridge History of the English Language, Volume 4: 1776–1997*. Cambridge: Cambridge Universtiy Press, pp. 92–329.

Fries, U. (2010). Sentence length, sentence complexity and the Noun Phrase in the 18th-Century News Publication. In Kytö, M., Scahill, J., and Tanabe, H. (eds), *Language Change and Variation from Old English to Late Modern English: A Festschrift for Minoji Akimoto*. Bern: Peter Lang, pp. 21–34.

Gildea, D. (2001). *Corpus Variation and Parser Performance. Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Pittsburgh, PA, pp. 167–202.

Görlach, M. (1991). *Introduction to Early Modern English*. Cambridge: Cambridge Universtiy Press.

Grewendorf, G. and Sternefeld, W. (eds), (1990). *Scrambling and Barriers*. Amsterdam/Philadelphia: Benjamins.

Grover, C. (2008). *LT-TTT2 Example Pipelines Documentation*. Edinburgh: Edinburgh Language Technology Group, July 24.

Hundt, M. (2004). Animacy, agentivity, and the spread of the progressive in Modern English. *English Language and Linguistics*, **8**(1): 47–69.

Hundt, M., Denison, D., and Schneider, G. (2012a). Retrieving relatives from historical data. *Literary and Linguistic Computing*, **27**(1): 3–16.

Hundt, M., Denison, D., and Schneider, G. (2012b). Relative complexity in scientific discourse. *English Language and Linguistics*, **16**(2): 209–40.

Hunston, S. and Francis, G. (2000). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia: Benjamins.

Kaplan, R. M., Maxwell, J. T.III, Holloway King, T., and Crouch, R. S. (2004). Integrating finite-state technology with deep LFG grammars. In *ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP* (ComShaDeP 2004), Nancy, France.

Köstenberger, A. and Patterson, R. D. (2011). *Invitation to Biblical Interpretation: Exploring the Hermeneutical Triad of History, Literature, and Theology*. Grand Rapids: Kregel.

Lehmann, H. M. and Schneider, G. (2009). Parser-based analysis of syntax-lexis interaction. In Jucker, A. H., Schreier, D., and Hundt, M. (eds), *Corpora:Pragmatics and Discourse:Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29), Ascona, Switzerland, 14–18 May 2008* (Language and computers; no. 68). Amsterdam: Rodopi.

Lehmann, H. M. and Schneider, G. (2012a). BNC Dependency Bank 1.0. In Ebeling, S.O., Ebeling, J., and Hasselgård, H. (eds), *Studies in Variation, Contacts and Change in English, Volume 12: Aspects of*

Corpus Linguistics: Compilation, Annotation, Analysis. Helsinki: Varieng.

Lehmann, H. M. and Schneider, G. (2012b). A large dependency bank. In *LREC 2012 Conference Workshop "Challenges in the Management of Large Corpora"*, Istanbul, Turkey, pp. 23–28, 22 May 2012.

López-Couso, M., Aarts, B., and Méndez-Naya, B. (2012). Late Modern English syntax. In Bergs, A. and Brinton, L. J. (eds), *Historical Linguistics of English: An international handbook*, vol. I. (Handbooks of Linguistics and Communication Science [HSK] 34.1). Berlin: Mouton de Gruyter, pp. 869–87.

Mair, C. (2003). Gerundial Complements After Begin and Start: Grammatical and Sociolinguistic Factors, and How they Work Against Each Other. In Rohdenburg, G. and Mohndorf, B. (eds), *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter, pp. 329–45.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, **19**: 313–30.

Nivre, J. (2006). *Inductive Dependency Parsing*. Text, Speech and Language Technology 34. Dordrecht, The Netherlands: Springer.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). *The CoNLL 2007 Shared Task on Dependency Parsing. Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*. Prague, Czech Republic: Association for Computational Linguistics, pp. 915–32.

Rayson, P., Archer, D., Baron, A., Culpeper, J., and Smith, N. (2007). "Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora". *Proceedings of Corpus Linguistics*, UK: University of Birmingham, pp. 27–30 July 2007.

Rissanen, M. (1999). Chapter 3: Syntax. In Romaine, S. (ed.), *The Cambridge History of the English Language, Volume 3: 1476–1776*. Cambridge: Cambridge Universtiy Press, pp. 187–331.

Schneider, G. (2008). *Hybrid Long-Distance Functional Dependency Parsing*. Ph.D. thesis, University of Zürich.

Schneider, G. and Hundt, M. (2009). "Using a parser as a heuristic tool for the description of New Englishes". In *The Fifth Corpus Linguistics Conference*, Liverpool, UK, pp. 20–23 July 2009, online.

Schneider, G. (2012). "Adapting a parser to Historical English". In Tyrkkö, J., Kilpiö, M., Nevalainen, T., and Rissanen, M. (eds), *Studies in Variation, Contacts and Change in English, Volume 10: Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*. Helsinki: Varieng.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tieken-Boon van Ostade, I. (2009). *An Introduction to Late Modern English*. Edinburgh: Edinburgh University Press.

van Noord, G. and Bouma, G. (2009). *Parsed Corpora for Linguistics. Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, Athens, Greece. Association for Computational Linguistics, pp. 33–39.