Visualizing the development of prose styles in Horse Manuals from Early Modern English to Present-Day English

Thijs Lubbers, Bettelou Los

Journal of Data Mining and Digital Humanities

Visualizing the development of prose styles in Horse Manuals from Early Modern English to Present-Day English

Thijs Lubbers[1], Bettelou Los[2*]
1 Independent Scholar, The Netherlands
2 University of Edinburgh, UK

Visualizing the development of prose styles in Horse Manuals from Early Modern English to Present-Day English

Thijs Lubbers[1], Bettelou Los[2*]
1 Independent Scholar, The Netherlands
2 University of Edinburgh, UK

*Corresponding author: Bettelou Los B.Los@ed.ac.uk

Abstract
This paper offers a data-driven analysis of the development of English prose styles in a single genre (instructive writing) dealing with a single topic (the correct way of feeding a horse) in 13 texts with publication dates ranging between 1565 to 2009. The texts are subjected to three investigations that offer visualizations of the findings: (i) a correspondence analysis of POS-tag trigrams; (ii) an association plot analysis; (iii) hierarchical clustering (dendograms). As the period selected – Early Modern English to Present-Day English – does not involve any major changes in English syntax, we expect to find developments that are predominantly stylistic.

keywords
historical corpora; text analysis; stylistics; correspondence analysis; stylometry; n-grams; visualization techniques

INTRODUCTION

The present paper attempts a data-driven investigation into the historical development of English writing styles. When cultures develop writing, successful communication – normally facilitated by immediate context, body language, intonation, cues from the audience, etc. – will have to rely on words only, and the conventions of speech will have to be modified to face this challenge. There was a general increase in literacy levels throughout Western Europe in the 15th C, so much so that the market for books became large enough to sustain the development of printing presses (Caxton famously set up his printing press in Westminster in 1476). The availability of more texts in turn further fuelled literacy rates. The spoken and the written word diverged, readers became familiar with conventions associated with the various genres of writing, and would model themselves on these conventions when they themselves became authors. The changing styles of the Early Modern Period (ca. 1500-1750) was further driven by conscious stylistic and rhetorical choices made by influential writers (see e.g. [Lenker 2010]. Some of Lenker's examples illustrate how deliberate stylistic efforts in the 18th-century have partly given shape to why, from a modern point of view, Chaucer's *Treatise on the Astrolabe*, an instruction manual on the use of this instrument written in 1391, seems so markedly different from Adam Smith's 18thC prose, and what differentiates both from most of what has been written since [Lenker 2010, 233ff]. This suggests that the Early Modern English period should be a fruitful object of study for an investigation into changing styles.

Another reason why style can be studied to advantage in this particular period is the fact that it does not involve any major changes in English syntax. The word order upheavals of the OV/VO change in Early Middle English, and the loss of Verb-Second as a dominant word order pattern

in the 15thC, have been completed, which means that any other syntactic developments that we find are more likely to represent stylistic than syntactic change.

The explorations in this paper are data-driven. When faced with the choice of which elements to study, researchers of necessity have to do a bit of bootstrapping to predict which features will be significant – and this selection tends to proceed initially from intuitions and assumptions. While this approach can be fruitful particularly for charting large-scale diachronic developments (e.g. [Biber & Finegan 1989, 1997]; [Pahta & Taavitsainen 2011]; [Taavitsainen & Pahta 2004], it also presents researchers with a logical dilemma. [Stubbs 2005] refers to this as the "Fish fork", after [Fish 1980]:

> Either we select a few linguistic features, which we know how to describe, and ignore the rest; or we select features which we already know are important, describe them, and then claim they are important. Since a comprehensive description is impossible, and since there is no way to attach definitive meanings to specific formal features, stylisticians are apparently caught in a logical fork [Stubbs 2005, 6]

Although Fish's concerns about the relationship between (literary) text and interpretation are somewhat removed from the quantitative linguistic work in diachronic investigations, the logical dilemma is nevertheless valid. Stubb's Fish fork serves as a warning that working (by necessity) from a pre-determined set of assumptions about what constitute relevant linguistic features means that such investigations may well uncover high frequencies of these features, but that such findings can be unsatisfactory because unsurprising. A data-driven, exploratory approach can aid in detecting latent patterns that will only present themselves as specific to a particular writing convention in a specific period after a statistical computational investigation. A text corpus gives us a different perspective: "language looks rather different when you look at a lot of it at once" [Sinclair 1991, 100]. In our case, this different perspective is achieved particularly through visual representations of the data.

Some of the techniques demonstrated in this paper derive from the research field of stylometry, or computation stylistics, the investigation into recurrent idiosyncratic patterns of language. Authorship attribution research, the usual goal of stylometric investigations, applies such techniques in order to distinguish between authors, on the basis of the hypothesis that "every writer has a unique and verifiable style" [Rudman 2006, 611], which "can be understood as the totality of all the conscious and subconscious choices he or she makes during the process of writing" [Tyrrkö 2013, 186]. Conscious control of one's writing style requires a high level of metalinguistic awareness, and will vary per author. Forensic stylometric investigations particularly rely on subconscious choices, a "stylistic fingerprint" that is unique to a particular individual [see e.g. Holmes and Kardos 2003]. In a way, our purpose puts this methodology on its head: we do not want to just discriminate between texts; we want to use stylometric methodology to investigate what it is about the prose writing of particular periods that gives the sense of a shared style, as a way of revealing the development of writing style conventions. Our task is a text classification task rather than an authorship attribution task, but the methodology behind both approaches is comparable (cf. [Argamon, Koppel and Avneri 1998].

Our claim of the investigation being data-driven requires a qualification. As our investigation is focusing on elements at the intersection of style, discourse and syntax, we are not taking as our data the lexicon (individual words, lemmas) but the underlying layer of morphological information (Part-of-Speech tags, POS tags for short). Strings of POS tags will be informative enough to recover the syntactic structure of the clause. This means that we cannot present the

analysis as completely data-driven, because the POS tag-set represents a layer of linguistic interpretation which already specifies that nouns are different entities from pronouns, prepositions are distinct from conjunctions, auxiliaries are different from verbs. This is unavoidable, as an analysis requires tools; the findings themselves, of course, will ultimately also have to be expressed using the toolkit of syntactic description: non-finite clauses, passives, complex noun phrases, etc.

I METHODOLOGY

1.1 Corpus

Texts may not only differ because they reflect historical developments but also because they reflect different genres or different subject matter. The corpus compiled for this study, therefore, selected texts from the same genre (instructive writing) dealing with the same topic (how to look after a horse). This topic was chosen because horse manuals were a popular genre, with many different works produced every century. Within those texts, we selected samples that dealt with one particular feature of horse care: feeding. The texts themselves were obtained from the digital repository of early English printed publications, *Early English Books Online* (EEBO), freely accessible web repositories and other sources in the public domain (like http://www.archive.org/ and the Google Books project, http://books.google.com). Focusing on feeding a horse as a single topic allows us to zoom in on "agnates", i.e different ways of constructing clauses to express the same meaning [Halliday 2004]. The texts are listed in Table 1.

Date	Author	Sample size (words)	Utterances (*n*)	Mean utterance length (words)	Source
1565	Blundeville, Thomas	4209	127	33.99	EEBO
1585	Clifford,Christopher	4639	89	52.96	EEBO
1607	Markham, Gervais	4439	100	45.31	EEBO
1618	Baret, Michael	4386	102	43.92	EEBO
1697	Speed, A.	4652	161	29.74	EEBO
1721	Gibson, William	5028	195	26.7	EEBO
1796	Hunter, J.	4480	130	35.28	EEBO
1823	Kirby, Jeremiah	4457	200	23.58	online
c1840	Skeavington, George	4507	187	24.87	EEBO
1886	Fleming, George	4437	193	23.74	EEBO
1921	Matheson, Darley	4599	193	24.57	online
1977	Leighton-Hardman, A.C.	4588	217	21.6	publ. libr.
2009	Davies, Zoe	4308	273	16.13	publ. libr.

Table 1. Descriptive statistics of the texts of our corpus.

The distribution of the texts over the centuries is not ideal, due to the fact that we are restricted to texts with a digital copy in the public domain. The texts by Hunter (1796) and Kirby (1823) are somewhat exceptional for their publication types: Hunter's text is written as a dictionary of farriery, and Kirby's text appears in a 19th-century edition of the Encyclopaedia Britannica. Nevertheless, these texts are included here because they seem to generally correspond to the

other texts in the corpus in terms of subject matter and textual composition: Hunter's text has large enough entries to consider these topics on farriery as paragraphs or small chapters, while Kirby's text is an entry which could easily have appeared
in monograph form, being 155 pages in length.

1.2 Corpus sampling and sample size

Our decision to keep not only the genre but also the topic stable across our texts, we are confronted with the fact that the amounts of text per author given over to "feeding the horse" varies substantially. As n-gram generation requires samples of an equal size, this meant that we had to go with texts that, after lexical stripping, contained exactly 4,000 POS tokens each. The selection of 4,000 tokens per text was taken from the beginnings of the sections on feeding, so includes topic introductions but not necessarily equal amounts of closing sections. This was found to be an acceptable sacrifice, since it would maximally result in one incomplete clause per text sample. We did not attempt to balance the section length of subtopics, like "hay" and "watering" – the size of these sections varied per text.

1.3 Spelling Normalisation

After manually entering the data and some minor cleaning (lower casing, expansion of abbreviations like *y^t*, *&c.* into *that* and *etcetera*, deletion of illegible sections, of punctuation markers around roman and arabic numerals), these texts were standardised for spelling using the Variant Detector VARD 2 [Baron and Rayson 2008). The process did not involve any separate batch training of VARD, and normalisation settings were kept at the standard F-score weight (1.0) for spelling standardisation, combined with a high auto-normalisation rate (80%). As a result, the semi-automated process was restricted to only the most unambiguous items in the standard EModE VARD dictionary (confidence scores above 80%). This effectively meant that there was a high manual involvement in the standardisation of spelling. Unless otherwise stated, all string processing and statistical analyses were carried out using the statistical environment R [R Core Team 2014].

1.4 POS Tagging and trigram generation

After normalisation, the online CLAWS4 tagger in combination with the CLAWS-5 tagset (60+ tags)[1] was used to enrich the data with POS tags. This limited number of tags avoids the fine-grained tagging errors of more elaborate tagsets and boost frequency counts, which increases statistical reliability of the results.[2] For some of the texts, a cross-check was available in the form of manually annotated POS-files in the Penn-Helsinki corpus, which revealed only minor inconsistencies.[3]

Some further post-processing removed markers for quotations and bracketing from the data. All other punctuation markers are part of a single category with tag PUN, as per the standard CLAWS-5 tagset. An alternative could have been to follow the practice of the Penn-Helsinki

[1] See http://ucrel.lancs.ac.uk/claws5tags.html/

[2] For comparison, the same data was also tagged using two other taggers (the CLAWS-7 and the Penn-Treebank NLP tagger available through Python and the Apache OpenNLP tagger in R, package "OpenNLP"), but the CLAWS-5 tagset produced the most robust results.

[3] The Penn-Helsinki .pos files use a tagset of about 90 tags, including 4 for punctuation and another 7 `other' tags, e.g., unknown, CODE, line break, foreignword ; see https://www.ling.upenn.edu/hist-corpora/annotation/labels.htm.

corpora, which distinguish sentence-medial and sentence-final punctuation, but the punctuation system in our earlier texts differs in important respects from PDE practice (the colon sign seems much more of a sentence-final rather than a sentence-medial marker in the earlier samples, to name just one difference), which carried the risk that the data-driven analysis might pick up on variations in punctuation rather than on a linguistic distinction. The remaining tagset had exactly 60 possible POS categories (for an overview of the CLAWS-5 tagset, see http://ucrel.lancs.ac.uk/claws5tags.html).

Example (1), taken from the Baret text (1618), illustrates the various steps of the procedure.

(1) a. [after cleaning special characters]
Now whereas it hath bene a custome to water a running Horse in the house, and to have him drinke but once a day, and likewise to put Liquoras, or such like, into the water to helpe his winde, all these I doe except against, and why?

b. [VARD spelling regularisation]
now whereas it has been a custom to water a running horse in the house, and to have him drink but once a day, and likewise to put liquorice, or such like, into the water to help his wind, all these I do except against, and why?

c. [CLAWS5 tagging]
now_AV0 whereas_CJS it_PNP has_VHZ been_VBN a_AT0 custom_NN1 to_TO0 water_VVI a_AT0 running_AJ0 horse_NN1 in_PRP the_AT0 house_NN1 ,_PUN and_CJC to_TO0 have_VHI him_PNP drink_VVB but_CJC once_AV0 a_AT0 day_NN1 ,_PUN and_CJC likewise_AV0 to_TO0 put_VVI liquorice_NN0 ,_PUN or_CJC such_DT0 like_AJ0 ,_PUN into_PRP the_AT0 water_NN1 to_TO0 help_VVI his_DPS wind_NN1 ,_PUN all_DT0 these_DT0 I_PNP do_VDB except_VVB against_PRP ,_PUN and_CJC why_AVQ ?_PUN

d. [removing lexical material, leaving only POS tags]
AV0 CJS PNP VHZ VBN AT0 NN1 TO0 VVI AT0 AJ0 NN1 PRP AT0 NN1 PUN CJC TO0VHI PNP VVB CJC AV0 AT0 NN1 PUN CJC AV0 TO0 VVI NN0 PUN CJC DT0 AJ0 PUNPRP AT0 NN1 TO0 VVI DPS NN1 PUN DT0 DT0 PNP VDB VVB PRP PUN CJC AVQ PUN

The resulting strings of POS tags served as the basis for the calculation of POS frequency averages as well as the generation of n-grams (i.e., POS grams). N-gram generation was carried out in R using the `RWeka` library [Hornik et al. 2014]. The size of the n-grams was set to three, based both on a number of diachronic studies which report that trigrams strike an optimal balance between linguistic interpretability, statistical power and computational costs [Gries, Newman and Shaoul 2011, 10]. Note that there are no stop signs for the generation of trigrams until the end of each text sample is reached. Trigram generation results in overlapping strings, as exemplified for the string AV0 CJS PNP VHZ VBN AT0 NN1 in (2).

(2) AV0 CJS PNP -- CJS PNP VHZ -- PNP VHZ VBN -- VHZ VBN AT0 -- VBN AT0 NN1

The underlying assumption is that such recurring bundles of three successive elements, even when they are part of more complex grammatical clusters, indicate the rate of occurrence of frequently used constructions and habitual patterns of language use (and, in our case, of structure). The trigrams generated can be grouped and tabulated by frequency. A cumulative list of the POS-trigram frequencies found across these 13 texts serves as the basis for a correspondence analysis (cf. [Benzécri 1973]; [Greenacre, 1984]; [Greenacre 2017]; [Murtagh, 2005], as we will discuss in the next section.

1.5 What is a correspondence analysis?

Correspondence analysis has a wide range of applications, from corpus linguistics (e.g. [Ernestus, Mulken and Baayen 2006]; [Tummers, Speelman and Geeraerts 2012]; [Tummers, Speelman and Geeraerts 2014]) and forensic linguistics (cf. [Bécue-Bertaut et al. 2014]) to stdies in chemistry, ecology, epidemiology, marketing and tourism (cf. [Beh and Lombardo 2014] for an overview). It is an exploratory multivariate scaling or ordination technique and as such is related to Principal Component Analysis (PCA), factor analysis (FA) and multidimensional scaling (MD. Its particular use, however, is in the application of such ordination techniques on data sets that contain frequency counts (i.e., categorical data as found in contingency tables). Given the data set obtained using the procedure above, with cells containing counts for the number of times a certain POS trigram occurs, both in total and per text sample, this provides a ready candidate for the application of correspondence analysis. In addition, as the main purpose of the technique is to uncover and visualise associations between the rows and columns of the contingency table (i.e., POS trigrams and text samples), correspondence analysis can aid in visualising significant clusters in the data across these two sets of variables. Using this technique, we hope to establish correspondences between clusters of texts, in addition to clusters of POS trigrams in the vicinity of such text sample clusters.

If we consider every row or column of a data matrix as a "dimension", the purpose of correspondence analysis is to reduce a high-dimensional space to a small number of significant, latent dimensions. For the particular mathematics behind the reduction of the original numberof dimensions to a few latent dimensions, see e.g., [Greenacre 1984]; [Greenacre 2017]; and [Murtagh 2005]. One of the main advantages of the method is to facilitate inspection of high-dimensional data by way of a graphical display using a low-dimensional plot. This plot is usually restricted to the two or three latent dimensions that account for most of the difference in the intra-row and intra-column distances (calculated as chi-squared distances). The total variance in the data matrix is measured by what is known as the "inertia", calculated on the basis of the relative differences between rows and columns in terms of observed and expected frequencies. Since every latent dimension contributes to inertia, the output of the process provides the principal inertia (or *eigenvalue*) for each dimension, as well as a percentage for how much it contributes towards total inertia.

Based on the data in the rows and columns in the contingency table, two distance matrices are drawn, much like the distance matrix between cities in a geographical map. In correspondence analysis, one matrix contains the distances of rows by rows (here: the text samples in our corpus), while another contains the distances of columns by columns (here: POS trigrams). The points drawn in the subsequent plots reflect the distances between points in these matrices. For example, rows that are far removed in the distance matrix are also far removed from each other in the plot (i.e., these texts are very dissimilar in terms of trigram frequencies), and vice versa for rows (texts) that show small distances in the distance matrix. The correspondence analyses for our date are computed in the statistical environment R using the libraries `ca` (cf. Nenadic and Greenacre, 2007; Nenadic and Greenacre, 2014) and, as a cross-check, `languageR` (Baayen, 2014). A step-by-step guide to (re)producing correspondence analysis plots can be accessed at the following stable URL: http://datashare.is.ed.ac.uk/handle/10283/2912,

1.6 Accumulation of error

N-gram generation is fully automated and not subject to errors. Any error in its results will have been caused by errors in previous steps. Our concern is not so much with errors in the transcription phase; although some errors undoutedly remain, these are unsystematic human errors, and their effect is going to be negligible in comparison to the danger posed by more systematic errors that might have been introduced in the successive rounds of spelling normalisation and POS tagging. In theory, the axis in the dimensional reduction produced by correspondence analysis could not reflect the historical development of English style, but rather the degree of tagging error, as the early texts represent data points that are relatively high in error, whereas for the most recent texts the tagger is, naturally, much more accurate. We have tried to pre-empt it by carrying out regular cross-checks of the VARD-ed and POS tagged text files. Other scholars working in this area have noted that that "as long as a language analysis system is consistent in the errors it makes, machine learning techniques can pick up on correlations between linguistic features and style even though the label of a linguistic feature (the 'quality' it measures) is mislabeled" [Gamon, 2004].

Another advantage of automatic tagging is that it allows our corpus to be compared to other texts, which, tagged with the same tagset labels by the same tagger, can be used as a reference corpus. This would have been much more difficult if our corpus was tagged manually. [Tyrrkö 2013, 191], reporting a pilot study by [Hiltunen and Tyrrkö 2012] using a similar methodology to ours, notes that most errors were found particularly in the more fine-grained level of tags (inherent in the CLAWS-7 tagger compared to the CLAWS-5 ; also see footnote 2). For in-depth discussions of the degree of error allowed in POS tagging, see [Mair et al. 2002] and particularly [Rayson et al. 2007] for the accuracy of POS tagging in combination with the use of VARD as applied on Early Modern English texts.

II Correspondence analysis

2.1 Rationale

The first correspondence analysis is based on the 309 most frequent trigrams which cover 50% of the trigram tokens in the current data set. With the tag set used for the current experiment, the number of possible POS trigrams amounts to 216,000 (60^3), so it is surprising that the total number of different tag trigrams found in the current data set is 7,305, which is only a fraction($\approx$ 3.38%) of the total number of possible tag permutations. Although it is clear that some tag combinations are unlikely to occur in natural language (e.g., CJC-CJC-CJC), it is nevertheless striking that the entire data set is covered by less than 4% of possible trigrams. Of these tag combinations, the number of hapax legomena (tag combinations that occur only once) amount to 3,374. In other words, nearly half (i.e. 46.18%) of the possible tag combinations that are attested in the current corpus occur only once in the entire data set. Another 1,125 tag combinations occur only twice (i.e., dis legomena POS trigrams), comprising 15.40% of the possible tag combinations. Surprisingly, these figures roughly correspond to the typical "Zipfian" distribution of lexical items in natural language data, with the occurrence rate of hapax legomena ranging between 40-50%, and the rate of dis legomena between 10-15% (cf. [Kornai 2008].

Given this "Zipfian"-like distribution, and the high percentage of hapax and dis legomena trigrams, we needed to restrict the data for the correspondence analysis to the more frequenct items, for the purposes of computational efficiency (to avoid low frequency cell counts). Knowing what we know about style, our assumption is that it will be particularly the more frequently recurring patterns, rather than the large number of combinations appearing only once

or twice in a text or group of texts, that will cumulatively add up to what intuitively feels to be a particular style of writing. As a first try, we decided to include only those trigrams that make up half of the observations in the full data set. The total number of observations (tokens) over the 7,305 POS trigram combinations (types) in the data matrix is 51,974, which corresponds to 3,998 trigrams per text – the number of trigrams generated is necessarily n-2 of the number of tags in the sample, as the first and last element of every 4,000 POS-sample cannot feature as a central element of a tag trigram – i.e., no trigrams can be generated on the basis of these two tags, as there is no prior element for the first, and no following element for the last POS tag. Of the total of 51,974 POS trigrams tokens, the 309 most frequent trigram types in the current data set turned out to cover 26,005, i.e. just over half. We saw at the beginning of this section that the total number of different tag trigrams found in the current data set (7,305) is only a fraction (≈ 3.38%) of the total number of possible tag permutations (which is 216,000 (60^3)) ; with the 7,305 reduced to 309, it is surprising to find that half the the observations in the data are covered by less than one percent (309/216,000*100 = 0.1430556) of all possible combinations of three consecutive POS tags. The reason is, of course, that the strings of POS tags are not randomly distributed but reflect the underlying sentence structure that the syntax makes available.

2.2 A correspondence analysis: POS trigrams, 50% of tokens

Dimension	value	%	cum%	scree plot
1	0.171576	34.8	34.8	*********
2	0.068336	13.9	48.6	***
3	0.058488	11.9	60.5	***
4	0.035653	7.2	67.7	**
5	0.032732	6.6	74.4	**
6	0.026308	5.3	79.7	*
7	0.021909	4.4	84.2	*
8	0.019884	4	88.2	*
9	0.017303	3.5	91.7	*
10	0.015024	3	94.7	*
11	0.013245	2.7	97.4	*
12	0.012698	2.6	100	*
Total	inertia	0.493153	100	

Table 2. Scree plot for correspondence analysis (50% of tokens).

A scree plot of the latent dimensions in the data and their (cumulative) percentages shows that the total number of dimensions is 12 (cf. Table 2). Using the method for determining significant dimensions proposed by [Bendixen 1996, 26], the expected average "inertia" (= the extent to which a particular dimension accounts for the variation in the data) is in our case 100/(14-1) ≈7.69% for the rows (text samples), and 100/(309-1) ≈ 0.32% for the columns (POS tag trigrams). Only the first three dimensions have percentages above the highest of these values. These dimensions explain respectively 34.8%, 13.9% and 11.9% of the inertia, to a total of 60.5%.

2.2.1 Inspection of the symmetric biplot

The symmetric plot that is drawn on the basis of this correspondence analysis is provided in figure 1. The POS trigram data points are indicated by red triangles, and the texts are indicated

by blue dots. The shading indicates the relative contribution of data points to the dimensions, with darker shades indicating a higher contribution.

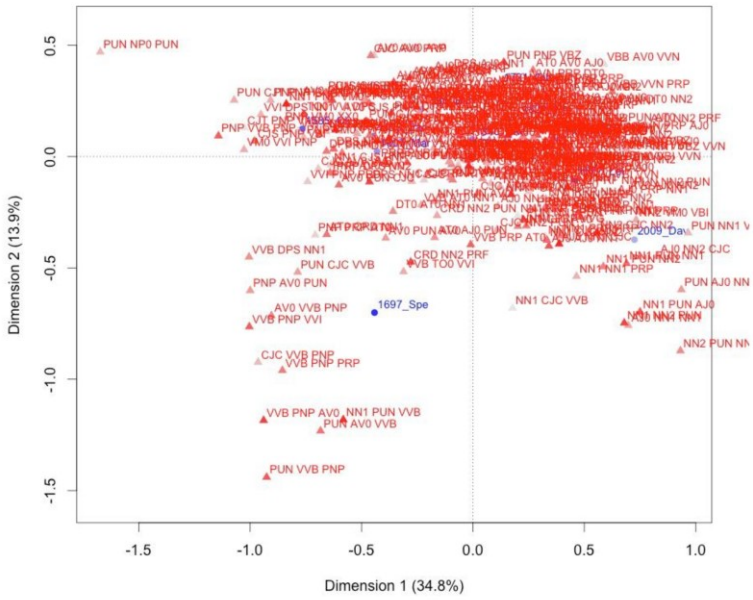

Figure 1. Symmetric plot of trigrams and sources (50% of tokens).

Figure 1 is so cluttered that the labels indicating the sample texts are almost completely obscured by the cloud of tags representing the data points of the POS trigrams. Omitting the POS trigram labels, as in Figure 2, makes the visualization of the positions of the texts clearer.

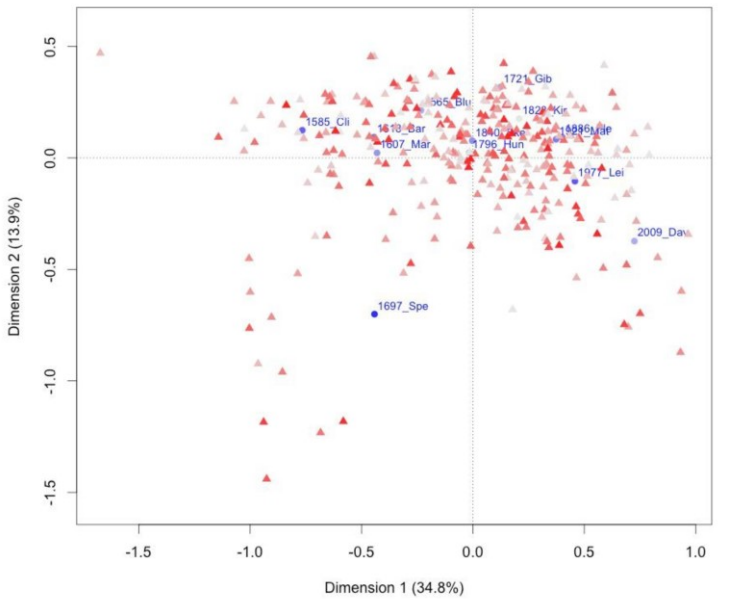

Figure 2. Symmetric plot of trigrams and sources (50% of tokens, no POS labels).

The labels of the axes show the dimensions and their principal inertias. Because correspondence analysis seeks to reduce the geometry of a number of multi-dimensional points to a two-

dimensional display, these two principal inertias indicate how accurate the two axes in the current plot are in accounting for the inertia in the data set. The two dimensions plotted here account for 48.6% of the variation.

What can be gleaned from Figure 2 is that the data points do not appear to be plotted completely at random: there appears to be a progression, with the 16th- and 17th-century texts clustering in the upper left quadrant, the 18th- and 19th-century texts in the centre near the origin, and the contemporary text (Davies' text, from 2009) on the far right. This slope is far from perfect, however: the Blundeville text from 1565 is positioned near the centre of the plot, away from its contemporaries, and the mid-19th-century text by Skeavington is similarly displaced. Other peculiarities include the relative position of Gibson compared to the other 18th-century text by Hunter. The 17th-century manuals all appear with similar coordinates on dimension 1, although their position on dimension 2 varies, with the texts of Baret and Markham appearing with positive coordinates, and the text by Speed with the most negative coordinate on dimension 2 in the entire corpus.

2.2.2 Discussion of dimension 1

The first dimension is associated with the roughly chronological ordering of the text samples, and is also the dimension which accounts for most of the inertia in the dataset as a whole. It suggests that this factor is related to the historical development of the style of this genre. The question is whether the POS trigrams and their location in the plot will give us an idea as to what sets the texts of each century apart from the others. The plot suggests that trigrams appearing towards the left end of the cloud of points seem to contain conjunction (CJC) tags with considerable frequency. On the other hand, trigrams in the bottom right quadrant of the plot seem to contain a fair amount of noun tags; both singular (NN1) as well as plural (NN2) noun tags, or a combination of both. But much of the information is lost in the general cloud of POS tags.

More specific information regarding these findings can be obtained from the numerical output of the correspondence analysis, particularly the quality, contributions and (squared) correlations of the individual columns (POS trigrams) in relation to the factors. Table 3 lists ten trigrams with high correlations on the first dimension, five for each direction (positive or negative) in the correspondence analysis, as ranked by correlation.

#	POS trigram	Quality	Correlation	Coordinate
23	PRP-DPS-NN1	839	826	-569
58	TO0-VVI-PNP	829	783	-571
72	VM0-VVI-PNP	945	817	-1012
12	DPS-NN1-PUN	821	803	-657
125	CJS-PNP-VBI	830	754	-980
31	VVN-PRP-AT0	877	840	+480
133	VBZ-VVN-PRP	792	785	+530
253	NN2-PUN-AT0	820	763	+569
206	VVN-PRP-AJ0	762	742	+725
28	AJ0-NN1-PRF	732	721	+562

Table 3. Highest POS trigram correlations on dimension 1 (in permille).

These ten trigrams provide an additional guide to an interpretation of dimension 1. The "quality" score indicates that there is a fair degree of certainty (over 50% for a data point over 500) that the position of the data point is being represented accurately by the dimensions chosen. A high "correlation" means that a the POS trigram in question is strongly associated with the dimension (this is also indicated by the shading in the plot). Coordinates for these trigrams are provided here to indicate whether it is positioned in the positive or negative domain of the horizontal axis, and thus to illustrate whether it is associated with the later (+) or rather the earlier (-) period in the corpus.

The first POS trigram in the negative domain indicates the combination of a preposition (other than *of*), a possessive determiner and a singular noun (#23: PRP-DPS-NN1; cf. example (3)).

(3) I Beseech you shew me what forrage and provender is best for mine horse to eat,... (Clifford, 1585)

Example (4) shows the trigram that is listed as the fourth combination DPS-NN1-PUN, also in the negative domain:

(4) [. . .] sithence you denie me to let my horse bloud in the spring time, which cannot sincke into my head, but to the good, ... (Clifford, 1585)

This combination is also preceded by a preposition,like (3), making it a 4-gram of the form PRP-DPS-NN1-PUN. The second POS trigram with a negative coordinate in the list represents the use of an infinitive marker, an infinitive of a lexical verb and a personal pronoun: TO0-VVI-PNP (#60). It can be found in sentences such as seen in Baret's *Vineyard of Horsemanship*:

(5) a. you shall adde to his Oates Beanes; for they will increase strength and lust, and so keepe him till you intend to hunt him; ... (Baret, 1618)
b. Now for the quantity that you should give your Horse at one time, there cannot be any certaine limitation thereof, but it must bee proportionated according to his appetite; onely be sure to give him his full feeding, for that will keepe his body in better temper, ... (Baret, 1618)

Another POS trigram in the early section of the plot is the combination CJS-PNP-VBI: a subordinating conjunction, a personal pronoun and a infinitive form of *be* (#125). The trigram contains a tag for a (subordinating) conjunction which, based on the plot, we would expect to find quite frequently on this end of the dimension.

(6) if he be laid downe, you shal not onelie your selfe refraine from comming unto him, but also have care no noise or tumult be neare the stable, ... (Markham,1607)

Such examples serve to illustrate the continuative style of these earliest equine manuals, with rather long sentences and a high frequency of conjunctions, either coordinating or subordinating (see also [Burnley 1986] on this type of writing in prose). The last trigram in this list, the combination of a modal verb, the infinitive form of a lexical verb and a personal pronoun VM0-VVI-PNP (#72) we will be able to see more clearly in the plot of the second correspondence analysis below. One example is provided in (7):

(7) Even so do I wishe also that the heye, strawe, or garbage, whereof the horse feedyth all the daye, be gyven hym by lytle and lytle, even as he dothe spende it, and not to

be layde before him all at once, for that will lothe him, and take away his appetyte, ... (Blundeville, 1565)

On the positive end of the scale, the third highest ranking POS trigram in terms of correlation with the first dimension and a positive coordinate also contains a punctuation marker. However, in this case it is in second position in the trigram, and is preceded by a plural noun and followed by an article (#253), found in example (8):

(8) Unless the food contains a sufficient proportion of these substances, the body must be inefficiently nourished, ... (Fleming, 1884)

In this case the punctuation mark separates a subclause from a main clause, but there are many other configurations which would give rise to this trigram, like lists, so it is difficult to interpret the significance of this trigram. This is not the case with the other four trigrams. Three of them contain past participles of lexical verbs (VVN); the first one is a combination of a past participle (VVN), a preposition other than *of* (PRP) and an article (AT0; #31):

(9) a. problems can arise if they are brought into a stuffy loose-box on a hot summer evening (Leighton-Hardman, 1977)
b. organic fertilisers are released at a slower rate than artificial fertilisers (Leighton-Hardman, 1977)

Although such sentences may not strike the modern reader as particularly remarkable, the use of the past participle in such examples turns out to be important markers of style in the later part of our corpus. They are all likely to reflect passive constructions, a well-known feature of contemporary informative prose, particularly scientific writing (e.g., [Biber et al. 1999]; [Halliday 2004]; [Huddleston 1971]; [Swales 1990]). The second one, VBZ-VVN-PRP (#133), represents an -s form of the verb *be* (so either *is* or -*s*), a past participle form of a lexical verb and a punctuation marker, as in (10a-b); the presence of the auxiliary *be* shows that this trigram can definitely be identified as proceeding from a passive construction:

(10) a. Water is lost from the horse's body via urine, feces, sweat and evaporation from the lungs and skin. (Davies, 2009)
b Haylage is preferred for horses in hard work or with known respiratory conditions (Davies, 2009)

The third POS trigram containing a past participle further contains a preposition other than *of* and an unmarked adjective (i.e. not a comparative or superlative; #206: VVN-PRP-AJ0). Two examples, (11a) from Matheson and (11b) from Fleming, may suffice:

(11) a. but all this superfluous flesh has to be got rid of by about the end of October, being substituted by hard muscles for soft ones (Matheson, 1921)
b. Should an excess of this material be given for any length of time, and no requirement for it be created by corresponding increase of work, disease must result. (Fleming, 1884)

A second example from Matheson (1921) illustrates the last POS trigram highlighted for this region on dimension 1, AJ0-NN1-PRF (#28). It hints at the importance of large nominal clusters containing both pre- as well as postmodification on this end of the dimension, with a

combination of an unmarked adjective, an singular noun and an *of*-preposition (cf. also the phrases *corresponding increase of work* in (11b) and *sufficient proportion of* in (8)).

(12) The comparatively small size of a horse's stomach, and the short time that food remains within it, clearly indicate ... (Matheson, 1921)

The rise of such complex noun phrases have been discussed in depth by [Halliday 2004] as the transition from a "Doric" style to an "Attic" style of science writing. Typical for the Attic style is the development of the "grammatical metaphor", "a participating entity, a process, and then a second entity participating directly or circumstantially" [Halliday 2004, 104]. His example is given in (13) (adapted from [Halliday 2004, 105]); the Doric style is exemplified by (13a), its Attic agnate by (13b):

(13) a. If you invest in a new facility for the railways you will be committing funds for a long term
b. Investment of funds in a rail facility implies a long-term commitment

Quite characteristic of the construction in (13b) is that processes, expressed by verbs in (13a), are expressed by nouns, witness *investment* and *commitment* in (13b), which are nominalisations of verbs. This allows the grammatical metaphor to capture the relationship between two processes in a highly abstract and concise way, and allowing the use of a third process (encoded by the main verb *implies* in (13b)) to describe or further specify this relationship.

2.2.3 Discussion of dimension 2

Providing an interpretation of the second dimension, realised as the vertical axis in the plots above and accounting for 13.9% of total inertia in this CA, proves somewhat more difficult. For both positive and negative coordinate values, Table 4 lists three POS trigrams with the highest correlation on this dimension.

#	POS trigram	Quality	Correlation	Coordinate
54	NN1-PUN-VVB	931	588	-1182
50	PUN-VVB-PNP	940	584	-1440
129	PUN-AV0-VVB	970	656	-1232
212	VM0-XX0-VBI	480	429	+312
41	PRP-DT0-NN1	500	367	+272
201	DPS-AJ0-NN1	397	368	+386

Table 4. Highest POS trigram correlations on dimension 2 (in permille).

A general problem for the trigrams listed in the positive domain of dimension 2 is that these show a fairly low quality (i.e., 500 points and lower), which indicates that there is a high probability that their position in the correspondence plot is not entirely correct. Dimension 2 primarily seems to mark the distinction between an outlier (Speed, 1697) and the other texts in the corpus, rather than a diachronic progression of earlier to later texts.

All three negative tags contain a punctuation marker, which might at first suggest that this axis is related to idiosyncratic practices of punctuation or sub-register-specific conventions of usage,

for example heavy versus light punctuation (cf. [Nunberg, Briscoe and Huddleston 2002]). The fact that all three negative tags can be found near the sample by Speed, the outlier, however, makes it much more likely that these tags represent a strong association with this particular text. The sequence of trigram #50, PUN-VVB-PNP, i.e. a punctuation marker, a base form of a lexical verb and a personal pronoun reflects Speed's typical sequences of instructions, such as in (14).

(14) . Give him a due proportion of provender, litter him very well, and let him be clean rubbed down... (Speed, 1697)

The VVB here reflects an imperative, and the pronoun reflects the use of *him* to refer to the discourse entity of the generic horse. Both tags seem particularly indicative of Speed's recipe-like horse manual, and the reason why he is such an outlier. Two of the three trigrams probably represent overlapping sequences of NN1-PUN-VVB and PUN-VVB-PNP of the 4-gram NN1-PUN-VVB-PNP (*provender, litter him* in (14)). Speed's manual has a remarkable 'recipe-like' character, even in comparison to other sample texts in the corpus, so that he produces these sequences much more frequently than the other authors.

For tags in the positive region, it is particularly the use of modals and pronouns which stands out. Table 4 contains two of such POS trigrams, e.g., #212: a modal verb, negative marker and a *be* infinitive, as in *shall not be* (Clifford 1585) (VM0-XX0-VBI) and the combination of a possessive determiner (e.g., *your*, *his*), general adjective and a singular noun, as in *his proper place* (Blundeville, 1565) (#201: DPS-AJ0-NN1). The third POS trigram with a high correlation on dimension 2 has the form of a preposition, general determiner (which means a demonstrative, like *these*, *some*, as articles have their own POS tag, AT0) and a singular noun (#41: PRP-DT0-NN1). Our corpus shows a decrease in frequency for this particular POS trigram after the onset of the 20th-century.

(15) a. and take heede when you will swim your horse in this sort, that you bridle him with a watering bit or snaffle, or else with a paire of false raines at his ordinarie bit, ... (Clifford, 1585)

b. with this exercise and sharp diet, I haue in short space made mine horse so strong of stomacke, that he woulde eate eight handfulles a daie ... (Clifford, 1585)

What the correspondence analysis has managed to pick up on here is the fact that there is a subtle change with respect to how "given" information – information that links back to previous referents in the discourse, which is the function of the demonstrative in DT0 – is positioned, as well as how it is expressed. In (15a), the prepositional phrase, which means 'in this way' and refers back to a method explained in the immediately preceding text, is positioned at the end of its clause, which is fine from the perspective of syntax but not ideal from the perspective of information structure, where the natural flow of information is from given to new; the earlier manuals are much more likely to be less strict about information flow. The other prepositional phrase, in (15b), in clause-initial position, is fine in terms of the easy flow of information (as the clause now starts off with given information) but we know from other work ([Pérez-Guerra 2005]; [Los and Dreschler 2012]) that clause-initial prepositional phrases are increasingly dispreferred as encoders of given information; links to the previous discourse are either expressed by subjects (cf. an agnate like *This exercise and strict diet has given my horse such a strong stomach that*...) or by clauses (cf. an agnate like *By maintaining this exercise and strict diet, I have in a short space*...).

2.3 A correspondence analysis: POS trigrams with 100+ count

The earlier correspondence analysis was done on a selection of POS trigrams determined by frequency (only including the most frequent 309 trigrams), based on the assumption that the "look" of a stylistic profile of a particular period will be based on the structures and stylistic choices that occur with some frequency, and we had found that the number of hapaxes was quite high. This section will use an even smaller set of POS trigrams, i.e. only those for which the cumulative cell frequency in the corpus lies at 100 observations or more. To compare, [Ernestus, Mulken and Baayen 2006] only considered the top 35 trigrams of their corpus. These 58 POS trigram types still cover approximately a quarter of the tokens in the data (23.50598%). That is, reducing the number of types by some 80% (from 309 to 58 types) brings the number of underlying tokens down by roughly only half (and recall that the total number of types, including hapax legomena, lies at 7,305). In particular, the plot in Figure 3, based on these 58 most frequent trigrams offers a better visual inspection than Figure 1.

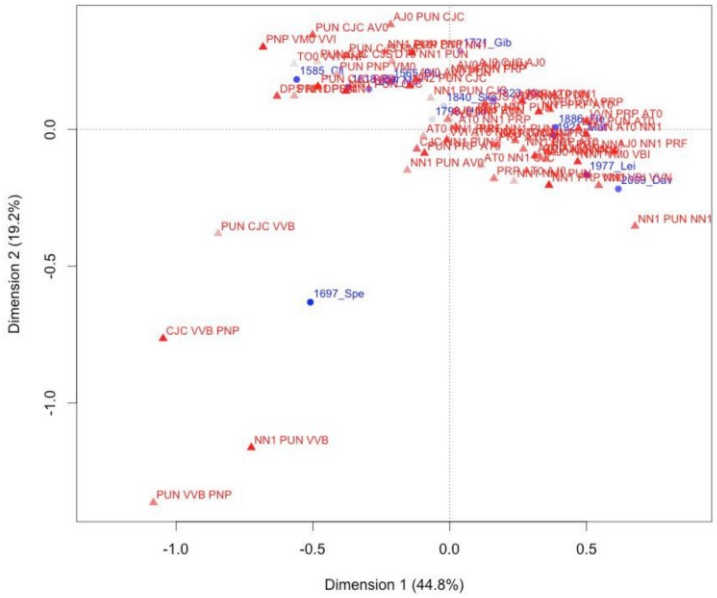

Figure 3. Symmetric plot of trigrams and sources (100+ observations).

The symmetric plot of the correspondence analysis, Figure 3, shows a similar pattern to the ones of Figures 1 and 2. Both patterns point to a similar overall positioning, revealing clusters of texts and tag trigrams that are stable over both subsets of the data (i.e., roughly 25 % and 50% of tokens), but some of the POS trigram labels are now more easily identifiable in the plot.

The results for dimension 1 are similar to those of the earlier correspondence analysis. For dimension 2, the trigrams that define the outlier in the bottom left-hand quadrant, Speed's text, now show up more clearly. Two (PUN-CJC-VVB, with 39 observations in Speed out of a total of 134 in the corpus, and CJC-VVB-PNP, with 41 observations out of 100) can be identified as overlappings of the 4-gram PUN-CJC-VVB-PNP, as in (16).

(16) ....., and give him three of them a day (Speed, 1697)

The other two trigrams are similarly overlappings of a 4-gram, PUN-CJC-VVB-PNP; they were discussed in the previous section (see example (14)).

Dimension 2, flagged up in the earlier correspondence analysis as primarily signifying Speed versus not-Speed, now also becomes informative for the not-Speed group. The cloud of tags shows that the earlier texts are distinguished from the later group by tags containing either a coordinating conjunction (CJC) or an adverb (AV0) after a punctuation marker (PUN); an example of the latter is given in (17):

(17) let him be watered, and that wilbe about the ix houre of the day, and then cast him an other bottel of heye, ... (Blundeville, 1565)

Another tag associated with the early group is a personal pronoun with a modal auxiliary and a lexical verb infinitive (PNP-VM0-VVI):

(18) Now for the quantitie which you shall allow; I thinke for great Horses, or Princes or Gentlemens privat saddle horses, ... (Markham, 1607)

Pronouns and modal verbs were also noted as a feature in the previous correspondence analysis (see e.g. example (7) above), and can be related to the more personal (as opposed to impersonal) Doric style (see also *you will be committing funds* in (13a) above).

A feature that is picked up in this correspondence analysis as significant for the later group is the singular noun before and after a punctuation marker (i.e., NN1-PUN-NN1):

(19) ... legumes such as alfalfa contain higher amounts of protein, calcium and magnesium for example. (Davies, 2009)

So although dimension 2 primarily expresses the difference between Speed and not-Speed, there is also a trace of a diachronic difference.

The final point to note about both correspondence analyses was that dimension 3 also made a significant contribution (11.9% of total inertia in the 50% tokens correspondence analysis, 10.1% in the 100+ observations one). That contribution will become relevant in section IV, and will be mentioned there.

III Association plot

This section reduces the set of POS trigram types even further – to the 10 most frequent POS trigrams. The association plot of figure 4 visualizes the degree to which these trigrams are associated with each individual text. These 10 most frequent POS trigrams together account for approximately 8.77% of all trigram tokens in the data.

Association plots do not visualise the absolute frequencies of trigrams per text but rather their residuals (i.e. the difference between the observed and expected frequencies, as calculated on the basis of the row and column totals). For example, a bar in green above the centre line indicates that a trigram has more observations in a particular text than would be expected based on the average across the corpus. As a result, its residual is positive. A bar below the centre line in red indicates that there are fewer observations for this trigram in a text than expected. Because the graphical output in figure 4 does not allow the plotting of labels for all text samples (horizontal) and POS trigrams (vertical), these labels can be derived from Table 5 below it (with

the first row in the table corresponding to the top line in the plot, and so on).Table 5 also provides absolute frequencies for these 10 POS trigrams.[4]

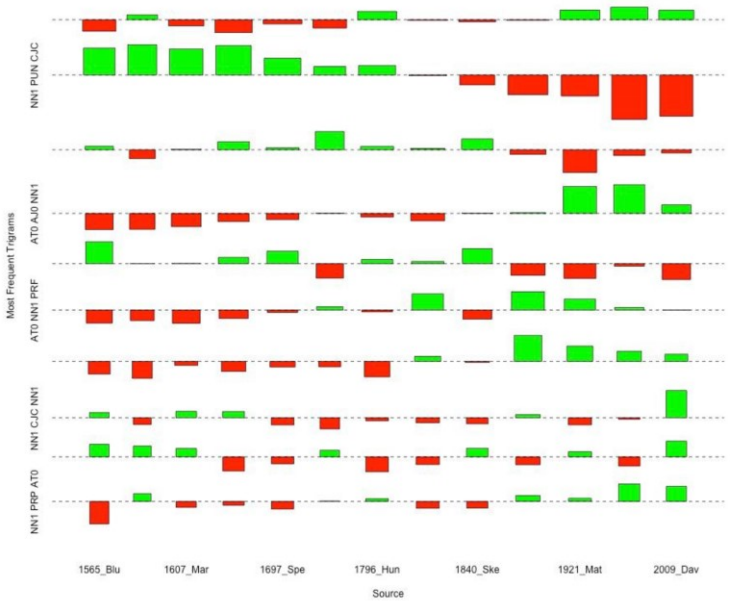

Figure 4: Association plot of 10 most frequent POS trigrams and sources.

POS trigram	Blun	Clif	Mark	Bare	Spee	Gibs	Hunt	Kirb	Skea	Flem	Math	Leig	Davi
PRP-AT0-NN1	36	44	47	45	59	40	74	68	56	75	78	72	62
NN1-PUN-CJC	67	61	73	83	75	51	68	61	42	45	37	7	7
AJ0-NN1-PUN	37	22	39	50	48	50	50	51	52	50	26	37	35
AT0-AJ0-NN1	20	16	26	34	39	34	42	41	41	54	73	67	44
AT0-NN1-PUN	48	26	35	44	53	20	46	47	51	38	30	36	21
AT0-NN1-PRF	19	17	22	29	37	32	38	57	28	64	51	38	32
NN1-PRF-NN1	12	7	20	18	23	17	16	33	24	53	40	32	27
NN1-CJC-NN1	19	10	22	24	17	10	20	20	16	28	18	19	34
PRP-NN1-PUN	23	19	23	11	17	20	12	18	25	20	26	14	27
NN1-PRP-AT0	3	17	14	17	16	16	23	18	15	29	24	30	26
Total *n*	284	239	321	355	384	290	389	414	350	456	403	352	315

Table 5: 10 most frequent POS trigrams in the corpus (*n*).

The association plot reveals exciting results for identifying a chronological progression. For example, the signal observed for the second trigram from the top, NN1-PUN-CJC (the sequence *stomach, and* in (12) would be an example), shows a remarkably positive association with the early half of the corpus and a negative association with the second half: Early Modern texts

[4] A chi-squared test was carried out on this table, which indicates that there is an association between these trigrams and text samples (χ^2= 1,258.4, df=129, p < 0.001, with Cramer's *V* effect size = 1.577366). However, this result should probably be taken with a pinch of salt given that the counts for the trigrams shown here can hardly be assumed to be independent. It can be observed that more than one trigram in this table may overlap with other trigrams listed, for example. Another potential problem is that these 10 trigram cell frequencies and their row and column totals are a subset of a larger contingency table (see [Gries 2014, 376]).

appear with green bars above the centre line indicating that the POS trigram NN1-PUN-CJC is found with a higher frequency than expected in the early section of the corpus, and inversely, is found less than expected in the second half of the corpus. The strength of this signal may come as somewhat as a surprise, as the absolute figures in Table 5 indicate that this POS trigram is frequently used throughout the corpus.

The best example of the opposite distribution is the fourth row, representing the trigram AT0-AJ0-NN1 (e.g., *a young horse*). In the middle period in our corpus this tag occurs more or less with expected frequency (i.e., almost no deviation from the centre line), but can be seen to occur more frequently in the last three texts, as well as somewhat less often in the early half of the corpus.

Somewhat of a similar pattern appears with the sixth and seventh POS trigram, respectively AT0-NN1-PUN in (20a) and AT0-NN1-PRF in (20b):

(20) a. available minerals in the soil, ... (Leighton-Hardman, 1977)
b. matters .... which pertain to the welfare of all classes of horses ... (Matheson, 1921)

Both patterns are attested less than expected in the early half and more than expected in the second half of the corpus. What is particularly interesting about these POS combinations, however, is that these trigrams share the first two of their three tags (i.e., the bigram AT0-NN1), occurring either before a punctuation marker or preposition *of*. The reason is that both POS trigrams may be assumed to occur in larger chunks in combinations with other tags, for example the one in the first row: PRP-AT0-NN1. And indeed, both POS trigram examples displayed here are found in 4-gram chunks of the form PRP-AT0-NN1-PUN (*in the soil*, (20a) or PRP-AT0-NN1-PRF (*to the welfare of* (20b)), which is confirmed by the roughly parallel red and green patterning of rows 6 and 7 in the association plot. Such nominal postmodifications strategies of varying complexity may therefore be particularly indicative of the prose in the later period of our corpus (see also the earlier discussion of the Doric and the Attic style, and example (13a-b)).

It is interesting that the same patterning is not seen in row 1, although its trigram PRP-AT0-NN1 (*to the welfare*) also represents a sub-set of the larger 4-gram. Instead, PRP-AT0-NN1 may well overlap with the trigram NN1-PRP-AT0 of row 10, which does follow a similar patterning to row 1. This suggests that rows 1 and 10 reflect another 4-gram, NN1-PRP-AT0-NN1 (*food for the majority*; Davies, 2009). It seems likely, then, that a thorough inspection of the tokens underlying the PRP-AT0-NN1 trigram of row 1 will uncover that a greater proportion of this particular combination will be preceded by a singular noun (NN1) rather than followed by either a punctuation marker (PUN) or a prepositional phrase headed by the preposition *of* (PRF) in this corpus. Rows 1 and 10 also show little sign of a chronological progression, suggesting that the 4-gram NN1-PRP-AT0-NN1 is not particularly conditioned by the style of a certain period.

Rows 1 and 10 have an interesting mirror image in row 3, the row of trigram AJ0-NN1-PUN (an adjective followed by a singular noun followed by a punctuation mark), as in (21):

(21) or to giue him the intrayles of a Barble or Tench, with whyte wyne. (Gibson 1721)

It is alsmost as if 1 and 10 are the negative of row 3: what is green in 1 and 10 is red in 3, and vice versa. This almost complementary distribution appears to indicate that 1 and 10 on the one

hand and 3 on the other are part of different (idiosyncratic) strategies or styles of writing. Texts that employ the use of an adjective followed by a singular noun and a punctuation marker more than expected, the use of a general preposition, article and singular noun is used less than expected based on average frequencies across the corpus. Adjectives premodify a noun while prepositional phrases postmodify it, and these modifications can have the same function of restricting the referent – white wine refers to a subset of all wines (cf. (21) while minerals in the soil are a subset of all minerals (cf. 20a). English allows agnates where the same information is expressed either by premodifying restriction or by postmodifying restriction (as in 22a), or where a head noun in one agnate can be expressed by a prepositional phrase in another (22b).

(22) a. A young horse/A horse under the age of 2
b. A deficient diet/A deficiency in the diet

The presence of a punctuation marker after the noun in the trigram of row 3 indicates that the noun in that sequence is not postmodified, so this might point to a personal preference of a writer for an premodifying or a postmodifying style. It seems remarkable that such a seemingly clear distributional pattern may be observed in two or three of the ten POS trigrams selected here, given that these 10 combinations reflect only a fraction (≈ 0.14%) of all the trigram types found across the corpus.

IV Hierarchical clustering

4.1 Rationale

Similar to correspondence analysis, hierarchical clustering analysis (HCA) is a multivariate method which seeks to reduce the complexity of a high number of points. As a classification method, HCA tries to describe this number of points in a lower number of classes. [Greenacre 1988, 50] has suggested the routine use of such hierarchical clustering methods when a correspondence analysis is carried out on a contingency table, particularly when there is a suspicion that the data is not being represented well due to the dimensional reduction inherent in correspondence analysis – either because correspondence analysis misses clusters that exist in high-dimensional space, or conversely, creates clusters in low-dimensional space that do not reflect the full complexity of the data set. The use of HCA can corroborate the visual output obtained in the biplot of figures 1 and 2 above. As hierarchical clustering of POS trigrams only tells us which trigrams go together, and not which trigrams go together in which periods, we will focus on the clustering of text samples, to see whether HCA can detect the chronological progression of style on the basis of the trigrams.

4.2 Method

We use an agglomerative hierarchical clustering method here, which means that we build our tree 'from the bottom up', starting out with every text sample as its own group (the 'leaves') and merging groups according to similarity until only one group is left. The opposite, called divisive hierarchical clustering, works from the top down by considering all texts as one group and splitting the trunk (as well as successive groups) on the basis of their dissimilarity, until only groups of one ('leaves') are left. For the current problem, agglomerative clustering seems the more appropriate approach.

For the linking method we use Ward clustering, which entails that clusters are joined in such a way as to minimise the increase of the error sum of squares for each merger (or conversely, the smallest increase in within-cluster variance; see also [Tyrrkö 2013]. [Greenacre 1988, 44] argues that Ward clustering is particularly suitable for approaching the data of a correspondence analysis contingency table, as this linking method remains close to the chi-square statistic for the original distance matrix. As [Baayen 2008, 158] notes, there is a variety of clustering settings and techniques available, and the dendrograms depicted are often chosen on the basis of the clusters which fit a researcher's assumptions best. In our case, we will use Ward's linking method for both HCAs in this section, so that they will only differ in how much of the data they include. A comparison between the two HCAs will then allow us to see exactly what the impact is of the selection of the data.

The distances between points in the contingency table is a weighted squared chi-square distance between cluster centroids (the weight of which is dependent on the profile masses in the clusters; cf. [Greenacre, 1984]; [Greenacre, 1988]. For the distances of the standardised POS frequency HCAs, we use two measures: squared Spearman correlations (cf. [Baayen, 2008, 152] and squared Euclidian distance (cf. [Greenacre, 2017]; [Murtagh, 2005]).

4.3 Results – POS trigram frequencies

The dendrogram on the contingency table containing POS trigrams with 50% or more tokens in the corpus is shown in Figure 4.

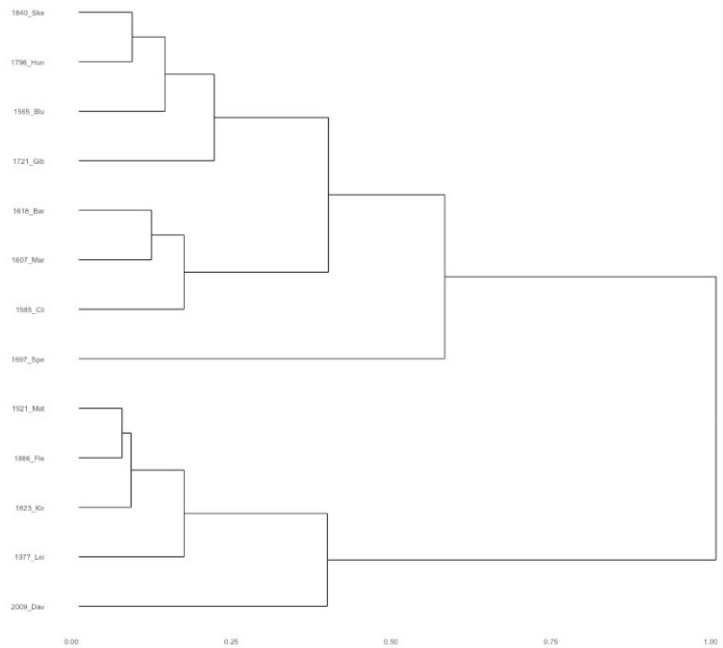

Fig. 5. Dendogram on text samples (POS trigrams, 100+ of observations, method: Ward.

The first thing to mote about dendrograms is that the branches of the dendrogram may be flipped, like those of a crib mobile, without changing the underlying structure of the clustering or the relative distance between text samples (cf. [Oksanen, 2014, 6]). The fact that the dates in the left hand column are not in chronological order is just a consequence of how the branches have flipped. It is not the order of texts items in the left hand column that indicates their relative distance, but their proximity in terms of the branching.

Figure 5 shows that there is a clear division in the texts in our data set: the two groups of manuals which are merged last consist of a Late Modern group (all texts as of the start of the 19th-century, except for Skeavington (c1840)) and an Early Modern group. That the text by Skeavington is grouped with texts published before the 19th-century may be surprising, but this has a correlate in the biplot of the axis with the highest inertias in the corresponding correspondence analysis (cf. figure 3): Skeavington is positioned towards the left on dimension 1 in the plot, and even to left of the origin, whereas all other texts published after the 18th-century (in addition to Gibson (1721)) are positioned in the positive domain of this dimension.

What about our outlier, Speed (1697)? He is shown as a marginal member of the Early Modern group, merged last (and notably, even after Skeavington). The remaining two larger clusters in the early section of the corpus consist of a late 16th-century and early 17th-century cluster of Clifford, Baret and Markham on the one hand, and a more varied group (in terms of date of publication) of the 18th-century texts (Gibson and Hunter) and Skeavington in combination with, somewhat inexplicably, the late 16th-century text by Blundeville. On the basis of the biplot of dimensions 1 and 2 in the correspondence analysis, the position of Blundeville is surprising. But a plot of dimensions 2 and 3 in the correspondence analysis base on 100+ observations (figure 6) shows a cluster of Clifford, Baret and Markham's texts, i.e., the earliest cluster of this HCA, in the top-left quadrant, whereas Gibson (1721), Hunter (1796), Skeavington as well as Blundeville (1565) are positioned in the top-right quadrant.

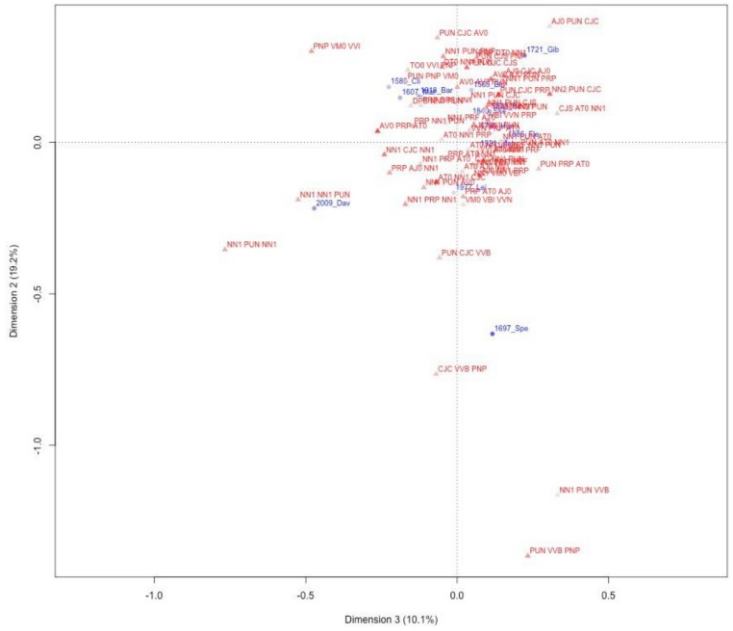

Figure 6. Symmetric plot of dimensions 2 and 3 (100+ observations).

When we carry out an agglomerative hierarchical clustering on the data set which covers a greater proportion (50%) of POS trigram tokens in the corpus, some slight variations in clusters are found (Figure 7).

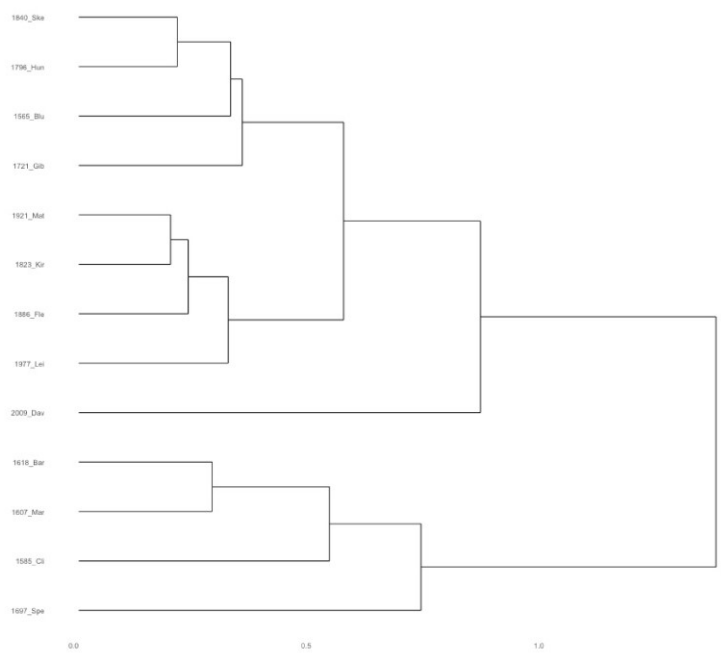

Figure 7. Dendogram on text samples (POS trigrams, 50% of tokens, method: Ward).

Figure 7 shows that the Baret, Markham, Clifford and Speed texts, positioned as a group in the centre of the previous dendrogram, are now positioned at the bottom of the tree. The other large cluster, the top-branch in the dendrogram, consists mostly of the Late Modern English text samples. Davies (2009) now appears as an outlier, as it was in Figure 2. Where the other texts swarmed as a cloud, in a ragged chronological procession, in Figure 2, this dendrogram, based on the same data, is more helpful: the larger Late Modern group is seen to consist of two remaining sub-clusters, one for the texts from the 19th- (barring Skeavington) and 20th-centuries, and one for the the 18th-century texts by Gibson and Hunter, as well as the texts by Blundeville and Skeavington. The similarity of the texts in this last cluster, based on the POS trigrams, is apparently quite robust: in both dendrograms this cluster is built up in the same way, with Hunter and Skeavington merged first, then Blundeville, and then Gibson, before this entire group is merged with other clusters (either the clearly earlier section in the dendrogram in figure 5, or the 19th & 20th-century section in the 50%-of-tokens model in figure 7)

As the difference between these two dendograms is not based on a difference in clustering technique, linking method or distance measure (all of these have remained the same), but only on an expansion of the POS trigram types and underlying frequencies, we can conclude that for HCA, expanding the number of trigrams to include a wider range of possible types, as in Figure 7, is the preferred method. Including more data does not change the structure of the dendrograms, while at the same time giving a more accurate representation of the complexity of the data set.

V CONCLUSIONS

This paper has investigated the development of English written styles by means of a corpus of text samples from different periods all on the same topic, "feeding horses". The data-driven investigation used POS labels rather than lexical words for n-gram generation and correspondence analyses, as a way of identifying the underlying structure of the styles of the individual writers. Correspondence analysis offers a visual display of both angles on the same

data: as POS-trigrams across diachronically ordered texts, and as text samples according to frequency of use for POS clusters, greatly facilitating a visual inspection of the associations between both sets. After establishing that more than half of the trigrams contained hapax legomena or dis legomena, we experimented with two different settings as cut-off points, to include only trigrams that were occurring with some degree of frequency. Remarkably, even on the basis of parts-of-speech, an extremely simplified form of grammatical information, a signal in the data can be picked up that arranges text samples roughly in chronological order, and these patterns turned out to be robust when subjected to various clustering techniques: correspondence analysis, association plots, and hierarchical clustering analysis (dendograms). The variation found in the current data, then, are likely to represent genuine stylistic differences that chart the development of a written style for this particular genre ("instructive writing"), rather than idiosyncratic authorial differences.

The challenge of the results of these visualizations is how to interpret them. Using the techniques described above can only aid in detecting or uncovering latent patterns that cannot be noticed with the naked eye; they cannot replace theory. Some of the findings could be linked to stylistic change noted in the literature, like [Halliday 2004]'s development of the Doric into the Attic style. That data-driven methods confirm what is already known or suspected, is not a bad thing: "Indeed, in developing a new method, it is perhaps better not to find anything too new, but to confirm findings from many years of traditional study, since this gives confidence that the method can be relied on" [Stubbs 2005, 6].

The outliers in the study – Speed (1697) and Davies (2009) – provide interesting food for thought. Even though we tried to keep to the same genre, Speed and Davies are reminders of an insiduous change in register: the instructive, possibly procedural (Speed!) writing of the earlier publications increasingly shades into science writing, and hence starts to reflect the emergence of the more concise style of that register (Davies). This style is slowly convergenced on by writers and readers over the years, as a response to a change in the type of referents that need to be tracked in the discourse: no longer human protagonists, as in narrative, but scientific processes (see particularly [Halliday 2001, 2004]). This change is driven not only by a general increase in scientific discovery, but also by a change in readership over the years – in this case, not only with respect to educational levels (with writers increasingly taking it for granted that readers will have a basic grounding in biology and chemistry), but also a change from a general readership to the much more select stratum of society which is now representing horse ownership. This reflects a change in the role of the horse in society, now taken over by motor vehicles. Trying to keep the genre stable, then. does not mean that the readership of that genre will remain the same over the centuries, and this will also affect the development of style.

Primary sources (source texts)

Baret, M. An hipponomie or the vineyard of horsemanship. monograph. 1618.
Blundeville, T. The fower chiefyst offices belongyng to horsemanshippe: part ii - the order of dietynge of horses. Monograph. 1565.
Clifford, C. The schoole of horsmanship. Monograph. 1585.
Davies, Z. Introduction to horse nutrition. Monograph. Wiley-Blackwell (Chichester), 2009.
Duberstein, K. & Johnson, E. L. How to feed a horse: understanding the basic principles of horse nutrition. Electronic. 2009/2012.
Fleming, G. The practical horse keeper. Monograph. 1884.
Gibson, W. The true method of dieting horses. Monograph. 1721.
Hunter, J. A complete dictionary of farriery and horsemanship. Monograph. 1796.

Kirby, J. Farriery. Chapter in Encyclopaedia Britannica, 6th Edition, edited by Charles Maclaren. Constable and Co (Edinburgh), 1823.
Leighton-Hardman, A. C. A guide to feeding horses and ponies. Monograph. Pelham (London), 1977.
Markham, G. Cauelarice, or the english horseman. Monograph. 1607.
Matheson, D. (1921). The horse: in health, accident & disease. Monograph. C. Arthur Pearson (London), 1921.
Morgan, N. The perfection of horse-manship, drawne from nature; arte, and practise. Monograph. 1609.
Skeavington, G. The modern system of farriery. monograph. c1840.
Speed, A. The gentleman's compleat jockey; with the perfect horseman, and experienc'd farrier. Monograph. 1697.

References

Argamon, S., Koppel, M., and Avneri, G. Routing documents according to style. *Proceedings of the First International Workshop on Innovative Internet Information Systems*, 1998; iii-98.
Baayen, R. H. *Analyzing linguistic data: A practical introduction to statistics using R*. CUP (Cambridge), 2008.
Baayen, R. H. `languager' (R manual). 2014.
Baron, A. and Rayson, P. VARD2: a tool for dealing with spelling variation in historical corpora. *Proceedings of the Postgraduate Conference in Corpus Linguistics*. 2008.
Bécue-Bertaut, M., Kostov, B., Morin, A., and Naro, G. Rhetorical strategy in forensic speeches: Multidimensional statistics-based methodology. *Journal of Classification* 2014; 31(1), 85.
Beh, E. J., & Lombardo, R. *Correspondence Analysis: Theory, practice and new strategies* (Wiley Series in Probability and Statistics). Wiley (Chichester), 2014.
Bendixen, M. A practical guide to the use of correspondence analysis in marketing research. *Marketing Research On-Line* 1996; 1(1): 16-36.
Benzécri, J. P. *Handbook of Correspondence Analysis* (Transl: T.K. Go-palan). Marcel Dekker (New York), 1992.
Biber, D. and Finegan, E. Drift and the evolution of English style: A history of three genres. *Language* 1989; 65(3): 487-517.
Biber, D. and Finegan, E. Diachronic relations among speech-based and written registers in English. *To explain the present: Studies in the changing English language in honour of Matti Rissanen*, Nevalainen, T. and Kahlas-Tarkka, L. (Eds). Mémoires de la Société Néophilologique de Helsinki. Société Néophilologique (Helsinki), 1997.
Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. *Longman Grammar of Spoken and Written English*. Longman (London), 1999.
Burnley, J. D. Curial prose in England. *Speculum* 1986; 61: 593-614.
Ernestus, M., van Mulken, M. and Baayen, R.H. Ridders en heiligen in tijd en ruimte: Moderne stylometrische technieken toegepast op Oud-Franse teksten. *Taal & Tongval* 2006; 58: 70-83.
Fish, S. E. *Is there a text in this class?: The authority of interpretive communities*. Harvard University Press (Cambridge, Mass.), 1980.
Gamon, M. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. *Proceedings of the 20th International Conference on Computational Linguistics* (Geneva), 2004.

Greenacre, M. *Theory and Application of Correspondence Analysis*. Academic Press (London), 1984.
Greenacre, M. Clustering the rows and columns of a contingency table. *Journal of Classication* 1988; 5(1): 39-51.
Greenacre, M. *Correspondence Analysis in Practice*. 3rd edition. Chapman & Hall (London), 2017.
Gries, S. T. Frequency tables, effect sizes, and explorations. *Corpus methods for semantics: quantitative studies in polysemy and synonymy*. Glynn, D. and Robinson, J. (Eds), Benjamins (Amsterdam/Philadelphia), 2014.
Gries, S. T., Newman, J., and Shaoul, C. N-grams and the clustering of registers. *ELR Journal* 2011; 5(1).
Halliday, M. A. K. The language of science. *Collected works of M.A.K. Halliday* Vol. 5, Webster, J. (Ed.). Continuum (London), 2004.
Hiltunen, T. and Tyrrkö, J. Normalising and tagging Early Modern English medical writing (1500 - 1700): A pilot study. *Presentation at finsse* (Joensuu, Finland), 2012.
Holmes, D. I. and Kardos, J. Who was the author? An introduction to stylometry. *Chance* 2003; 16(2): 5-8.
Hornik, K., Zeileis, A., Hothorn, T. and Buchta, C. RWeka: An R Interface to Weka. R package version 0.4-34. 2007. http://CRAN.R-project.org/.
Huddleston, R. D. *The sentence in written English: A syntactic study based on an analysis of scientic texts*. (Cambridge Studies in Linguistics). CUP (London), 1971.
Kornai, A. *Mathematical linguistics* (Advanced Information and Knowledge Processing). Springer (London), 2008.
Lenker, U. *Argument and rhetoric: Adverbial connectors in the history of English* (Topics in English Linguistics 64). Mouton de Gruyter (Berlin), 2010.
Los, B. and Dreschler, G. The loss of local anchoring: From adverbial local anchors to permissive subjects. *Rethinking Approaches to the History of English*, Nevalainen, T. and Traugott, E.C. (eds). Oxford University Press (New York), 2012.
Mair, C., Hundt, M., Leech, G., and Smith, N. Short term diachronic shifts in part-of-speech frequencies: A comparison of the tagged lob and f-lob corpora. *International Journal of Corpus Linguistics* 2002; 7(2): 245-264.
Murtagh, F. *Correspondence analysis and data coding with Java and R*. CRC Press. 2005.
Nenadić, O. and Greenacre, M. Correspondence analysis in R, with two- and three-dimensional graphics: the ca package. *Journal of Statistical Software* 2007; 20. https://goescholar.uni-goettingen.de/handle/1/5892. Retrieved 10 October 2017.
Nenadic, O. and Greenacre, M. J. `ca' (R manual). 2014.
Nunberg, G., Briscoe, T., and Huddleston, R. (2002). Punctuation. *The Cambridge Grammar of the English Language*, Huddleston, R. and Pullum, G.K. (Eds.). CUP (Cambridge), 2002.
Oksanen, J. *Cluster analysis: Tutorial with R.* 2014. Retrieved from http://cc.oulu.fi/~jarioksa/opetus/metodi/sessio3.pdf
Pahta, P. and Taavitsainen, I. An interdisciplinary approach to medical writing in Early Modern English. *Medical writing in Early Modern English* (Studies in English Language), Taavitsainen, I. & Pahta, P. (Eds). CUP (Cambridge), 2011.
Pérez-Guerra, J. Word order after the loss of the verb-second constraint or the importance of early Modern English in the fixation of syntactic and informative (un-)markedness. *English Studies* 2005; 86: 342–69.
R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. http://www.R-project.org.

Rayson, P., Archer, D., Baron, A., Culpeper, J., and Smith, N. Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. *Proceedings of Corpus Linguistics* 2007, July 27-30, University of Birmingham, UK. Retrieved from http://ucrel.lancs.ac.uk/publications/cl2007/paper/192 Paper.pdf

Rudman, J. Authorship attribution: statistical and computational methods. *Encyclopedia of Language & Linguistics* (Second Edition), Brown, K. (Ed.) Elsevier (Oxford), 2006.

Sinclair J. *Corpus, concordance, collocation*. Oxford University Press (Oxford), 1991.

Stubbs, M. Conrad in the computer: Examples of quantitative stylistic methods. *Language and Literature* 2005; 14(1): 5-24.

Swales, J. *Genre analysis: English in academic and research settings*. CUP (Cambridge), 1990.

Taavitsainen, I., and Pahta, P. (Eds). Medical and Scientific Writing in Late Medieval English. CUP (Cambridge), 2004.

Tummers, J., Speelman, D., and Geeraerts, D. Multiple Correspondence Analysis as heuristic tool to unveil confounding variables in corpus linguistics. *Proceedings of the 11th International Conference on the Statistical Analysis of Textual Data.* Presses Universitaires de Louvain (Liège), 2012.

Tummers, J., Speelman, D., and Geeraerts. Spurious effects in variational corpus linguistics: Identification and implications of confounding. *International Journal of Corpus Linguistics* 2014; 19(4): 478-504.

Tyrkkö, J. Exploring part-of-speech profiles and authorship attribution in early modern medical texts. Meaning in the History of English: Words and Texts in Context (Studies in Language Companion Series 148), Jucker, A.H., Landert, D., Seiler, A. and Studer-Joho, N. (Eds). Benjamins (Amsterdam), 2013.