THE UNIVERSITY
of EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Towards Profiles of Periodic Style

Discourse organisation in modern English instructional writing

Thijs Hendrikus Johannes Bernardus Lubbers

Thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Philosophy, Psychology and Language Sciences
The University of Edinburgh
2016

Abstract

A notorious challenge in the study of the diachrony of English is to determine whether developments in syntax, including changing frequencies of a particular construction, or word-order changes as suggested by perceived patterns in extant texts, represent genuine linguistic changes or are due to changes in conventions of writing. What is intuitively clear, however, even to a casual eye, is that a piece of English prose from, say, the 16^{th}-century differs markedly from texts from the 18^{th}-century. Yet such judgements cannot be based on syntactic changes alone, since essential grammatical features of Present-Day English are in place already by the end of the Late Middle English period. As a result, these differences are often simply ascribed to the notoriously elusive domain of style.

The current study attempts to come to grips with the issue of period-specific conventions of writing by focusing on features of discourse structure and textual organisation as of the Early Modern English period. It can be positioned at the meso-level between large-scale quantitative approaches of sentence-level linguistic features and detailed, small-scale discourse-analytic studies of individual texts. Texts selected for the current purpose, manuals for equine care, derive from a sub-domain of instructional writing with a long history in the vernacular. As these texts share similar communicative purposes and deal with the same "global" topics of feeding and looking after a horse, any differences between them cannot be attributed to different genres or differences in subject matter. This permits us to zoom in on 'agnates', different ways of expressing the same meanings, and allows us to see how the stylistic options selected by authors achieve the various communicative goals that have to be negotiated, such as discourse coherence or the transition to new topics.

The three main sections in this dissertation offer different ways to identifying developments in discourse organisation. The first section explores the traditional corpus-based approach that is frequently used to measure the parameter of "personal involvement", an indicator of periodic style. Initially, this approach restricts itself to measuring the contribution of frequencies of individual lexical items like first and second person pronouns. Next, this section will focus on the presence and linguistic realisation of the interlocutors of these instructional texts, i.e. the writer and the reader. The second main section will try to diagnose such varying styles by employing a completely data-driven, quantitative methodology which offers a linguistically unbiased and theory-independent perspective on the data in the corpus. This second approach offers cues as to how 'subliminal' patterns of grammar may affect perceptions of style, and how quantitative measures may aid in assessing whether the texts in our corpus cluster in expected or unexpected ways. The third section draws on theories of referential coherence and textual progression. By charting the variation with which texts from different periods in the history of English apply conventions for discourse organisation, it offers an insight into developments of hierarchical discourse structures (i.e., coordinated versus subordinated discourse relations) and practices of co-reference.

Taken together, these three independent measures offer a novel, multi-angled approach to stylistic developments in prose writing. Combining features 'above the sentence' level which involve discourse and information structural changes, this dissertation affords a glimpse into the emergence of written textual conventions, or 'grammars of prose', in the history of English.

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, either in whole or in part, in any previous application for a degree. Except where otherwise acknowledged, the work presented is entirely my own.

Thijs Hendrikus Johannes Bernardus Lubbers

Voor ma

Acknowledgements

The writing of this dissertation is associated with three different locations, and so, too, are the people I wish to thank for their support.

I am tremendously grateful to Bettelou for inviting me to come to Scotland when she accepted a chair at the University of Edinburgh. With her enthusiasm for research into the history of English, her keen insights and supportive pushes and nudges, she has contributed beyond measure to both my academic development as well as my personal growth. Likewise, I am grateful to Linda for looking over my shoulder when needed, as well as for her general kindness and helpful but unscrupulous eye regarding my ideas and writing. Graeme Trousdale and Ursula Lenker, my examiners, I thank for a stimulating *viva* and useful suggestions for the final version of this dissertation. There are far too many esteemed LEL staff members whose warmth, friendliness and intellect leave me astounded to this day. Nevertheless, special mention must go to Rhona and Meg for making me feel at home in the Dugald Stewart Building and Edinburgh from the start. I feel very fortunate to have enjoyed their company, as well as that of their partners-in-different-crimes, Davy and Roger.

I have been very fortunate to find myself part of the DSB room 1.15 community. Those that were there with me from the start, James, Mark, Matt, Misnadin, Kevin, Soundess and of course Steph, but also George, Daniel and so many others, including Michela in particular, I thank for their help, insights and friendship over the years. I am indebted to my good friend John-Sebastian for his chat and companionship during countless hours in CSE, in swimming pools, on trails and in woodlands. With their contagious enthusiasm for physical exercise, he and Hedwig have shown me parts of Edinburgh and Midlothian undoubtedly unknown to many locals. In addition to them, Judith also served as an important haven of 'Dutchness' in a sea of Scots (although we probably both hate to admit it!). Be it for spinning, nachos, home-made pizza or just a pint, dehydrating and re-hydrating in her company was always a great pleasure. Ever since a maiden traverse of Aonach Eagach, which will be etched in my memory forever, John has been an indefatigable friend. Ross I am indebted to for being a great flatmate and my trusted buddy all these years. So many fond memories since my very first days in Edinburgh, and but one complaint: ironing won't ever be the same!

The second stop in my journey to finishing this dissertation is Germany, where Tonjes Veenstra and Frau Beckmann facilitated my working at the library of the *Zentrum für Allgemeine Sprachwissenschaft* (ZAS). I wish to thank especially my father, Behzad, but also his wife Rajaa and daughter Guity, for their generous social and financial support during my stay in Berlin. I am particularly grateful for the fact that my PhD allowed me to spend time in their proximity.

My siblings Daria, Maren, Bidjan and Masiar, their respective partners and my nephews and nieces Jasmina, George, Sophie, Simon, Chris and Madeleine, I thank for their support, very welcome distractions and general family time. In addition, I wish to express my gratitude to their *omama*, Jutta, for her unexpected but enduring warmth and encouragement.

All members of the Shotokan Ohshima karate dojo Berlin, but in particular Claus, Charles, Jacky and Nina, I thank for their support and for allowing me to keep body and mind in sync – in the dojo as well as afterwards at the *Späti*. A special thanks goes to my sparring partner and good friend Daniel and his buddy Marc for their friendship and generous sharing of thoughts. Manfred I thank for showing me that age poses no barrier to remaining an engaging conversationalist. Lastly, my time in Berlin would not have been the same without the help of Alex and Felix. Providing me with a desk and innumerable nutritious lunches was the least of Felix's support, for which I will remain to be indebted.

The last stop in my dissertation journey, the Netherlands, and in particular the Twente region, hosts the people who have supported me throughout my life. Members of the Shotokan Ohshima karate dojos in the Netherlands, and in particular my friends Jeroen, Kees, Rob and Stefan, have supported me with advice and continued friendship. I am truly fortunate to have found such great people on my way. I thank my outdoor friends, *modderkruipers* Patrick, Paul, Rene and Stephan, for starting a yearly tradition by coming over to enjoy the Scottish Highlands with me. My friends Bas, Wouter, Luuk, Ivo and Arjan, whom I have known since secondary school, have each in their separate ways contributed to my development. Over the years, they have ensured my staying in touch with my roots. Special thanks goes to Luuk for bestowing on me the honour to act as his *paranimf*, and for making the effort to come visit me at each and every stop along the way. Like Luuk, my old uni buddy Frans has provided me with a shining example of how a PhD dissertation is defended. Who would have thought, back in the day! I should also wish to thank Erwin Komen and Ans van Kemenade at my old *alma mater*, the University of Nijmegen, for their moral and intellectual support at various points.

The tremendous support and encouragement I have received from my family should not go unmentioned. I am deeply grateful for the close-knit *Lubbers clan*, which is a support network without equal. In particular, I thank the *noabers* from next-door, Rita, Henk and MaVinLeonie for reality checks over countless cups of coffee, and for making the loneliness of a write-up more bearable. Dionne I thank for always remaining my little *sousin*, and for coming over to celebrate an unforgettable birthday with me on 12-12-'12. My godparents Annette and Edwin, and their respective partners Ben (*in memoriam*) and Jacqueline, deserve no less mention for looking out for me since I was a wee lad. In case they are watching down from up high, I will undoubtedly have also received support from my grandparents and uncle Jos.

A very special thank you goes to my mum, for her unrelenting support and for showing me, on a daily basis, the meaning of the word perseverance. I am more proud of her than I can ever express, and it fills me with great joy that I am able to dedicate this dissertation to her.

Lastly, I would like to thank Friederike for brightening up my life and for being a beacon of moral support for which I feel undeserving. Although my talent for understating how much I care for her has few equals, she has been my source of sunshine on many a *dreigh* PhD day.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

A relatively new direction for historical linguistic research which has gained considerable traction in recent years involves a focus on information structure and discourse structure. These two foci, essentially subfields of pragmatics which are sometimes also referred to as operating at the syntax-discourse interface (cf. Erteschik-Shir, 2007), are concerned not so much with syntactic structure or grammatical order but rather with the way in which informational content is managed. While information structure (IS) is mainly concerned with the ordering of such elements at the level of the sentence, e.g., what is new information, and how and where is it realised in the sentence, discourse structure (DS) seeks to study such phenomena at the level above the sentence, investigating the connection between information-structural units.[1] Although research on information structural tendencies in West-Germanic, and particularly English, is far from new (cf. Behagel, 1909; Chafe, 1976; Givn, 1983; Halliday, 1967; Mathesius, 1928/64; Prince, 1981), it is only fairly recently that these notions have attracted the serious attention of historical syntacticians. Recognising that such features may have had an impact on fundamental diachronic developments, recent contributions, e.g., Bech (2001), Los (2009), Taylor and Pintzuk (2012), van Kemenade and Los (2006), van Kemenade (2012), attempt to take such information structural features into account explanations of word order changes in the history of English.

In a collection of explorations on the topic of syntactic change relating to informa-

[1]Since these are relatively new fields of study, definitions and units of analysis are still under negotiation and subject to frequent re-definition. One such example is the appropriate level at which information structure should be analysed. Particularly in light of variation in practices of punctuation across particular periods and registers of English (cf. Parkes, 1992; Smith, 2012), it is not altogether clear whether sentence and clause level provide optimal structural unit of analysis for IS, resulting in the frequent use of the "utterance" as the primary structural unit (Los, Lpez-Couso, & Meurman-Solin, 2012, p. 5).

tion and discourse structure, Los et al. (2012) flag up a number of important challenges in the diachronic study of such features. One primary problem for research into information structure is the identification of patterns in writing as opposed to speech. Since information structure may be marked by prosody, intonation or even gestures in face-to-face spoken interaction, the lack of such communicative sings in writing necessitate written modes to develop other, functionally equivalent, routines. Beginning writers of English, for example, have to acquire routines to make up for the lack of intonation, an important device for signalling informational salience in English (cf. Birner & Ward, 1998; Erteschik-Shir, 2007; Lambrecht, 1994; Prince, 1981). That it is relevant to look into the different ways of expressing information structure in these two modalities of language is also underscored by the statement by Weinert and Miller that "[t]he terms 'spoken language' and 'written language' do not refer merely to different mediums but relate to partially different systems of morphology, syntax, vocabulary, and the organization of texts" (Weinert & Miller, 1998, p. 4-5).

With respect to the history of English, Los et al. observe that many phenomena observed in older texts can be traced back to "characteristics of preliterate oral text" (Los et al., 2012, p. 10). However, it seems worth noting that it took some time before the more puzzling aspects of pre-Modern literature, e.g., its extraordinary parataxis, mystery particles (cf. Brinton, 1996), conspicuous anaphora and repetitions, "proleptic" topicalizations, and jarring alternations of tenses (Fleischman, 1990, p. 10), were suddenly revealed as extremely interesting in their own right. This re-evaluation is probably in no small part due to the fact that written styles have become so important and pervasive in the Modern period that investigations into oral speech styles somewhat astonishingly demonstrated the extent to which these medieval textual phenomena, which are so bewildering for a modern linguist, derive from and can be explained by characteristics of a pre-literate oral culture. "Studies of oral versus literate strategies seem to suggest that in written literate traditions 'the meaning is in the text', in the actual written words, while in oral situations 'the meaning is in the context', and in the implications of communicative acts (Fleischman, 1990b:9, quoting Goody & Watt, 1968; see also Olson 1977, Bauman 1986)." (Los et al., 2012, p. 10). That this is not only true for historical periods is underscored by Tannen (1982), who illustrates how oral and literate discourse strategies may be identified in contemporary narrative. Such insights underscore the importance of studying the development from an oral into a written literary tradition from a linguistic perspective. In addition, they may afford a glimpse into the emergence of conventions of prose writing, i.e., a "grammar of prose" (Perret, 1988). E.g., "[w]hen speakers become authors and hearers readers, such conventions compensate for the loss of prosody and intonation to achieve communicative purposes" (Los et al., 2012, p. 10). It is thus assumed that these conventions readily

involve features pertaining to the structuring of discourse and information.

A second important challenge in the historical study of IS and DS which will only receive passing mention here is that of quantifying information structural change (cf. Los et al., 2012, p. 12). Since synchronic studies have shown that information structural patterns represent "options rather than absolutes" (Los et al., 2012, p. 12), low frequency attestations may reflect more fundamental causes than may be expected on the basis of their rates of occurrence. Added to the fact that native speaker judgements or psycholinguistic experiments are not available for the study of historical records, the researcher has to tread carefully when analysing such patterns. Interpreting information or discourse structural features often requires an analysis of alternative realisations, or information-packaging alternatives (cf. Ward, Birner, & Huddleston, 2002), making it a relative rather than absolute study of variation (cf. Biber & Finegan, 1989, p. 488).

Another challenge for investigating IS and DS phenomena in the mapping of spoken routines onto the written mode is that it can be misleading to take patterns which appear to be motivated by information or discourse structure as evidence for diachronic change, whereas in reality these patterns reflect stylistic (aesthetic) variation (Los et al., 2012, p. 11). Although it may be difficult to disentangle the motivation for such patterns, i.e., either stylistic or communicative/functional, both may appear as part of the same set of conventions that characterise a particular period or certain register.[2] It nevertheless indicates that the association between procedural functions and perceptions of style are intimately connected, and together may reflect characteristic features of period prose styles in the history of English. One possible way to progress from this situation, as Los et al. (2012) suggest, is to rely on synchronic studies of information-packaging constructions to identify the likely communicative purpose of a deviation from the regular – be it in terms of canonical word order at the level of the sentence or in the discourse structure above it. Another avenue is to rely on evidence that explicitly identifies certain textual conventions and their stylistic motivation for the organisation of discourse, for which particularly literary stylistic studies with a diachronic focus may be exploited (cf. Adamson, 2000).[3]

[2] From a certain point of view, aesthetic or stylistic motivations may also be termed functional in the sense that they may for example mark a certain piece of writing as belonging to a particular genre or text domain. We reserve the use of the label functional to refer to a particular communicative or procedural function here, however, and will not use it to refer to such linguistic marking related to the wider socio-cultural context.

[3] With respect to the use of the terms discourse structure and discourse organisation, we remark that we use the latter in a more general sense to refer to the way discourse is organised in text or speech, and the former as a notion for an as yet ill-defined, formal theory or set of theories that may account for structural patterns in such an organisation of discourse.

1.1.1 Approaching periodic styles in Modern English prose

In the introduction of a computational method for measuring differences between (historical) text samples, Juola and Baayen remark that,

> "A person completely familiar with 20th century English may still find Shakespeare somewhat daunting, an effect of three centuries of language drift, but will be more comfortable than a German speaker with no English knowledge whatsoever."
>
> (Juola & Baayen, 2005, p. 61)

This assertion, which harks back to observation above that modern-day readers may be challenged by interpreting prose from an earlier era, neatly situates the issue with which the current dissertation is concerned, as well as introducing the notion of language drift. Language drift refers to a term associated with the work of Sapir (1921) and, particularly with respect to more recent explorations of historical developments in English written language variation, that of Biber and Finegan. Biber and Finegan (1989) define such drift as "a cumulative series of gradual linguistic developments in a consistent direction" (Biber & Finegan, 1989, p. 489). While the definition of drift by Sapir (1921) pertains to structural and unconscious developments,[4] Biber and Finegan (1989) make a theoretical distinction between conscious and unconscious motivations for such patterns of development and target functional variation and its underlying motivations. This leads them to conclude that, in lieu of Sapir, "patterns of stylistic drift are often consciously accessible (and consciously advanced or resisted) rather than progressing below the level of consciousness" (Biber & Finegan, 1989, p. 516). Although these developments need not necessarily be taken to be deterministic, i.e., developing in a specific direction for a particular purpose, they will be motivated by situational demands. Based on Biber (1988), Biber and Finegan amply argue that "co-occurrence [of linguistic features] reflects shared [communicative] function" (Biber & Finegan, 1989, p. 488) and that directions of development are therefore to a considerable degree determined by the functional demands of circumstances in the communicative situation. One such direction is the particular linguistic trajectory observed in the development of specialist, professional registers such as scientific or medical prose, which will be mentioned below.

At the time of writing, however, Biber and Finegan (1989, p. 488) note that situational linguistic variation is not an object of study for most historical linguists, with the main body of research in this direction carried out in the literary oriented field of diachronic stylistics. The authors mention work by Gordon (1966) and Adolph (1968),

[4] "The drift of a language is constituted by the unconscious selection on the part of its speakers of those individual variations that are cumulative in some special direction" (Sapir, 1921, p. 155).

among others, who have attempted to "[describe] the linguistic and rhetorical characteristics of 'period styles' in English" from a literary angle (Biber & Finegan, 1989, p. 488).

Gordon (1966) devotes particular attention to the 15^{th}-century, noting both the pivotal role of this period for the development of a particular English prose style, as well as the popularity of manuals of instruction (see for a similar observation Taavitsainen, 1999). The 15^{th}-century is characterised as an "age of instruction", with "the demand in the fifteenth century for manuals of instruction [being] insatiable" (Gordon, 1966, p. 63). In general, English prose is described primarily as a utilitarian language, with this function and style going back to writings in Old English. Particular fond the author seems to be of the compact Alfredian prose, described as an effective functional prose of exposition, which itself harks back to Anglo-Saxon instructional writing (cf. Gordon, 1966, pp. 35-40, 58-64): "There was no literary pretension about this kind of writing [i.e., secular instructional manuals]. Prose was there to be useful" (Gordon, 1966, p. 59). Gordon proposes that this style of prose is arguably closer to the everyday language than the 'trailing' prose that is derived from translating French works lexical item-by-lexical item, for example extensively studied by Burnley (1986), resulting in a "Latinised Anglo-Saxon" rather than true English (Gordon, 1966, p. 39).

With respect to major developments in the trajectory of literary prose, however, both Gordon (1966) as well as Adolph (1968) perceive the most fundamental shift to take place around the second half of the 17^{th}-century, when the impact of a movement towards 'plain prose' is being felt (see especially Adolph, 1968, who devotes considerable attention to the 17^{th}-century "plain style"). Traced back to roughly the third quarter of the 17^{th}-century, Gordon identifies "a remarkable agreement between writer and reader – whatever their social class – of what constituted an acceptable way of writing" prose with lasting effects, which forms the basis of what he sees as the "century of prose" (i.e., the period 1660-1760; cf. Gordon, 1966, p. 133ff).

Adamson (2000) continues the tradition of literary stylistic analysis of developments in Early Modern English in her extensive chapter in the *Cambridge History of the English Language III: 1476-1776*, and also analyses these developments which take place around the second-half of the 17^{th}-century.[5] Whereas at the start of the Renaissance

[5]However, note that given her focus on the literary canon, the chapter by Adamson (2000) is primarily concerned with *planned or deliberate* stylistic change, and not necessarily linguistic change, in the written mode. Adamson defends this choice on the ground that many other prose writings may be argued to derive their styles from such literary examples (Adamson, 2000, p. 540). Given the author's primary focus on literary stylistics, however, it is not entirely clear whether she sees this exemplary role as pertaining only to literary writings, or also to non-literary, utilitarian prose. With Adamson remarking that "just as the distinction between literary and technical genres is not clear-cut in the Renaissance [...] nor is the dividing line between stylistic and utilitarian borrowings" (Adamson, 2000, p. 573), we assume that this trickling down effect of 28/12/16 20/02/17 20/02/17 28/12/16literary prose may in principle extend to any text domain of Early Modern English, including the current selection

the classical rhetorical distinction between high, middle and low or plain styles is still being emulated and exploited in English (cf. Adamson, 2000; Smith, 2006), this gives way to a more singular, unified pursuit in a "redirection of rhetoric" at the end of the 17^{th}-century (Adamson, 2000, p. 599). The author notes a late 17^{th}-, early 18^{th}-century turn towards different ideals of classical antiquity: while *copia*, an abundant, enriched style which could employ a variety of formal figures of speech is the guiding principle for rhetorical composition in the Renaissance (cf. Adamson, 2000, p. 545ff), the start of the neo-classical period 1660-1776 is guided by the concept of *perspicuity*. In addition, this shift reflected a move from a courtly to a middle-class audience, resulting in a "democratic plain style" (Adamson, 2000, p. 598).

Two important aspects or ideals which form the basis of the movement for perspicuity, according to Adamson, i.e., mutual intelligibility and referential transparency, are aimed at creating "a language of transparent reference" (Adamson, 2000, p. 600). Although not couched in exactly the same terms, Adamson (2000) analyses this new stylistic mode particularly in terms of changes in phenomena relating to phenomena of information and discourse structure. For example, a feature of renaissance writing that receives ample criticism in the later part of the 17^{th}-century, particularly with respect to too direct translations of Latin texts, is that of the "perverse distortion of natural word order" (Adamson, 2000, p. 603). 'Natural' was increasingly becoming regarded as the basic word order of *English*, and no longer that of classical models: "There was a general preference for maintaining an SVO sequence and for placing adjective before noun, verb before adverb and main clause before subordinate adverbial clause" (Adamson, 2000, p. 603). Interestingly, Adamson comments on the fact that although there was no technical term for what is now regarded as the domain of information structure, the concept is discussed by neo-classical grammarians. "They recognise that word order often performs the function of distributing the writer's emphases and enabling the reader to discriminate between given and new information" (Adamson, 2000, p. 604). A note on end-focus by Priestly illustrates this point,

> Priestley's advice reflects an understanding that natural stress and focus fall at the end of an information unit, which means that there are times when "it favours perspicuity" for the adverbial clause to precede the main clause [...]: 'for were those circumstances placed after the principal idea, they would either have no attention at all paid to them, or they would take from that which is due to the principal idea' (Priestley 1777: 282)
>
> (Adamson, 2000, p. 604)

On the other hand, the discourse structural features mentioned by Adamson can be roughly divided into what is termed referential and relational coherence in recent

of utilitarian prose.

linguistic and psycholinguistic theories of discourse (Sanders & Maat, 2006, p. 592). Referential coherence refers to the way that texts may cohere through units which relate to entities or referents in a mental or discourse representation model of discourse (e.g., nominal expressions and discourse anaphora), whereas relational coherence refers to the coherence established by the way in which text segments (e.g., sentences, utterances or clauses) are connected, and the specific relations between such text segments (i.e., clausal relations; cf. Sanders & Maat, 2006, p. 592). Both forms of coherence are moreover intimately connected with discourse and information structural phenomena. That these features of discourse coherence might prove particularly influential for perceptions of periodic styles is underscored by the observation by Adamson (2000) regarding the marked stylistic shift in the evolution of Early Modern English prose:

> "It is perhaps more than anything the new attention paid to connective strategies that causes the sea-change in prose which everyone notices in passing from renaissance to neo-classical styles."
>
> (pp. 604-5 Adamson, 2000)

In terms of relational coherence, this can be directly related to observations in a study by Lenker (2010) on adverbial connectors in the history of English and changes advocated by a group known as the Scottish Rhetoricians, which included Locke (see also Lenker, 2010, pp. 233ff, especially pp. 243-6, which is in turn partly based on the chapter by Adamson, 2000). A comparison of the frequency of adverbial connectors with an additive function (e.g., *and, also*) in various genres reveals that such connectors were particularly frequent in academic writing, probably because there was a deliberate effort to make links between connects as explicit as possible, in line with the ideal of transparency (see also Adamson, 2000, p. 607). Lenker shows how these developments led to syntactic change, with adverbial connectors and logical linkers shifting from clause-initial to clause-medial position (Lenker, 2010). However, the author also notes that certain changes in adverbial connection occurring after the Early Modern English period reflect stylistic rather than typological concerns (Lenker, 2010, p. 226).[6]

While the use of the term "connective strategies" in the previous quote is particularly associated the use of clausal connection to foster rhetorical coherence by way of conjunctions and conjunctive adverbials (cf. "adverbial connectors" in the terminology

[6]The conscious efforts of 18^{th}-century Scottish proponents of New Rhetoric and members of the *Royal Society* are mentioned by Lenker (2010, p. 233ff) as instrumental in the diachronic account of adverbial connection, explaining for example the divide between the use of adverbial connectors by Chaucer and Adam Smith. Without explicit evidence of prescriptive writings by the Scottish Rhetoricians on the use of adverbial connection, however, it might have been impossible to interpret certain developments as stylistic rather than typological. This underscores how important it is to bear in mind how explicit stylistic conventions can obscure language change.

by Lenker, 2010), it also relates to the use of strategies for referential coherence.[7]

Anaphora are discussed with specific reference to the resolution of ambiguity from the perspective of the receiver (i.e., the reader or audience). In Early Modern English prose, the pronoun serves to 'rehearse' the antecedent, and topic continuity is often also achieved by the use of synonymous noun phrases. Adamson (2000), however, notes that this "poses a double threat to the perspicuity of a text: readers have to establish sameness of sense in order to establish grammatical coreference; and they may have difficulties in interpreting the information structure of the message (in terms of its given/new relationships) since a new linguistic form may or may not signal a new topic" (Adamson, 2000, p. 605).

The anaphoric function of relativisers was also acknowledged by authors of the time, according to Adamson, with Swift remarking that "one of the greatest difficulties in our language, lies in the use of the relatives; and the making it always evident to what antecedent they refer (cited in Bately 1964: 282)" (Adamson, 2000, p. 605).[8] In addition, Adamson notes a prevalent use of restrictive relatives in the new plain style versus the much more frequent use of non-restrictive relatives in earlier periods Adamson (2000, p. 602). Relative clauses thus also gain an important referential function, with relatives in the new plain style being able to *define* or uniquely identify an antecedent, next to adding descriptive *elaboration* and digression (itself much more associated with the ideal of *copia*; Adamson, 2000, p. 602).In addition, *wh*-relativisers are favoured over *th*-forms, according to Adamson (2000), not so much due to the wish to follow a Latinate model but rather to independent considerations of perspicuity. Since *wh*-relativisers allow the inclusion of additional information that facilitates antecedent identification or disambiguation, e.g., in terms of the animacy of the antecedent (*who/which*), a pronoun's syntactic role in the clause (*who/whom*), and the fact that *that* can be confused with its use as a complementiser or a demonstrative, *wh*-relativisers are preferred over the use of *th*-forms (Adamson, 2000, p. 605).

The increased importance assigned to discourse coherence also underlies the prominence given to demonstratives and other discourse deictics, which may bind a discourse together in a similar way to anaphoric pronouns (Adamson, 2000, p. 606). Adamson (2000) briefly discusses an example drawn from Steele's (1711) essay *The Death of a Friend* in which every sentence contains a new subject. According to Adamson (2000),

[7] Burnley (1986, p. 596) has also identified certain linguistic features which promote referential coherence in the curial style of the late 15^{th}-century. Adamson (2000) associates the heavy use of (non-restrictive) relativisers, participial clauses and the use of anaphoric conjunctions which link clauses into larger units (although ambiguous between coordinating and subordinating function; cf. Adamson, 2000, pp. 583-6) with this curial style. However, the resulting 'trailing' style is often difficult to bring back to sentence units that would be acceptable in present day English prose, according to Adamson (2000, p. 583).

[8] Part of this confusion may stem from the distance between relativiser and antecedent, cf. Truswell (2011) on EModE relatives with a leftward island.

this practice runs the risk of making the discourse appear fragmented, which can be averted by a successful use of discourse deictics (Adamson, 2000, p. 606). At the risk of *Hineininterpretierung*, Adamson (2000) comments on the use of demonstratives in the passage by Steele by noting that they,

> "enhance cohesion by formally binding each sentence to its predecessor and they enhance comprehension by signalling that the new lexical material of the noun phrases they introduce is to be construed as given information [...]. In addition, they guide the reader through the topic-flow of the discourse, the distal deictic *that* marking the receding topic, the proximal deictic *these* marking the topic of continuing relevance or more immediate personal involvement."
>
> (Adamson, 2000, p. 606)

Such discourse deictics thus set up a "network of textual signposts", to which the use of the existential construction can also be added, according to Adamson (2000, p. 606). Below we will return to perceptions of (in)coherence in discourses of consecutive sentences in terms of the use of such new referential entities in subject position, noting here that this will be a key issue for matters discussed in chapter 6. Before we turn to grammatical strategies as identified by for example Halliday and Kastovsky for ensuring coherence in the face of such 'new subject'-referents, however, we draw attention to the quantitative linguistic studies of Biber and Finegan on historical developments in registers of English prose.

1.1.2 Style, densification and genres/registers in the history of English

Whereas the work by diachronic stylisticians such as Adamson (2000) is mainly focused on literary prose, Biber and Finegan and Taavitsainen have sought to approach the issue of changes in English genres and styles from a quantitative linguistic angle by applying statistical techniques such as factor analysis. Already mentioned above, Biber and Finegan (1989) approach the issue of situational linguistic variation in the history of English by analysing letters, essays and fictional writing from the mid-17^{th}-century up to the second half of the 20^{th}-century on a broad range of linguistic features. As the studies by Biber and Finegan underscore, it is necessary to keep the distinction between such different text domains in mind, since their trajectories of development may take different paths.

Although we will not go too deeply into the specifics of the factor analyses by Biber and Finegan (1989), it must be noted that these multidimensional studies are based on the empirical clustering of 67 linguistic features in a synchronic study of contemporary registers by Biber (1988), revealing a small set of underlying dimensions of linguistic variation. Biber and Finegan subject the historical texts in their corpus to a similar

analysis, and rank these according to the three dimensions in Biber (1988) that are associated with a contrast between roughly oral and literate styles: a dimension of Informational versus involved production (dimension A), a dimension of "Elaborated versus Situation-dependent reference" (dimension B) and a third dimension that is simply termed "Abstract style" (dimension C Biber & Finegan, 1989, p.).[9] Given this broad oral–literate distinction, the general results of these analyses show that English prose gradually developed towards a greater reliance on oral features in being less elaborate, less involved and less abstract over the time span covered by the corpus (cf. Sapir's language drift mentioned earlier Biber & Finegan, 1989).

With reference to their earlier results, however, Biber and Finegan (1992) add to the three genres in their previous study two additional genres with speech-like characteristics, i.e., dialogue in fiction and drama (i.e., plays Biber & Finegan, 1992, p. 689). The global pattern of development seems to be in line with the 1989 results, with 17^{th}- and 18^{th}-century texts starting out fairly literate, followed by a move towards more oral characteristics in the 19^{th}-century, and a continuing development towards distinct oral features in the 20^{th}-century. However, Biber and Finegan (1992, p. 695) also note "extreme experimentation" with both more oral and literate features in the middle periods of their corpus. In general, it is observed that the speech-like genres seem to be the most focussed (i.e., their range of internal variation is less extreme), and that the written genres of fiction and letters on the other hand show the most variation within each respective time period (Biber & Finegan, 1992, p. 699). For the speech-like genres, this lack of variation is most pronounced on Dimension C: Abstract Style, although an equivalent pattern can be detected in terms of Elaborated Reference (dimension B).

Comparing the "modern" section (i.e., texts written between 1865-1950) with texts taken from the contemporary *London-Lund* corpus (transcribed face-to-face conversation), Biber & Finegan show that the latter corpus "shows a wider range of variation than the literary dialogue texts within any given period on both of these dimensions" (1992: 699). One explanation for this could be that actual conversation shows "a wider range of purposes and situations" than that captured in literary dialogue, but the authors stress the genuine similarity between both literary and real-time dialogue (Biber & Finegan, 1992, pp. 699-700). Also, whereas 17^{th}-century dialogue in fiction seems to be fairly literate (information-focused) and shows a steady progression towards more

[9] These dimension labels are based on features which 'load' on the factors in the statistical solution. Features that are associated with positive weights on dimension A are for example nouns, attributive adjectives, prepositions and (longer) word length, with private verbs, 1^{st}- and 2^{nd}-person pronouns and *that*-deletion associated with the negative domain of dimension A. Dimension B has time and place adverbials and "other adverbs" in its negative domain, with nominalisations, phrasal coordination and *wh*-relative clauses in the positive domain. Dimension C only has only positive features with considerable weights, which contain agentless passives, conjuncts and *by*-passives, among others. The labelling based on these features is based on Biber (1988).

oral characteristics, dialogue in drama reflects the more general pattern of a fairly oral style becoming more literate over the 18^{th}-century before progressing towards a more oral style again (Biber & Finegan, 1992, p. 700). Unlike modern face to face conversation, a dialogue sample taken from a mid-18^{th}-century play shows extreme detail and information focus in its description of setting and events; e.g. through nouns, PPs, and an "extremely high frequency of attributive adjectives" (Biber & Finegan, 1992, p. 701). The elaborate and informationally adorned text seems to focus on providing vivid details in a planned fashion, something which is largely absent in real-time conversation (Biber & Finegan, 1992, p. 701; no doubt due to on-line processing constraints). These findings by Biber and Finegan (1992) thus situate the developments in written prose styles of English across various text domains, employing the use of certain linguistic features as a quantitative metric.

However, in a third study with the same aims but a completely different and extended corpus (i.e., ARCHER), Biber and Finegan note that the general patterns of drift in prose writing attested in their previous studies,

> "do not hold for all written registers: in particular, specialist, expository registers follow a different developmental course from popular written registers. It turns out that the written registers analysed in our earlier studies – essays, fiction and letters – were all popular kinds of writing produced for a wide, general readership."
>
> (Biber & Finegan, 1997/2001, p. 81)

This necessitates Biber and Finegan (1997/2001) to broadly discern two clusters of written registers, illustrating the markedly different trajectories of development which both groupings have undergone. It turns out that using this extended corpus, the popular, informal cluster of registers becomes considerably more literate and less speech-like over the course of the 18^{th}-century, before this direction of development is reversed in the 19^{th}- and 20^{th}-centuries as popular writing adopts more oral linguistic features again (Biber & Finegan, 1997/2001). The more formal registers, however, show a consistent drift towards less oral and more register-specific linguistic characteristics, and it is argued that "these specialist registers have come to exploit the resources of the written mode in innovative ways, resulting in styles of discourse not previously attested" (Biber & Finegan, 1997/2001, p. 82). As an additional consequence, the authors note that the audience for such specialist registers must have become increasingly specialised as well, "requiring extensive background training to be able to comprehend these texts effectively" (Biber & Finegan, 1997/2001, p. 82).[10] Based on these findings by Biber and Finegan, popular written registers such as diaries, private correspondence, news reporting and fictional writing may thus be differentiated on the basis of prevalent linguistic

[10]Such findings link up with observations by Halliday on the socio-cultural ramifications of the trajectory of scientific writing into a "discourse of the expert" (cf. Halliday, 2004, p. 95).

features and associated trajectories of drift from specialist and professional expository registers such medical, scientific, and legal prose (Biber & Finegan, 1997/2001, p. 82).

Taavitsainen has carried out similar multifactorial studies using Late Middle English and Early Modern English texts in the *Helsinki Corpus* (HC), focussing particularly on linguistic features that express interpersonal involvement as indicators of diverging text domains in the history of English prose writing (cf. Taavitsainen, 1993, 1997). One result appears to be, for example, that on the basis of specific uses of pronouns referring to reader and writer, one may successfully distinguish between fictional texts and texts from other, adjacent text domains (Taavitsainen, 1997, p. 257).[11] Whereas Taavitsainen (1997) focuses on involved features of dimension A, Biber and Clark (2002) and Biber (2003) choose to investigate historical developments such as noun phrase modification, which mainly involve features that are associated with the other, informational side of this first dimension identified by e.g., the use of nouns and adjectives, Biber (1988). For example, Biber, Conrad, and Cortes have observed that "[a]s a productive grammatical strategy, the dense use of complex noun phrase constructions has dramatically increased in informational written registers over the past 100 years" (Biber et al., 2004, p. 399-400).

Biber and Clark (2002) note that the most relevant diachronic developments regarding in specialist, expository registers pertains to dimensions A and C (Biber & Clark, 2002, p. 50). Registers such as scientific, medical and legal prose make use of increasing amounts of nouns, adjectives and prepositions (dimension A) and relative clause constructions (dimension C). Sampling four written registers along a continuum of formality (drama, fiction, news, medical prose), the authors note that a reliance on premodifying elements is particularly pronounced for the professional registers of news and medical prose, with medical prose leading the way for attributive adjectives and news reporting showing the highest frequency of noun+noun frequencies overall (see also Biber, 2003, who investigates these trends in news writing in particular). Although the frequency of such premodification strategies remains roughly the same over the course of the 17^{th}- to 20^{th}-centuries for the popular registers of drama and fiction, professional registers see a dramatic increase of such features, and particularly in the 19^{th}- and 20^{th}-centuries. With respect to the use of nominal postmodifying strategies, relatively little change is observed for prepositional phrases across these four registers, while prepositional phrases have become more frequent over the four centuries covered by the ARCHER corpus (Biber & Clark, 2002, p. 57; but see below for a caveat).

Such developments are interesting in light of the demands of lexical compactness and colloquialisation which may be used to characterise the two separate trajectories of

[11] These results, as well as other features of (inter)personal involvement and personal affect, will be more extensively covered in chapters 3 and 4.

the groupings of written registers identified by i.e., specialist and popular prose writing; Biber and Finegan (1997/2001). Biber and Gray (2012, p. 326) note that for fiction and specialist registers such as scientific writing, detecting the influence of such demands is fairly straightforward; with the former, popular register being subject to demands of colloquialisation and the latter, specialist register subject to an increasing "economy of writing" (Biber, 2003) which relies heavily on nominal features.Features associated with such a trend towards colloquialisation are for example the use of contractions and various verbal elements such as (semi-)modals and *get*-passives (cf. Biber, Johansson, Leech, Conrad, & Finegan, 1999; Biber & Gray, 2012). However, the register of news is marked as particularly interesting for combining the informational purpose associated with the professional registers (e.g., medical or scientific prose) whilst at the same time being written for a large, and generally popular, audience. Demands of lexical compactness and colloquialisation may be seen to compete in this register in particular, therefore. Since "newspaper prose has participated in only weak increases in the use of most colloquial features [whereas it has seen] strong increases in the use of most economy features" (Biber & Gray, 2012, pp. 326-327), the authors conclude that the stronger influence is seen to extend from the informational communicative purpose behind this drive for lexical compactness.

Leech, Smith, and Rayson, too, notice an increase in "nouniness" of written language over the course of the 20^{th}-century, suggesting that "nouns, as key parts of the noun phrase, have been taking a bigger role in written syntax." (Leech et al., 2012, p. 78). This trend is seen particularly in light of a choice in style, according to Leech et al. (2012). Although many definitions of style are available (e.g., Enkvist, 1973; Carter & Nash, 1990; Leech & Short, 2007; Sandig & Selting, 1997), the authors opt to use two everyday definitions of style, that is, as a form of "language variation within the standard language" which may be seen as either (1) "a particular way of using the language" or (2) "a particular way of expressing meanings" (Leech et al., 2012, p. 70). The authors concentrate mainly on this second definition, as it provides the more revealing results for corpus analytic approaches: because "there is an implicit comparison of varieties in any discussion of style" (Leech et al., 2012, p. 70), this second definition allows the comparison of the frequency for one stylistic alternative "in relation to another choice [whilst keeping] the meaning constant" (Leech et al., 2012, p. 73). "That is, using style in [this second sense], we can say that *I don't know* and *I do not know* differ 'merely in style', whereas, for example, *I don't know* and *I do know* differ more radically – in terms of meaning" (Leech et al., 2012, p. 73). Although measuring such "meaning-preserving" relative variation and change is methodologically decidedly more difficult, it should be linguistically more revealing (Leech et al., 2012, p. 73).

With respect to stylistic trends towards nounniness in written styles, Leech et al.

(2012) note that only specific elements associated with the noun phrase and noun phrase modification have increased in frequency, with prepositions generally on the decline (particularly *of*-PPs; cf. Leech et al., 2012, p. 78). In addition, Leech et al. (2012, p. 79) note that certain registers may be more permissive of particular strategies than others, with the lowest frequencies for noun+noun sequences and *s*-genitives found in fiction, and the highest frequencies not in learned prose but in news (cf. results by Biber & Clark, 2002). This trend towards nounniness, which is more appropriately termed "densification", seems primarily driven by the necessity to compress as much semantic content into as few words as possible (Leech et al., 2012, p. 78). Rather than using (postnominal) phrasal elements for noun phrase modification, Leech et al. (2012, p. 78) identify an increasing tendency for modification of the noun phrase head to be achieved by single-word premodifiers in English (see also Berg, 2011, on such patterns of pre-head modification). Under certain contextual circumstances, three different strategies for such head noun modification are presented by Leech et al. (2012, pp. 78-9) as *stylistic alternatives*, i.e., a postmodifying *of*-phrase (1-a), an *s*-genitive (1-b) or a noun+noun sequence (1-c):

(1) a. the behavior of a patient
b. a patient's behavior [Brown J34]
c. patient behavior
(example taken from Leech et al., 2012, p. 79)

Leech et al. (2012) seem to perceive these alternatives as equivalent rewordings, although (1-a) is taken to be the most explicit wording (and the most condensed, (1-c) arguably the most implicit). Although not referenced as such, the examples in (1) seems very closely linked to what Halliday has termed "agnates" (cf. Halliday, 1988), or a "paradigm of agnate forms" (Halliday, 2004, p. 29). Although the term agnation has a particular meaning in systemic-functional linguistics (cf. Matthiessen & Halliday, 2014, p. 49), we use it here in the general sense of a continuum of rewordings which all encode the same (truth-conditional) meaning. The notion of agnates is thus a useful concept, since it provides a perspective on grammatical encoding as a choice between equivalent stylistic alternatives.

1.1.3 Developments in specialist writing, discourse coherence and the grammatical metaphor/syntactic recategorisation

With specific reference to developments in a specialist register, Atkinson (2001) provides an account of how findings from an multidimensional (MD) analysis can be interpreted in qualitative terms, applying both an MD approach as well as a rhetorical

analysis on the Modern English *Philosophical Transactions of the Royal Society of London* in the period 1675-1975. In line with general trends discussed by Biber and Finegan (1997/2001), the author signals that an "extreme shift from comparatively involved/verbal discourse to highly informational/nominal discourse on Dimension [A] of the MD analysis closely parallels the movement from an author-centered to an object-centered discourse" (Atkinson, 2001, p. 61). This gradual movement in the Royal Society's *Philosopical Transactions*, and in scientific writing more generally, ties in with the evolution of a highly abstract and passivised discourse style. Although the MD results and rhetorical analyses show slightly different cut-off dates, the most marked shift in this trend takes place in the 19^{th}-century according to Atkinson's data, while even as late as at the end of the 19^{th}-century it is possible to observe a "mixture of emphases" relating to the centring of the discourse around either the author or the object of study. According to the author, such mixing is not remarkable even in the early scientific literature given the ultimate focus of scientific writing, which is to describe nature and natural phenomena: "no matter how personalized or author-centered, scientific texts attempt to tell their users ultimately about the object of study, and that object in its various aspects will thus frequently occupy a grammatically or information-structurally prominent place in sentences and clauses, prototypically the grammatical subject/theme position" (Atkinson, 2001, pp. 61-2). The shift to object-centred reporting lays the groundwork for some of the well-known present-day textual norms for scientific writing, according to Atkinson (2001), with "[t]exts grow ever more informational and non-narrative linguistically, and more impersonal and effaced rhetorically" (Atkinson, 2001, p. 63). In addition, although narrative was never a major feature of prose intended for the Royal Society's *Philosopical Transactions*, it is attested in the earlier periods but gradually loses even more terrain (Atkinson, 2001, p. 62). Perhaps even more revealing is that Atkinson points to historians who have noted a conscious agenda by members of the Royal Society to distance their writing in this new approach to natural philosophy and science from that of the (late) Scholastic tradition, which resulted in the development of a "rhetoric of immediate experience" (Atkinson, 2001, p. 62). The beginning of the 19^{th}-century saw a change from such Early Modern scientific attempts to the establishing of a scientific discourse proper, marked by a jump in the MD metrics that signals an increased informational production as well as being marked by general non-narrativity (Atkinson, 2001, p. 63).

Work on the evolution of the "language of science" by (Halliday, 2004) offers a useful functional interpretation of the diachronic developments of specialist registers noted by for example Biber and Finegan (1997/2001), Atkinson (2001). Halliday argues that the increased idiosyncrasy of the scientific register, as illustrated in (2), is an essential feature of such a contemporary scientific language:

(2) The rate of crack growth depends not only on the chemical environment but also on the magnitude of the applied stress (Michalske & Bunker, 1987, p. 81; as cited in Halliday, 2004, p. 141)

The 'Attic' style of (2) is contrasted with the 'Doric' style of earlier scientific writing (see below), and both styles are offset against everyday discourse (Halliday, 2004, p. 102). Where the Attic style of writing is based largely on packaging lexical information in (complex) nominal groups (cf. Biber & Clark, 2002; Leech et al., 2012), the Doric style is characterized by expressing its information in a more clausal form (Halliday, 2004, p. 103). In terms of meaning, moreover, these two forms, or 'agnates' in Halliday's terminology, constitute points along a continuum of formality, as the one form can usually be reworded into the other.

A typical Attic construction and, according to Halliday, "the hallmark [...] of scientific discourse in English" (Halliday, 2004, p. 104). is the development of the "grammatical metaphor". In general semantic terms, this is the combination of "a participating entity, a process, and then a second entity participating directly or circumstantially" (Halliday, 2004, p. 104). An example is given in (3):

(3) Investment in a rail facility implies a long-term commitment (Halliday, 2004, p. 105)

A more Doric version of such a sentence would be (4):

(4) If you invest in a new facility for the railways you will be committing [funds] for a long term (Halliday, 2004, p. 105)

Quite characteristic of the construction in (4) is the conception of processes (commonly: verbs) as entities (nouns), such as *investment* and *commitment* in (3). This, in turn, 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) to describe or further specify this relationship (cf. Halliday, 2004).[12]

With respect to diachrony, an investigation of the origins of the Attic mode reveals that early English scientific writing such as Chaucer's *Treatise on the Astrolabe* (c. 1390) actually shows little resemblance to what would later develop as the sophisticated scientific discourse style, Halliday (2004, p. 144) observes. This largely 'Doric' Middle English text does, however, contain specific technical nouns exclusive to the fields of astronomy and mathematics. A type of construction that persists in scientific texts to

[12]Note that Halliday (2004, pp. 39, 93) observes that grammatical metaphor is not only rewording, but can also entail "re-meaning" in a semogenic sense. Whereas rewording to Halliday is comparable to 'regrammaticising', the reconstruing of experience which takes place in the history of scientific writing by way of for example grammatical metaphor is also 'resemanticising' (Halliday, 2004, p. 95).

this day is the 'iterated-phrase-&-group Qualifier' construction (Halliday, 2004, p. 144). An example of this construction is given in (5):

(5) a change [in [transparency [of [the liquefied [substance]]]]]] (Opticks, Newton)

The text sample from Newton's work represents the next stage of development; the text is still largely Doric in nature, although Newton's theoretical conclusions contain the occasional Attic construction. It is not until Darwin's writing in the second half of the nineteenth century that we see the actual landslide development towards the Attic discourse style (Halliday, 2004, p. 114). Darwin's texts contain processes being expressed as nominalizations, often featuring the relatively new gerund, as for example (6):

(6) The security of the hive is known mainly to depend on a large number of bees being supported (The Origin of Species, Darwin, p. 255)

Halliday's qualitative analysis thus dovetails the quantitative findings of Biber and Finegan (1997/2001), Atkinson (2001), but also for example Leech et al. (2012), for a trend towards a more literate, idiosyncratic discourse style for specialised registers.

With respect to the concept of grammatical metaphor, however, it may be observed that both Halliday as well as Kastovsky draw attention to the use of comparable strategies for purposes of discourse cohesion and topic progression. The Hallidayan idea of a junction of grammatical classes is probably akin to what Kastovsky (2006) has termed *syntactic recategorisation* in the context of word-formation.[13] Although Kastovsky (2006) only speaks of recategorisation in the context of word formation, his syntactic recategorisation captures more than just strict nominalisation, and includes rewording in terms of syntactic/grammatical constructions in addition to phrasing "a complex lexical item [...] in nominal, adjectival or verbal form" (Kastovsky, 2006, p. 207): "word formation is also involved in another equally important function, viz. what might be called syntactic recategorisation, since it has syntactic properties. Here, a complex lexical item takes up the previous context and repeats it – almost like a pronoun – in nominal, adjectival or verbal form" (Kastovsky, 2006, pp. 206-207). Halliday (2004) mentions nominalisations as the first and most straightforward type of grammatical metaphor (e.g.,"decoupling 'qualities' and 'processes' from their congruent realizations of as adjectives and verbs, and recoupling both these meanings with nouns"), and Kastovsky (2006), too, regards such nominalisations as the most frequent type of recategorisation.

An example mentioned by Kastovsky (2006), reprinted here as (7), provides such a

[13] Note that with reference to grammatical metaphor, Halliday (2004, p. xvi) speaks of a junction between *category meanings*, and not just *word* meanings as involved in canonical (lexical) metaphor.

case of syntactic recategorisation, whereby the plural noun *bottles* is recategorised as an adjective in the following sentence (see Kastovsky, 2006, p. 207 for more examples):

(7) Finally he put the juice in bottles. Bottled juice was easier to store. (Kastovsky, 2006, p. 207; emphasis in original)

As a result of such a use of syntactic recategorisation, "the contents of a previous sentence can be easily modified without a clumsy syntactic construction" (Kastovsky, 2006, p. 207). In addition, the author claims that such word formations are frequently exploited for purposes of text cohesion, pronominalisation and information condensation (Kastovsky, 2006, p. 207).

Halliday (2004, p. 61) similarly notes how the grammatical metaphor may be employed to achieve textual coherence. In the example in (8), "the second sentence recapitulates the preceding point but in a grammatical form such that it can serve as the departure point for the next step in the reasoning" (Halliday, 2004, p. 61). This facilitates the setting up of chains of reasoning and logical progression of the discourse, which is one of the five central building blocks of the language of science and technology (Halliday, 2004, pp. 94-95).

(8) ... until *one essential nutrient in the medium* falls to a *very low value*, approaching exhaustion. At this limiting nutrient concentration, ... (Halliday, 2004, p. 61; emphasis mine)

In this way, both Kastovsky as well as Halliday independently indicate that these comparable strategies of syntactic or grammatical recategorisation may lead to ensuring discourse coherence in fairly typical examples of topic promotion (cf. Gregory & Michaelis, 2001; Lambrecht, 1994).

We may hark back to the discussion of an essay by Steele in Adamson (2000) above, where the use of new subjects is mentioned. It may be observed that the combined use of such word formation processes to create new referential entities may be another way to ensure the coherence of the discourse. This may be connected to the more significant function that Los (2012) associates with the entity encoding the subject after the loss of verb-second in English, with the subject role being increasingly associated with providing a link to the previous discourse, and in effect, the encoding of Given information (Los, 2012, p. 27).

But how, then, do the Attic and Doric styles in scientific writing and other formal prose relate to developments in informal registers?

Although it is asserted that depending on circumstance language users can alternate between different discourse styles at their disposal ('Doric' and 'Attic', but also 'Ionic',

'Arcadian', etc.), Halliday considers the clausal Doric style to be the source of the Attic style (Halliday, 2004, p. 123). This rests on the view that the Doric style itself remained largely unchanged, i.e., the basic template was clausal rather than nominal; an assumption which has not been tested exhaustively.[14]

This suggests that the Doric style itself is not a static entity but is subject to change. Informed by qualitative diachronic studies of informal registers, a study of linguistic features might reveal more details about the pathways of development in the 'Doric mode'.

Apart from the development of scientific expository writing, then, the history of English vernacular writing can be traced by investigating expository texts of a less specialist kind. An example of a text domain in this register that has an unbroken history from the 9^{th}- to the 14^{th}-century and beyond is the remedybook. These texts including commonplace books, medical treatises, and practical manuals, and started to be readily copied and disseminated across the country in what Taavitsainen has termed the "vernacular boom" in English writing in the 14^{th}-century (Taavitsainen, 2001b, p. 189). The same expansion is reflected in the nascent of new genres such as riddle books, instructive miscellanies, and texts on language teaching being added to the already established genres of religious and devotional prose. The register of instructional manuals is particularly interesting in that it can be considered to be both popular and informative – thus combining more informal linguistic form with a purpose, i.e. information transfer, that is often considered to be more detached in nature, and thus reflective of formal communicative pressures. The expectation, then, is that utilitarian manuals and proto-scientific texts must have been quite similar in this early period, both in purpose and linguistic form; practical instructional texts and Chaucer's *Treatise* may together have belonged to the Late Middle English text type equivalent of Biber's 20^{th}-century Information Interaction or Learned Exposition (Biber & Finegan, 1989). Studying the linguistic forms found in instructional manuals could 1) test the hypothesis that (informational) popular writing was subject to a different linguistic evolution than scientific writing, as Halliday has claimed, and 2) further delineate the dividing line between formal and informal register clusters as drawn by Biber and Finegan (1997/2001). An instructional genre that lends itself perfectly for this line of research is the equine manual – a prolific vernacular sub-register of instructional writing as popular today as it was in the later Middle Ages.

[14] In fact, this might not be exactly what is happening if we consider the data of Biber and Finegan (1997/2001): personal or non-specialist (Doric) writings becoming more formal (Attic) over the course of Modern English, before the pendulum swings back in the 20^{th}-century under pressure from campaigns for a plainer style (cf. Biber & Finegan, 1997/2001, and see also the *Conversationalization*-project at the Free University of Amsterdam).

1.2 Aims of this study

It may be clear from the above that scientific prose features quite prominently in the quantitative study of linguistic features of registers, genres and texts in the history of English (e.g., Biber, 2006; Halliday, 2004; Huddleston, 1971). The stylistics literature, on the other hand, has traditionally focused mainly on the study of aesthetically pleasing texts in poetry and prose (cf. Adamson, 2000; Adolph, 1968; Gordon, 1966). In between these two textual domains, however, there is a large field of popular writing left largely unstudied, especially from a diachronic point of view. This may partly be explained by the assumption that informal writing somehow has not changed significantly and merely forms the baseline from which other, more elevated forms of prose have developed and diverged independently. The study of colloquial writing is interesting in its own right, however, even if only to test whether this implicit 'baseline-hypothesis' is correct. In addition, and not in the least due to the availability of corpora such as *A Representative Corpus of Historical English Registers* (ARCHER), the *Corpus of English Dialogues 1560-1760* (CED) and the *Corpus of Early English Correspondence* (CEEC), there has been an increased attention in the literature on the history of English for colloquial forms of language.[15] Studies such as those by Fairman on 'naive writing' in the history of English serve to underscore that such colloquial, popular language is part of the diachronic linguistic landscape as well (cf. Fairman, 2000, 2002, 2003).

Authors such as Biber and Finegan have provided valuable insights in trends in linguistic variation that can be observed in (sub-)genres and registers in the history of English. There remains, however, a gap between the Historical Pragmatic approach and the Variationist approach regarding the question how linguistic form is linked to style and register in terms of the functions that a text has to perform. Diachronic variationist studies in the Finnish tradition, as well as research in the quite novel field of Historical Pragmatics, focus largely on lexical and morphosyntactic features and pay little attention to more syntactic and information structural units such as marked grammatical constructions with a specific procedural function (notable exceptions include Taavitsainen, 1997, 2004).

The primary aim of the current study is to establish how a specific English prose style may have developed over the course of the Early Modern period, and consecutively, whether we can profile various stages of this development successfully. It is assumed that informants' selection of available options in the store of marked word orders and grammatical constructions is not idiosyncratic but follows conventional practices. Using texts drawn from the same text domain and on the same subject matter should

[15] See for further information on these corpora the *Corpus Research Database* (CoRD) hosted at the University of Helsinki, http://www.helsinki.fi/varieng/CoRD/ .

allow us to develop a set of diagnostics (syntactic, discourse, and information structural indicators) that can identify such stylistic profiles. Informed by research on word order and information structural changes, the current dissertation may contextualise the diachronic evolution of (sub-)registers in terms of their procedural and textual functions. By charting stylistic developments in colloquial domains of informal and everyday writing such as popular lore, this study goes beyond an investigation of the oft-cited developments in registers of specialised prose or a study of literary sources.

1.3 Structure of this dissertation

Six chapters follow the current introductory chapter 1. Chapter 2 describes the derivation of data and makes some general observations regarding the corpus material on which the current dissertation is based. The core chapters of this dissertation are divided in three Parts, all exploring different aspects that contribute to perceptions of prose style: the expression of personal involvement (Part I), frequently co-occurring bundles or 'chunks' of grammatical features (Part II) and lastly the structuring of discourse and practices of referential coherence in Part III. Since these topics derive from different subfields of linguistics, they require different theoretical underpinnings and largely involve distinct methodological approaches. As a result, chapters 3 through 6 contain separate sections for reviewing relevant literature and outlining respective methodologies.

With respect to the particular chapters in this dissertation, chapters 3 and 4 in Part I of this dissertation generally follow the traditional variationist corpus linguistic approach for studying changes in English. Although these chapters are related and jointly contribute to the degree of personal involvement conveyed in a text, we have chosen to separate their contents based primarily on the methodological perspective taken. Chapter 3 traces the frequencies of 1^{st}- and 2^{nd}-person pronouns and providing an interpretation of their use in context, as well as their contribution to personal involvement in a form-to-function mapping (cf. Jacobs & Jucker, 1995). The second chapter in Part I takes the opposite approach, i.e., a function-to-form mapping. More specifically, chapter 4 exploits a discourse purpose, that of instructing a reader to carry out a desired act, and uses groupings of associated linguistic realisations to determine the degree of personal involvement conveyed. Together, these two chapters provide different points of view on the parameter of personal involvement, an oft-mentioned indicator of periodic style (cf. Taavitsainen, 1993).

In Part II, consisting of chapter 5, the identification of such varying styles is approached by employing a completely data-driven, quantitative methodology. This second approach offers cues as to how patterns of co-occurring grammatical markers

may affect perceptions of style, and how linguistically unbiased quantitative measures aid in assessing whether the texts in our corpus cluster in expected or unexpected ways. As an almost theory-independent method, this chapter stands in stark contrast to the single chapter which constitutes Part III, chapter 6. This chapter draws on theories of Information Structure and Discourse Structure, and combines a formal, computational theory of referential coherence with a text linguistic framework of discourse flow and textual progression. By charting the variation with which texts from different periods in the history of English apply conventions for discourse organisation, chapter 6 connects to insight into developments of hierarchical discourse structures (i.e., coordinated versus subordinated discourse relations) and practices of establishing co-reference.

The main findings of this dissertation are summarised in chapter 7.

Chapter 2

Material

2.1 Introduction

An analysis of the textual organisation and discourse structure of any selection of texts, as well as any diachronic changes it is subject to, will first need to identify and demarcate the body of texts under investigation. In addition, such a study should acknowledge confounding factors, such as wider linguistic or generic changes, which may have affected the development of this particular group of texts. The first section of this chapter will describe the derivation of data for the current dissertation, outlining how the text samples that comprise the current corpus were selected, as well as a general description of their source texts. The second section will deal with the distinction between text linguistic classifications such as genre, register and text type, and how our selection of texts can be characterised given these labels (§2.3).

2.2 Corpus – selection and description

2.2.1 Selection of source material

The current study is based on a selection of instructive writing obtained through the digital repository of early English printed publications, *Early English Books Online* (EEBO), freely accessible web repositories[1] and other sources in the public domain (see Bibliography part I: Primary sources). Texts selected for our corpus, manuals for equine care, derive from a sub-domain of instructional writing with a long history in the vernacular. The reason for choosing these particular texts is twofold. First, in light of connections to linguistic explorations of many adjacent (sub-)registers which share features with this particular type of texts, equine manuals seem an unexplored sub-register in terms of linguistic analysis. In terms of their dominant text type form,

[1] E.g., http://www.archive.org/ and the Google Books project, http://books.google.com/.

that of instruction, they may be related to diachronic studies of religious instruction by for example Kohnen (2001b, 2008a, 2008b), as well as historical explorations of secular instruction such as the cookery recipe (e.g., Grlach, 1992). However, as a linguistically more varied and elaborate genre than cookery recipes, equine manuals chart unexplored areas of secular instruction. In addition, they are less subject to the influence of Latin and the formulaic or codified language associated with religious instruction.

Demarcating the relationship between early medical and scientific writing and the selection of popular informative writing selected here deserves some attention. In terms of content matter, this connection pertains in particular to the veterinary information conveyed in such equine handbooks. To the extent that handbooks on equine care and more specialised registers such as scientific or medical writing share a communicative aim to *inform* its audience, there is a marked difference in that the latter are written for a specialist, professional audience, whereas the former are written primarily for a popular audience.[2] Although Biber and Finegan (1997/2001) have identified an increasing divergence in English prose writing of such specialist registers in terms of grammatical features, the grouping of texts represented in the general or non-specialist domain is largely lacking in secular prose types which share this informative discourse purpose. Equine manuals may thus inform whether this trend is primarily to be associated with the discourse purpose (to inform) of these specialist registers, or rather with the professional versus non-professional axis.

Second, although many other forms of secular instructional prose may be identified for English, few seem to have as long a history in the vernacular as does the horse manual. This is a favourable factor with respect to the availability of sources, and given the many available linguistic explorations of adjacent text domains, this makes the horse manual an excitingly unexplored area of secular prose.

Although the text samples are relatively limited in terms of sample size, i.e., n= 4,209 words for the shortest and 5,208 words for the longest sample, these fragments are highly consistent in terms of semantic content. The primary reason for this consistency is that these samples are collected to study the handling of comparable discourse topics, e.g., (the feeding of) hay, or (providing) water (to the horse), for the grammatical and co-referential encoding, particularly with respect to the realisation of information packaging constructions (cf. Ward et al., 2002). Although these texts do not present a parallel corpus in a strict sense, that is, they do not provide a body of texts for which parallel sentences can be aligned, the current method ensures a certain comparability of information flow not achievable by random corpus sampling. In addition, as these texts share similar communicative purposes and deal with the same global discourse topics

[2]Halliday (2004) has excellently decomposed the societal consequences of codifying meaning in such a way that it requires professional training to understand and reproduce such specialist registers.

of feeding and looking after a horse, any differences between them cannot be attributed to different genres or different subject matters using such a topic-alignment approach. This permits us to zoom in on 'agnates', different ways of expressing the same meanings in the grammar (cf. Halliday, 2004), allowing us to see how the stylistic options selected by authors achieve various communicative goals that have to be negotiated, such as discourse coherence or the transition to new topics.

Increasing the total word count turns out to be a highly laborious procedure using such an approach, as it involves incrementing the number of global discourse topics appearing across all texts in the current selection. Since these text samples are unified in terms of their general purpose, e,g,. identifying the vital elements for effective equine care, they should be fairly comparable as to global discourse topics covered. Nevertheless, there seems to be an inverse relation between the number of source texts selected and the number of similar discourse topics (that is, the more texts are selected, the fewer topics occur across all texts).

In addition to these criteria, a number of other practical matters determined which texts were selected for our corpus. We ensured to include at least two texts per century, from the 16^{th}-century onwards. Ideally, this meant selecting one source from the first half and one source from the second half of every century. The 16^{th}- and 21^{st}-centuries are exceptions to this heuristic, with both texts from the earlier century deriving from the second half of the century, and only one text selected for the current century.

A very practical concern was the availability of sources in the public domain, with preference given to texts with wide public accessibility (either via online repositories or public libraries). So as to avoid possible grammatical differences between British and American English (cf. Algeo, 2006; Rohdenburg & Schlter, 2009), as well as potential differences in British and American editorial practices, publications by UK-based publishing houses and by British authors were preferred, given equal availability. This was particularly relevant for publications in the 20^{th}-century. Gender, however, was not a strict criterion, with the latest two authors, Leighton-Hardman (1977) and Davies (2009), being female authors. A variable that is difficult to control for is that of audience, with texts in the early half of the corpus assumed to be written primarily by and for an adult, literate gentlemanly audience (cf. Curth, 2013; Keiser, 1999).[3] An attempt was made to keep this feature somewhat consistent by primarily excluding horse manuals for which it can be assumed that they are written for children or young adults. The combination of finding popular-informative publications intended for an adult audience, written by a UK-based author and on the right discourse topics

[3]It may be noted, however, that Molony and Warwick (2013, p. 356) recount that William Smellie, editor of the 1^{st} edition of the *Encyclopaedia Britannica*, as late as 1771 chose to treat the subject of farriery at considerable length "because most of the men engaged in this profession were universally illiterate".

proved to be somewhat of a challenge. Especially in the later half of the Late Modern English era, such texts rapidly become rather scientific. Marketing concerns may also impact the global discourse topics being dealt with in these texts. Publishing for an adult audience of horse owners, who might be either experts or complete novices with respect to general horse management (cf. Curth, 2013), means having to deal with both the mundane as well as the latest insights from veterinary specialists, agricultural biotechnologists, biochemists, environmental scientists, etc., so as to be able to cater for as large a potential audience as possible.

2.2.2 Description of source texts

Two of the sources used for the current corpus, the 19^{th}-century veterinary-oriented works by Skeavington (c1840) and Fleming (1884), can also be found in the Helsinki Corpus (HC). The third equine text available in the HC, a 15^{th}-century treatise known under various labels (e.g., CMHorses, MS Sloane 2584, *Horse Leechynge*, A Late Middle English Treatise on Horses; see particularly Sfinhufvud, 1978) is not included here, particularly for the reason that the text is a remedy book and therefore mainly deals with charms and medical recipes. As a consequence, it cannot be analysed straightforwardly as an instructional text that contains the discourse topics which form the basis for our current selection (see for example Milner, 2013, for the status of charms in this particular manuscript). The same reasoning underlies the exclusion of another Late Middle English text on horse care, MS Harley 6398 or *The Boke of Marchalsi* (cf. Odenstedt, 1973). Older vernacular sources which mention horse care in a veterinary context, e.g., Bald's *Leechbook* and *Leechdoms, wortcunning and starcraft of Early England* (cf. Cockayne, 1864), have not been taken into account for similar reasons. All of these texts have been consulted on occasion, however, for cross-referencing of source material.

Another text not officially included in our sample is the Present-Day English text by Duberstein and Johnson (2009/2012). Although in terms of audience and content matter it is very much akin to our sample texts, it is a four-page pamphlet and therefore in terms of length and textual properties not readily comparable to other texts in the current corpus. The pamphlet nature of this particular text may impact its style to an extent which cannot be assessed here, for example in terms of succinctness and "economy of writing" (cf. Biber, 2003). Although occasional examples may be drawn from this work, it is not included in the corpus proper for this particular reason.

The thirteen texts in our corpus are summarised in table 2.1. Full titles for these works can be found in the first section of the bibliography containing primary sources. This table lists the texts in order of date of publication, in addition to some basic descriptive statistics in terms of sample length, number of utterances and mean utterance length in words.

Table 2.1: Corpus – descriptive statistics

Date	Author	Sample size (words)	Utterances (n)	Mean utterance length (words)	Source
1565	Blundeville, Thomas	4,209	127	33.99	EEBO
1585	Clifford, Christopher	4,639	89	52.96	EEBO
1607	Markham, Gervais	4,439	100	45.31	EEBO
1618	Baret, Michael	4,386	102	43.92	EEBO
1697	Speed, A.	4,652	161	29.74	EEBO
1721	Gibson, William	5,028	195	26.70	EEBO
1796	Hunter, J.	4,480	130	35.28	EEBO
1823	Kirby, Jeremiah	4,457	200	23.58	online
c1840	Skeavington, George	4,507	187	24.87	EEBO
1886	Fleming, George	4,437	193	23.74	EEBO
1921	Matheson, Darley	4,599	193	24.57	online
1977	Leighton-Hardman, A.C.	4,588	217	21.60	publ. library
2009	Davies, Zoe	4,308	273	16.13	publ. library

As can be seen, the two texts from the 16^{th}-century derive from the second half of the century. This is due to a lack of a suitable source for the first half of this century. In addition, not two but three texts from the 17^{th}-century are included here (i.e., the texts by Markham, Baret and Speed). Although this overrepresentation of this century may appear as slightly worrying, it is offset by the inclusion of three sources for one century in the later half of the corpus (Kirby, Skeavington and Fleming). Although the distribution of these texts across the 19^{th}-century is somewhat more spread out than that for the three manuals from the 17^{th}-century (with two from the early half and one from the very last years of the century), it needs to be observed that no specific date of publication can be found for the text by Skeavington, in the literature nor in bibliographical records, which makes this texts somewhat difficult to position in terms of chronology. The Helsinki Corpus records that this text is published roughly in the fourth decade of the 19^{th}-century, which we follow here.

The texts by Hunter (1796) and Kirby (1823) are somewhat exceptional for their publication types. The manual by Hunter is written as a dictionary of farriery, and the text by Kirby appears in a 19^{th}-century edition of the *Encyclopaedia Britannica*. Nevertheless, these texts are included here because they seem to generally correspond to the other texts in table 2.1 in terms of subject matter and textual composition: Hunter has large enough entries to consider these topics on farriery as paragraphs or small chapters, while the text by Kirby is an entry which could easily have appeared in monograph form, being 155 pages in length. See particularly Molony and Warwick (2013) for the provenance, bibliographical details and publication history of this text by Kirby.

2.3 Classifying horse manuals

2.3.1 Text type, genre and register

Distinguishing between text classification labels such as *genre*, *register* and *text type* is often confusing. It has lead to idiosyncratic labelling (cf. Lee, 2001), both in synchronic as well as diachronic studies that attempt to take into account the textual environment to investigate and explain the use of linguistic features. In addition, Taavitsainen (1999) notes that the labels genre and text type are often used interchangeably (Taavitsainen, 1999, p. 140), and Lee (2001, p. 41) makes a similar comment regarding the use of these labels by Stubbs (1996).[4] Although Taavitsainen (1993) acknowledges the difference between the notions of genre and text type following Biber (1988), the author explicitly chooses to use these notions interchangeably in her study of domain-indicating features (Taavitsainen, 1993, p. 172). We will not attempt a full summary of the literature on genre classifications here, but rather opt to provide a clarification of the definitions of text classifications as used in this dissertation.

Pioneering work in distinguishing between the notions of *genre* and *text type* in the field of corpus linguistics has been carried out by Biber (1988, 1989). Biber (1988) distinguishes between these two classifications of texts on the notion that their internal structure is based on different criteria. Genres are seen as socio-cultural artefacts which are subject to situationally-defined, language- or culture-specific norms. Hence, the linguistic features contained within these texts are not regarded as criteria for the differentiation between members. As a result, genres are defined as "categorizations [of texts] assigned on the basis of external context" (Biber, 1988, p. 70). Text types, on the other hand, are characterised as "groupings of texts that are similar with respect to their linguistic form, irrespective of genre categories" (Biber, 1988, p. 70).

In an attempt to identify the basic text types of contemporary English, Biber (1989) notes that while genre distinctions are based on text-external criteria such as textual function (e.g., instruction) and audience (e.g., lay persons or specialists), the categorisation of text types is based on more linguistic, text-internal features such as the use of personal pronouns as a marker of informality. Following this line of reasoning, the deductive text typology by Biber (1989) is derived by statistical clustering of such grammatical features found in a broad range of English texts, and an assigning of labels for the eight clusters that appear in his statistical solution to the problem. This methodology can be compared to the five classes in the text typology by Werlich (1976). The latter has approached the issue of a classification of English text types in an inductive

[4] Note that two pages after her observation, Taavitsainen (1999) appears to get caught in the same trap somewhat by using the terms 'text type features' and 'genre purpose' with reference to the same text domain (i.e., recipes).

fashion by a careful scrutiny of (exemplary) texts. His typology has influenced the text type labels as provided in the Helsinki Corpus, for example (Taavitsainen, 1997).

The classes for both text typologies are provided in table 2.2. Note that since the text types are listed in order of presentation in their sources, the rows in this table do not represent the most closely resembling text types across both typologies.

Table 2.2: Text typologies by Biber (1989) and Werlich (1976)

Biber (1989)	Werlich (1976)
1. Intimate interpersonal interaction	1. Description
2. Informational interaction	2. Narration
3. Scientific exposition	3. Exposition
4. Learned exposition	4. Argumentation
5. Imaginative narrative	5. Instruction
6. General narrative exposition	
7. Situated reportage	
8. Involved persuasion	

Although texts within each text type can demonstrate a degree of linguistic variation, it may be important to note that in terms of linguistic features, the statistical variation between text types is always greater than the variation within text types in the categorisation of Biber (1989). In addition, while certain text types can dominate certain genres, they certainly do not have to correspond directly: depending on their orientation, some business letters will be sorted under involved persuasion, although the majority will correspond rather to features of informational interaction.

Particularly for text classifications that are based on linguistic features, such as the notion of text type used here, the identification of the smallest meaningful unit that may be used to differentiate between class members is an important issue. Although Lee (2001) seems to be searching for a characterisation of the smallest meaningful unit on which to base a classification using linguistic features (in his system, a 'register'), he follows Couture (1986) in regarding registers as abstract, text-internal linguistic patterns which are independent of text-level structures, whereas genres concern entire texts (Lee, 2001, p. 47). To define registers based on such sub-sentence level linguistic patterns, however, seems much closer to the notion of text type rather than register in frameworks that take a Biberian perspective, and has its own share of problems. Collocations or particular syntactic constructions may appear in certain registers with a high frequency, e.g., noun compounding or relativisation in scientific writing, and may even characterise it to some extent (but see Ariel, 2007, for a more nuanced perspective), they are ill suited to offer the basis for a comprehensive system of classification.

The question of how a text, text fragment or utterance is to be classified if it contains linguistic patterns of two different registers in exactly equal amounts seems a thorny problem for such an approach. The same problem has been identified by various authors, among others Longacre (1996, p. 15-6), especially on Embedding and Mixing Surface text types, e.g., Biber and Conrad (2009, p. 72-3) and Taavitsainen (2001a, p. 140 for an historical example).

We take every utterance to be associated with a specific linguistic form (typical sentence forms; cf. Werlich, 1976). This surface form usually corresponds to, and is associated with, a particular discourse purpose or communicative intent in a certain cultural context (cf. Longacre, 1996; Smith, 2003). Embedding or mimicking of surface forms does not obscure this intent, however, and may even strengthen it. As Biber and Conrad (2009) observe, register-use-out-of-situational-context may be employed to comical effect, or to lend emphasis to a particular surface form, if used within reason. In such cases, a "linguistic form is deliberately chosen because it is associated with a particular situational context other than the actual context of the interaction" (Biber & Conrad, 2009, p. 37). In the context of registers, this is termed *register shift*. In the context of texts and text categories, such a shift would be a genre or text type shift, and depending on whether a later shift returns to the original genre or text type, it would be a form of genre/text type embedding.

A practical problem with studying diachronic changes in text types, in addition, is the fact that this type of categorisation is defined by text-internal linguistic features, and as such cannot be studied by means of (changes in) those same defining criteria (the associated cultural usage of a text type such as learned exposition is merely a label, and derivative of a text type's linguistic features). For example, the use of personal pronouns might cluster with other text type features of intimate interpersonal interaction in a factor analysis of a contemporary corpus, but the same factor analysis on a historical corpus of English might reveal those pronouns to cluster with quite different features in a different time period (cf. Biber, 2001). To study changes in the use of personal pronouns in the text type of intimate interpersonal interaction diachronically, therefore, would mean using this feature both as a dependent and independent variable at the same time.

Genres, on the other hand, are based on a categorisation of text-external information in the context of culture, and it thus seems viable to study changes in linguistic features in a categorisation based on such criteria – even if cultural opinions might gradually change on what the prototypical macro-textual features of a genre should be (Taavitsainen, 1993, p. 171). The linguistic dimension of such generic shifts is readily approached using a corpus-based methodology. Kohnen (2001a), Kohnen (2010), Moessner (2001) and Diller (2001), as well as other contributions in Diller and Grlach

(2001), provide a particularly good starting point for the interpretation of differences between text classifications, and diachronic developments of such categorisations, in the history of English using corpus-based methods.

A distinction between genre and text type by Taavitsainen (1993) that has received considerable traction in historical variationist corpus linguistics generally follows the approach taken by Biber (1989), with genres and text types both being defined as different "abstractions made on the basis of individual texts" (Taavitsainen, 2001a, p. 140). Taavitsainen adds to this view a historical dimension, as well as the notion of genre as a 'horizon of expectation' for both consumers as well as producers, a term borrowed from Jauss (1979) and Burrow (1982). That is, when readers become writers, such expectations shape new products to fit the category. In this light, Taavitsainen (1993) also points to Kress (1988), who has underscored the conditioning influence of the audience. When a text is produced for a specific target audience, which will have particular expectations about generic forms or genre features, "the expectations of the audience guide the writer to reproduce the significant features of the genre he aims at (Kress 1988: 136-138)" (Taavitsainen, 1993).

We argue that such expectations should primarily be understood at the macro-textual level: what topics a genre may cover, how the discourse or text may evolve, etc. Studying linguistic and stylistic developments of procedural features in writing on the micro-textual level, then, means that we attempt to delimit the external criteria (i.e., what is said) as much as possible – within a (sub)genre, and possibly a particular topic within that sub-genre – and study how this same type of information is transmitted diachronically ('how it is phrased') using the procedural signs available to writers of a certain era (cf. the study of agnates in Halliday, 2004).

Another frequently used text-external categorisation of (groups of) texts is register, which is a broad term referring to situationally-defined varieties of language use, which may encompass a number of genres. Biber et al. (1999) refer to register as a language variety "relating to circumstance and purpose", which is slightly more informative than the description of Taavitsainen (2001a, p. 141) of 'situational language use'. Note, however, that Biber in his later work has remarked that his earlier publications have largely ignored the theoretical distinction between *genre* and *register*. That is, the former in his early research and the latter in his later work have been used "as a general cover term to refer to situationally-defined varieties described for their characteristic lexico-grammatical features" (Biber, Connor, & Upton, 2007, p. 9). We reserve here this definition for register, and note that registers may typically indeed contain one or more genres. Legal, medical or scientific writing are usually defined as registers, as is news writing in contemporary English (cf. Biber et al., 1999; Taavitsainen, 2001a), and each of these may contain uniquely identifiable genres or sub-registers in addition. On

the other hand, particular conventions as regarding associated macro- or micro-textual features we see as primarily applying to genres.

A fourth term that we will occasionally use is that of text domain. Adopted from computational linguistic studies, we use it here as somewhat of a theory-neutral term to refer to members of a text classification when it is not entirely clear what we are dealing with – whether it is a genre, text type, register, or something else for which no suitable label is available in the literature. Given that the use of these terms is so often confusing, we choose to rather indicate when it is unclear, or irrelevant for the point we are making, rather than add to the inconsistency in use of these terms.

To summarise this section, definitions of text classifications as used in this dissertation are as follows:

Text type grouping of language products (minimally utterances/sentences) based on text-internal features or linguistic forms.

Genre grouping of language products (texts) based on text-external, discourse or cultural factors.

Register specific language use for a particular social context, also 'situational language' use. Broader term, which may encompass more than one genre.

Text domain General term for a member of a text classification, either based on text-internal (e.g., text type) or text-external features (e.g., genre, register)

2.3.2 Positioning horse manuals

With respect to the current corpus, several qualifications can be made based on the text classifications outlined above. Given that a genre rather than a text type approach appears better suited to the study of linguistic change, we define the data set in the current corpus along lines of text-external rather than text-internal features.

Given the close relationship between texts in the current corpus and texts identified as belonging to the medical or scientific register, it may seem fitting to categorise the texts in the current corpus to either of these registers. Such a categorisation seems particularly tempting in the early Modern era, with the introduction of the genre of handbooks and developments in early-scientific writing (cf. Taavitsainen, 1999), which have a shared communicative aim to inform the reader – whether specialist or lay. Such a classification of horse manuals as medical or scientific writing seems inappropriate, however. Manuals on this particular topic could only marginally be classed as medical given the general lack of purely medical content, and this classification would be particularly unjustified given our selection of specific discourse topics, which largely

belong to the domain of general horse care. The label 'scientific' appears equally ill-fitting: although the latest texts in our corpus may employ features that overlap with grammatical patterns frequently attested in scientific prose, its communicative purpose, i.e., writing to instruct and inform a non-expert audience, seems to set these manuals apart from true scientific writing. In particular, this 'informative feature' which seems to connect both text domains appears largely to be one of providing description (of processes, concepts, etc.).

In addition to the communicative purpose of informing the reader, however, there is an important component in terms of audience. Given that the consumer of a handbook is envisioned to be a non-expert (that is, maybe experienced enough to deal with the topic under discussion, but not expert enough so as not to have to resort to the reading of a handbook), we identify this type of writing primarily as intended for a non-specialist, or rather, popular, audience. The label "popular lore" used by e.g., Biber (1985a, 1985b) seems particularly appropriate for the texts in our corpus. Although the term popular lore is used as part of a text type categorisation by Biber, the description of this and other text types in the text classifications in these early studies is more compatible with the definition of register in later work. In addition, Taavitsainen (1999) defines handbooks as a new genre which is introduced in the English vernacular in the 15^{th}-century. Along these lines, we define handbooks as a sub-register of popular lore, and may even suggest that it is the prototypical form (i.e., genre) associated with this register at least in the Early Modern period.

Next to this popular-informative feature, however, an intrinsic communicative purpose of handbooks is to advise or instruct the reader on particular actions to take. As a type of popular (and secular) informative writing on a particular topic, the general care for horses, the handbook fragments in the current corpus are thus primarily defined as a sub-register of popular lore with an instructive purpose, which is an important feature to set it apart from other members of the class.

A tertiary communicative purpose, in addition to providing advice or instruction and general descriptive information, may be a focus on providing entertainment (especially in certain kinds of Early Modern English handbooks; cf. Taavitsainen, 1999). That is, we are dealing with a form of prose writing which meets the situational communicative criteria of being popular, informative (descriptive), providing advice (and instruction) and with a possible minor aim of being entertaining.

With respect to the analysis of horse manuals as a community of practice (cf. Kopaczyk & Jucker, 2013), we note for example the fact that Gibson (1721) recounts an anecdote about the feeding practices of King Philip of Spain's entourage on a visit to England, naming Blundeville (1565) as the source of this information. A century later, the *Encyclopaedia Britannica* in turn recounts the same anecdote, introducing

it by stating that "it is said that when Philip of Spain was in this country", etc., but notably without mentioning a particular source (cf. Kirby, 1823). Although a superficial case, it underlines that writers of horse manuals find themselves in a tradition of textual transmission of knowledge on a particular subject, which may have readily involved reading and engaging with what had been written on the subject before. In this way, both content as well as sub-register specific conventions may have been implicitly transmitted.

Part I

Personal Involvement

Chapter 3

From Form to Function: Pronouns as Personal Involvement Markers

3.1 Introduction

In an overview of discourse analytic research with an historical focus, Brinton (2015, p. 224) refers to both micro-level (e.g., discourse markers, performative verbs) as well as macro-level linguistic features or phenomena (e.g., speech acts, discourse genres, politeness). In the context of function-to-form and form-to-function mappings, micro-level features are readily used for the study of macro-level abstractions such as discourse, genre or text type, by investigating the functional contexts in which these individual micro-level forms appear. Studies which take a form-to-function approach may focus for example on the global function expressed by such micro-level forms, e.g., their textual/procedural (structuring) function, or an interpersonal function in expressing personal involvement (e.g. Brinton, 2015, pp. 225-6).

With respect to the expression of the macro-level phenomenon of style, one of the main dimensions that seems associated with perceptions of periodic styles (cf. Biber & Finegan, 1989) is that between 'involved versus informative/detached production' or 'affective mood' in the multi-dimensional/multi-feature framework by Biber (1988). Micro-level linguistic features associated with the Involved/Detached axis in for example Biber and Finegan (1989, 1992) are private verbs, 1^{st}- and 2^{nd}-person pronouns, hedges, contractions and *that*-deletion, whereas features on the other end of the scale include nouns, prepositions, attributive adjectives and generally longer words (cf. Biber & Finegan, 1989, 1992). Consecutive studies that exploit this framework in diachrony are for example Biber and Finegan (1992, 1997/2001) and Taavitsainen (1993, 1997).

Although labels in the respective studies slightly vary for this dimension which is associated with the expression of interpersonal involvement or personal affect, results largely point to the same underlying dimension. For example, factor A in Biber and Finegan (1989) and Biber and Finegan (1997) is termed 'Informational versus Involved production', whereas when Taavitsainen (1993) subjects texts in the Helsinki Corpus to the same approach, she uses the term "Affective Mood" to refer to this comparable factor in her statistical analysis. This axis has positive loadings for 1^{st}- and 2^{nd}-person pronouns, exclamations and wh-questions, among others, and is therefore a close correlate of the diachronic Involved/Detached axis of Biber and Finegan (1992). In particular, Taavitsainen regards such markers of personal involvement as distinctive stylistic traits which may be used to demarcate the boundaries of (sub)genres of Late Middle English prose (Taavitsainen, 1993, p. 174).

With respect to the expression of such features of personal involvement and particular prose styles in the history of English scientific and medical writing, Taavitsainen (1999) investigates the use of written dialogue forms in LME and EModE medical prose. Three major stylistic forms may be distinguished, which correspond to the evolution of (pre-)scientific thought-styles: philosophical dialogues dating back to Greek philosophy, question-answer patterns characteristic of the medieval scholastic style of writing, and the tradition of mimetic writing, that is, of imitating a dialogue between fictional characters.[1] This latter prose style gains in popularity as of the 15^{th}-century, and is used particularly for purposes of instruction (potentially with a secondary purpose of entertainment; cf. Taavitsainen, 1999, pp. 245-246). Taavitsainen (1999) observes that deontic modality, enumeration and use of 2^{nd}-person pronouns are features typical of the scholastic thought-style, whereas direct addresses and imperatives are associated especially with the later tradition of mimetic dialogues. In addition, prescriptive phrases of the form "thou oughtest ...", "you must note ...", "These things may not be forgotten ..." (cf. Taavitsainen, 1999, p. 253) are said to be a relevant feature particularly of texts in the mimetic tradition.

Many of these features seem to be referred to with a broad cover term as metadiscursive comments, or simply metacomments (cf. Swales, 1990, p. 18). These apply when an author breaks into the flow of discourse "to direct the reader's attention" (Taavitsainen, 2001a, p. 146). Such comments refer to the evolving text rather than the subject matter, and seem to have primarily two discourse functions: an interpersonal and a textual one (Taavitsainen, 2000, p. 193).[2] Focusing particularly on these

[1] See also Alcorn (2011, p. 102), mimetic from *mimesis*, the telling of a story by "imitation of another persons' words", as opposed to *diegesis*, the recounting of a narrative.

[2] According to Domnguez-Rodrguez and Rodrguez-lvarez (2015), the identification of these particular functions in metacomments goes back to Halliday (1970, p. 88), and may be connected particularly to a systemic-functional perspective of grammar.

textual effects, Taavitsainen illustrates how one of the discourse functions of such micro-level metacomments is to provide structure or organisation at the macro-level, i.e., by establishing referential links through anaphoric or cataphoric reference and introducing topic-shifts that guide the reader through the discourse (Taavitsainen, 2001a, p. 146). Conversely, they "may be used to emphasise and isolate some points in the text for the readers' benefit, or they may modify the propositional contents by expressing a subjective point-of-view e.g., by hedging" (Taavitsainen, 2000, p. 193). This latter function illustrates part of their interpersonal function and, not unimportantly, such metadiscursive comments may be used to "anticipate criticism and prevent misunderstanding, and thus be regarded as face-saving speech-acts" (Taavitsainen, 2000, p. 193; reference is also made to Gläser (1995)). As the ultimate purpose of such metacomments is to ensure that the communication between discourse participants is successful, however, the textual and interpersonal function usually overlap and cannot be completely separated in practice (Taavitsainen, 2000, p. 193).[3]

Although such metacomments thus appear as important devices in establishing personal involvement, the exact identification of linguistic surface forms which are associated with their usage seems fragmentary. Given their communicative purpose as authorial interventions to ensure smooth communication, a broad range of forms may potentially be considered as belonging to this category. As a starting point, features that are associated with the Involved dimension of the multidimensional framework may be investigated. The study by Taavitsainen (2000) is such a pilot study which tries to identify micro-linguistic features in the MD framework which may be identified as expressing a metadiscursive comment. In this context, Taavitsainen (2001a) has noted that,

> "The ways of expressing metadiscursive comments are connected with the distinctions between the involved versus the detached way of writing as well as the monologic versus dialogic discourse forms, from explicit dialogue between the author and the addressee to implicit guidance."
>
> (Taavitsainen, 2001a, p. 146)

3.1.1 Linguistic forms of personal affect in instructional writing

In this section, we will explore a number of surface forms considered as expressing metadiscursive comments and ways to establish addressee guidance – and specifically with reference to explicit interpersonal involvement. The use of 1^{st}- and 2^{nd}-person pronouns may be highlighted in particular, since even if the use of such linguistic markers

[3]This observation dovetails the dual function that Biber et al. (2004) identify for many of the lexical bundles expressing personal stance and discourse organisation in classroom teaching and academic textbooks.

does not directly match the communicative purpose of a metadiscursive comment, it may still be true that such pronouns are a strong indicator of the potential expression of interpersonal involvement between discourse participants. With respect to historical text fragments in the scientific register, for example, Atkinson (2001) regards the use of 1^{st}-person pronouns in contributions to the *Philosophical Transactions of the Royal Society of London* as a strong indicator of author-centred discourse.

Using fictional texts in the Early Modern English part of the Helsinki Corpus, Taavitsainen (1997) studies the function of personal pronouns as indicators of personal affect, reader involvement and proximity. 1^{st}- and 2^{nd}-person pronouns are primarily related to a dimension of "self" – "other". In addition, it is noted that the function of 2^{nd}-person pronouns should be seen primarily as expressing an orientation towards participants, although these may be either participants in the discourse context (the reader) or characters in the textual world (the discourse co-text; cf. Taavitsainen, 1997, p. 245).[4]

Somewhat unexpectedly, however, Taavitsainen (1997) notes that authorial comments that are unambiguously directed at the reader are surprisingly rare in the corpus (Taavitsainen, 1997, p. 245-6). Lacking such overt signals of direct authorial address, reader involvement is assumed to be achieved more covertly, for example by repetition of moral themes or commentary in prefaces or titles (e.g., "Now, dear Reader", Taavitsainen, 1997, p. 246). It is also observed, however, that although fictional texts may have didactic aims, 'direct reader guidance' in fiction is of a different magnitude and quality than that found in handbooks and 17^{th}-century scientific texts (Taavitsainen, 1997, p. 246). In fictional conversation, the line between endophoric and exophoric address ("direct reader orientation") is exceptionally blurry (Taavitsainen, 1997, p. 246). Fludernik (1993) has also commented on the multiple layers which exist in the interpretation of the addressee of 2^{nd}-person 'you' in (fictional) narrative.

In an earlier multifactorial study on texts in the Helsinki Corpus, Taavitsainen (1993) compares a number of texts from different genres in the corpus on the basis of indicators of "affective mood" (e.g., use of pronouns, interjections, wh-questions, private verbs) following the method outlined by Biber, Conrad, and Reppen (1998), Biber and Finegan (1989). Although the main registers studied are not directly related to the current selection of secular instruction (i.e., religious treatises, drama and romances), Taavitsainen highlights, among others, a text which is not entirely unrelated to our current corpus, the veterinary manuscript known as *Horse Leechynge* (cf. Sfinhufvud,

[4] In connection, we adopt here the distinction between text participants (i.e., interlocutors in the discourse *co-text*; endophoric co-reference) and discourse participants (i.e., exophoric co-reference, which primarily includes interlocutors in the context of the discourse situation). Note that the latter is a wider category which may encompass participants in both discourse context as well as discourse co-text (Brown & Yule, 1983, for exo- vs endophoric co-reference).

1978, and our comments on this text in chapter 2). In terms of features which express personal affect, this manual is claimed to be low on the use of 1^{st}-person singular pronouns compared to other texts in the corpus, mid-range on the use of the 2^{nd}-person singular pronouns, low to mid-range on private verbs and low in comparison to other texts in the use of plural forms for the 1^{st}- and 2^{nd}-person.

Using a diachronic corpus of the neighbouring register of medical prose, Taavitsainen (2004, p. 89) notes that 1^{st}-person and 2^{nd}-person pronouns, proximal deictic elements (e.g., *here, now, this*) and present tense verbs produce involvement on the part of the receiver, and as a result, make a text appear more interesting. The difference between the use of such pronouns in mimetic versus scientific texts is also underscored: in the former they are mainly used in speech acts (thanking, apologising, etc.), whereas in the latter their function is mainly restricted to textual content and the internal organisation of the discourse (e.g., commenting on the unfolding text/discourse; see also Hyland (2002b) in this context).

With respect to the occurrence of personal pronouns in combination with other grammatical features, Biber and Conrad (2009, pp. 64-69) provide an interpretation of the frequent attestation of 1^{st}- and 2^{nd}-person personal pronouns in combination with mental and desire verbs (e.g., *thought, want, would like*, etc.) in the context of contemporary classroom settings. It is argued that phrases such as '*Today I want you to read*' or '*I thought that for tomorrow we might discuss*' serve both as directives as well as discourse organisers (Biber & Conrad, 2009, p. 66). At the same time that lexical bundles such as "take/have a look at" express a directive speech act, they may thus have an additional function at the level of discourse organisation, by introducing new discourse topics (Biber et al., 2004, p. 391). Expressions which involve a desire verb plus a pronoun that refers to an interlocutor are associated with the situational characteristic of personal stance as well as directive use (personal pronouns are further associated with situational characteristics such as interactivity, and directive and procedural situations share imperatives as a surface feature; cf. Biber & Conrad, 2009, p. 68). In fact, the authors claim that, somewhat to their surprise, in classroom settings "the expression of personal stance is equally important to conveying informational content, [...] and is often intertwined with conveying content or giving instructions" (Biber & Conrad, 2009, p. 66).

This issue is explored more in detail in Biber et al. (2004), in which the authors investigate frequently occurring lexical bundles (i.e., n-grams with size = 4 and unit = 'word (token)') in university-level classroom teaching and textbooks. In this study, the authors focus their analysis on providing a functional interpretation of such lexical chunks, and categorise the most frequent bundles in three primary (discourse) functions: stance expressions, discourse organisers and referential expressions. Although

personal pronouns are usually associated particularly with bundles that express interpersonal stance, they feature readily in bundles with the discourse function of discourse organisation, or in referential expressions.

Of particular interest is the section on attitudinal/modality stance bundles, which can usually be termed personal (that is, overtly attributable to the speaker or writer) and include a subcategory associated with the expression of obligations and directives. These obligation/directive bundles generally "differ from other personal bundles in that they have a second person pronoun (*you*) rather than first person pronoun as subject" (Biber et al., 2004, p. 390). However, attestations can also include bundles with a 1^{st}-person pronoun, in which case the pronoun occurs often in combination with a verb of desire "conveying the speaker's desire that the addressee carry out some action, and thus functioning as directives" (e.g., *You need to know how* versus *I want you to know how* Biber et al., 2004, p. 390). Although they generally "[direct] the listener to carry out actions that the speaker wants to have completed", both obligation/directive bundles with and without 1^{st}-person pronouns are regarded as *personal* expressions of stance (Biber et al., 2004, p. 390). However, it is also noted that the 'directive force' of expressions that include a 2^{nd}-person pronoun is variable, and may be relatively indirect (e.g., "you might want to have a look at" Biber et al., 2004, p. 390).

A subcategory of attitudinal/modality stance expressions, intention/prediction bundles, deserves mention for including expressions that are "overtly personal, expressing the speaker's own intention to perform some future action" (Biber et al., 2004, p. 391). Often, these express intentions of joint action, e.g., 'what we're going to do now', which may be classified as a case of 'workshop-we' 1^{st}-person plural pronoun usage (cf. §3.2.1).

With respect to other particular grammatical features which may express personal involvement in the history of English, Taavitsainen (1994) chooses to investigate the use of imperatives and passives next to personal pronouns as indicators of personal affect in an early study on emotive features in scientific writing in the Helsinki Corpus. Although the author observes that "[p]ersonal pronouns of the first and second person provide a key to the more personal level of communication" (Taavitsainen, 1994, p. 332), these provide only part of the picture. The marked absence of personal pronouns in certain texts leads Taavitsainen to also take other surface forms into account. While imperative forms are noted as an additional important way to express the interpersonal affect in the absence of personal pronouns, the frequent use of passives and other impersonal expressions appears as distinctive for texts that are low on emotive features (cf. Taavitsainen, 1994, p. 333).[5] As this study by Taavitsainen is mainly concerned with a qualitative linguistic description of text samples, no quantitative results are

[5]That is, 'impersonal' in a more everyday sense of 'not personally involved', and not in the sense of for example Allen (1995), Mitchell (1985).

provided. Nevertheless, the author concludes that the markers of interpersonality and impersonality noted here give the impression that they represent "two complementary sides of the semantic structure of early scientific texts" (Taavitsainen, 1994, p. 340).

Coincidentally, one of the texts included in our sample is by the same author as a scientific text studied by i.e., Blundeville's (1597) *A Briefe Description of the Tables*; Taavitsainen (1994). Taavitsainen notes that 1^{st}-person pronouns occur particularly with verbs of cognition and perception in this text, in addition to appearing regularly in text-internal references (e.g., "as I have said before"; cf. Taavitsainen, 1994, p. 337). 2^{nd}-Person pronouns, on the other hand, are employed by Blundeville to take the viewpoint of the reader into account in explanations of the scientific process. Such pronouns also frequently occur in combination with modals of prediction, obligation and necessity ("you may perceive/must do", etc.; cf. Taavitsainen, 1994, p. 337). Constructions with 2^{nd}-person pronouns and modal verbs of this kind, as well as the imperatives mentioned earlier, may be identified as directive speech acts, or directives, which will be dealt with at length in chapter 4.

In a collection of papers on instructive writing in the history of English, Tanskanen, Skaffari, and Peikola (2009) specifically list imperatives, directive speech acts and modals of obligation as linguistic devices that "involve in various ways the reader(s) or addressee(s) in the discourse as targets of instruction" (Tanskanen et al., 2009, p. 5). Listing directive speech acts next to imperatives and modals of obligation seems somewhat confusing, as the latter are micro-forms that may even occur as instantiations of the macro-form of directive speech acts (cf. Brinton, 2015, above). Nevertheless, such directive speech acts may be compared to the category of directives in academic writing identified by Hyland (2002a, 2002b), which are defined as utterances that impart upon the reader "an obligation [...] to do or not do something" (Hyland, 2002b, p. 216). Directives are said to "help to construct readers as learners and learning as a one-way transfer of knowledge from primary-knower to neophyte" (Hyland, 2002b, pp. 222-223). Based on this assertion, it may well be that their occurrence thus helps to underscore the information asymmetry between writer and reader. Even more revealing in terms of their relationship with the discourse purpose of *instruction* is that directives are defined by Hyland as "utterances which *instruct the reader* to perform an action or to see things in a way determined by the writer" (Hyland, 2002b, pp. 215-216; emphasis mine). As such, "[d]irectives [...] permit authorial intervention in a discourse" (Hyland, 2002b, p. 232).

As far as the use of directives in certain genres or disciplines of academic writing is concerned, Hyland observes that natural science papers "not only contained far more directives, but these were also more likely to function as a means of guiding readers through a procedure and to the conclusions of the writer. Both these frequencies and

more impositional functions are partly influenced by traditions of precision, tight space constraints, and highly formalized argument structures in the hard fields [of science]" (Hyland, 2002b, p. 236). We will turn to the use of directives at length in chapter 4, but already note here the seemingly close relationship between the expression of personal involvement and the use of directive utterances in the current text domain of instructive prose.

With respect to the use of the extrapositions as a metacomment, that is, an "impersonal metadiscursive comment" in the sense of Taavitsainen (2000), it is noted that these constructions serve first and foremost a textual function in organising discourse which may be comparable to modern day sub-headings (and could even be accompanied by paralinguistic signs such as red ink in manuscripts; Taavitsainen, 2000, p. 194). Such *it is to + verb*-constructions appear to be relatively common in scholastic texts and are likely to be translations of Latin formulae for textual organisation, e.g., *restat dicere*, in proto-scientific medical texts. According to Taavitsainen (citing Blake 1992a), such collocations "[reflect] the attempt to transfer features of the learned writing in the source language to the new register in the vernacular" (Taavitsainen, 2000, p. 193).

In terms of the historical trends in the use of such metacomments and the discourse functions they express, it is noted that the general development is from textual to interpersonal, with the first occurrences of interpersonal use with 1^{st}-person singular pronouns attested in the mid-16^{th}-century (although trends are mediated by text domain and discipline within the larger genre of medical writing; cf. Taavitsainen, 2000, pp. 196, 202-203).

Taavitsainen (2001b, p. 195) provides frequent phrases of the form "it is to wit ...', 'it is to note ...', etc., as examples of prescriptive phrases. However, Pahta (2001, p. 213) points out that although prescriptive phrases of the form 'it is to be known' or it is to be noted' are indicators of a scholastic thought-style, these are often direct translations of Latin source texts (cf. Pahta & Taavitsainen, 1995; Taavitsainen & Pahta, 1995). Such Latinate forms should therefore be distinguished from directive extrapositions such as *it is important that ...*.

In addition, prescriptive phrases and directive extraposition constructions seem to put different obligations on the reader. For the former, it often seems to request a specific cognitive action on the part of the receiver ('note this', 'remember this', etc.), whereas directives are more circumscribed in the action that is required of the reader ('it is important that ...' conveys 'This is important. Make sure it happens!'). With respect to such required actions, Hyland makes a useful distinction in directives found in Present-day English academic articles between physical acts ('the temperature should be set to', etc.), textual acts (e.g., 'see section 4.1'),and cognitive acts ('it should be borne in mind'; for further distinctions within these categories, cf. Hyland, 2002b,

p. 218).

Note that the use of third-person pronouns ('he', 'she', 'it', 'they') and their case forms is not part of our current analysis. Although these pronouns undoubtedly contribute to the structure of the discourse (e.g. Brown & Yule, 1983; Chafe, 1985), we delay their treatment until section 6.1, where the matter of endophoric (text-internal) reference in relation to discourse structure and coherence is addressed more extensively. As Taavitsainen has noted, "[i]n general terms, first and second person pronouns mark the presence of a narrator and addressee, while third person pronouns mark relatively inexact reference to persons outside the immediate context of interaction." (Taavitsainen, 1997, p. 199).[6]

The current chapter thus focusses specifically on the relationship between discourse interlocutors, that is, the reader and writer, and how their interaction is realised linguistically. In conjunction with the next chapter, which provides a function-to-form mapping of the linguistic realisation of the orientation towards the reader, the two chapters in Part I highlight different angles on the expression of personal involvement in instructive texts. First, however, this chapter will outline the frequency of pronominal forms, illustrated with examples and an interpretation of their primary function in context.

3.2 Personal pronouns, involvement and discourse structuring

3.2.1 Frequency of 1^{st}- (sg.&pl.) and 2^{nd}-person pronouns

The most obvious source of personal involvement and reader engagement between sender and receiver can be found in the use of first and second person pronouns. As is well attested in the literature, such markers are associated with interpersonal styles of writing, both in contemporary registers as well as in the history of English (e.g. Biber, 1988; Biber & Finegan, 1989; Taavitsainen, 1993, 1997).

As a first step to identifying the discourse-structuring properties associated with these surface forms, an inventory is made of references to discourse participants involved in the production (the writer) and reception (the reader) of the text. The pronouns provided in table 3.1 here include both personal and possessive forms, as well as case distinctions (1^{st} (sg.): 'I', 'me', 'my', 'mine'; 2^{nd}: 'you', 'your', 'yours', and the 1^{st}-person plural forms 'we', 'us', 'our' and 'ours'). Spelling variations were also taken into account (e.g., 'wee', 'oure'), as well as *th*-forms (e.g., 2^{nd}-person 'thou', 'thee', 'thy',

[6] That an interpretation of the use and co-referential status of third-person pronouns in Middle and Early Modern English instructive text might be less than straightforward is underscored in Grund (2011) for alchemical texts.

'thine'). The latter are found in one text sample only, the late sixteenth-century manual by Christopher Clifford (see section 3.2.1). In addition, the sample includes a number of reflexive pronouns: one 'thy self' in Clifford, two counts of 'your self' in Speed and one 'yourself' in Skeavington. These counts were added to the general frequencies for the 2^{nd}-person pronoun lexemes.[7]

Table 3.1: Frequency of 1^{st}- and 2^{nd}-person pronouns in the corpus

Source	1^{st} (sg.)	2^{nd}	1^{st} (pl.)	Sample total (%)
1565 Blundeville	17	15	16	48 (1.14%)
1585 Clifford	126	179*	1	306 (6.60%)
1607 Markham	21	165	7	193 (4.35%)
1618 Baret	27	55	0	82 (1.87%)
1697 Speed	10	74	1	85 (1.82%)
1721 Gibson	7	5	50	62 (1.23%)
1796 Hunter	2	33	1	36 (0.80%)
1823 Kirby	0	0	14	14 (0.31%)
1840 Skeavington	18	42	9	69 (1.53%)
1886 Fleming	0	0	4	4 (0.09%)
1921 Matheson	0	1	3	4 (0.09%)
1977 Leighton	0	13	0	13 (0.28%)
2009 Davies	0	0	0	0 (0%)
Total	226	554	107	889 (1.51%)

*Includes 'thou' (n=28), 'thee' (5), 'thy' (17) and 'thine' (3)

Although the figures in table 3.1 suggest a decrease in the use of these personal pronouns over time, it needs to be pointed out that this finding might be a sampling effect rather than the result of actual register change. No strong claims are made with respect to these figures, therefore, and it is stressed that they should be taken primarily as a descriptive metric of the texts or text profiles included in our sample.

What may be clear from table 3.1, however, is that the absolute frequency of 1^{st}- and 2^{nd}-person pronoun use is highest in the earliest texts. There seems to be a sharp drop in pronoun use after 1800. If we regard the text by Skeavington (c1840) as an outlier for the moment, texts written after the start of the 19^{th}-century seem to generally refrain from the use of such pronouns. The use of the 1^{st}-person pronoun is already marginal in the eighteenth-century and seems to fall away completely a century later (again: disregarding the text by Skeavington). Patterns in the use of 2^{nd}-person pronouns and 1^{st}-person plural pronouns are not as clear-cut: 2^{nd}-person pronouns seem to be used generally throughout the period studied here. However, they occur with considerable

[7] Given that reflexives are, in referential terms, intimately connected with their deictic referents, they were included with the personal and possessive pronoun forms. A procedure as applied by Sily, Nevalainen, and Siirtola (2011, p. 185; Appendix A2), to combine reflexives written as two words and regard them as a single pronoun lexeme, would have been possible here. However, this did not seem particularly relevant given the relatively low number of reflexives in our sample.

variation between texts written in approximately the same time period, which seems to point to idiosyncratic usage. Nevertheless, 2^{nd}-person pronoun lexemes are by far the most prevalent pronoun type used in the corpus, representing roughly 62.3% of the total number of 1^{st}- and 2^{nd}-person pronoun lexemes.

With respect to the use of the 1^{st}-person plural pronoun, it may be observed that although the attestation of this form is generally low compared to the other pronouns, its use does not seem to be subject to the same sharp decrease in frequency. Although a marginal category throughout, 'we' and other 1^{st}-person plural pronoun forms are attested with somewhat similar frequencies before as well as after the start of the 19^{th}-century.

Note that in addition, Taavitsainen (1999, p. 247) has observed that 1^{st}- and 2^{nd}-person pronouns in early English medical writing (pre-EModE/pre-16^{th}-century) often tend to be found in clusters rather than evenly distributed across a sample. There are some indicators that this is also the case for the current selection of texts; with multiple pronouns frequently being used in isolated sentences, or with a high concentration of pronouns in a few pockets of consecutive utterances per text. For example, 10 of the 13 2^{nd}-person pronouns in Leighton-Hardman (1977) occur in the space of 5 consecutive utterances, on a total of just over 200 utterances in this text sample. There has not been any systematic scrutiny to ascertain whether this pattern is only attested in our samples, and is thus the result of sampling bias, or whether it is reflective of patterns found throughout the texts from which these samples have been drawn. Nevertheless, that Taavitsainen (1999) observes a similar pattern is food for thought, and establishes a relationship between medical writing and popular lore text in the current selection.[8]

Figure 3.1 overleaf provides an overview of the use of 1^{st}- and 2^{nd}-person pronouns as a proportion of each individual sample (cf. numbers in the stacked bars). These figures are based on standardised scores, i.e., number of tokens per 1,000 words. For ease of inspection, table 3.2 provides the corresponding standardised scores in tabular form.

At first glance, the diagram in 3.1 seems to suggest a sharp drop in the frequency of overt discourse participant realisation over time. If we disregard the texts by Clifford and Markham in this bar chart, however, the average number of 1^{st} and 2^{nd}-person pronouns does not seem to exceed the mark of 20 items per 1,000 words. Compared with the distribution of pronouns in contemporary registers, such figures fall somewhere in the mid-range between the highly formal written register of academic prose and that of other written registers (e.g., fiction and news writing; cf. Biber et al., 1999, p. 333-

[8] It is not improbable that this feature is associated particularly with instructive-informative writing, and this seems to warrant further study. For example, one may wonder whether a difference exists in intuitive judgements regarding personal involvement of texts if pronouns occur in clusters rather than evenly distributed across a text.

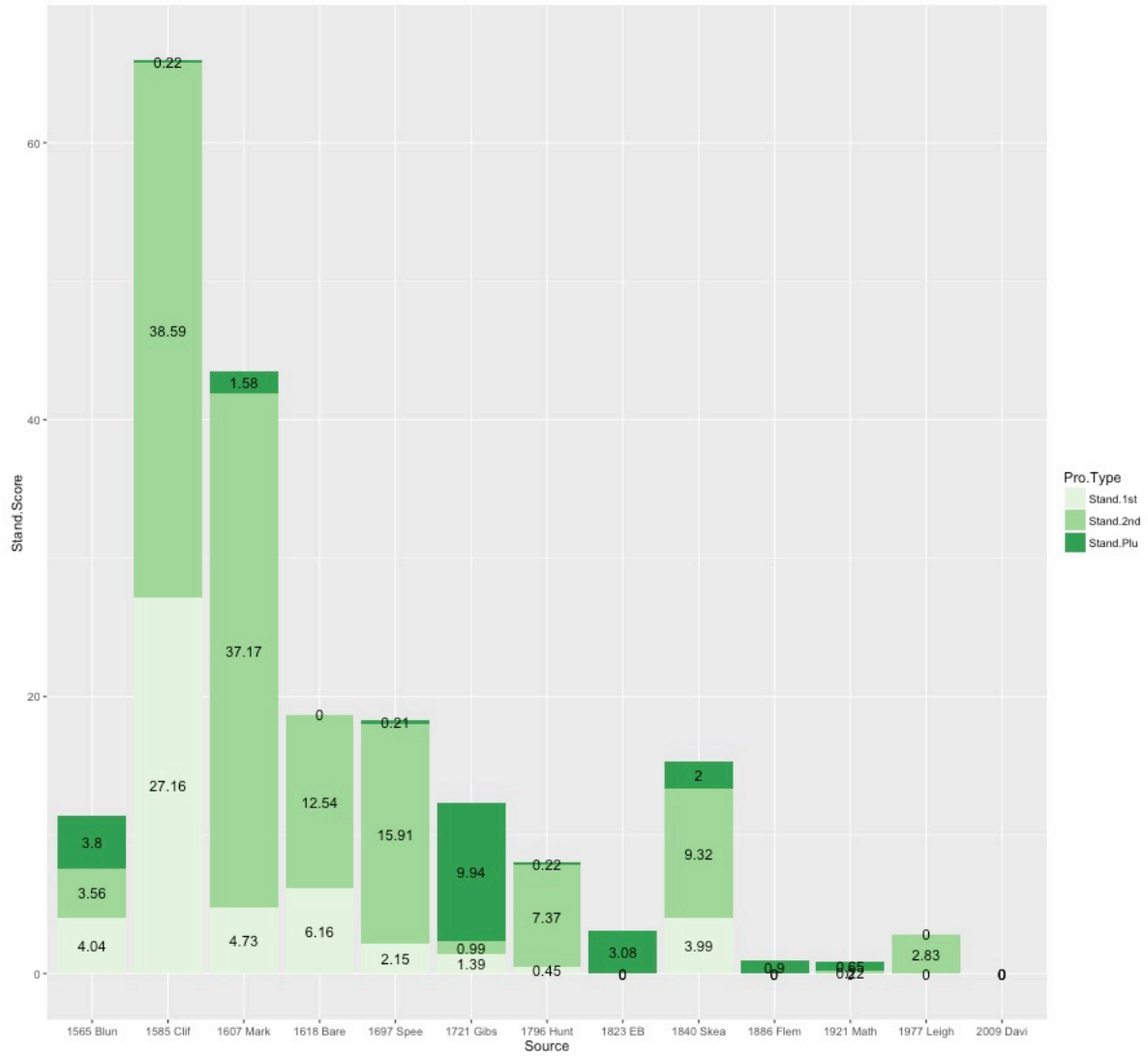


Figure 3.1: Distribution of 1^{st}- (sg.&pl.) and 2^{nd}-person pronouns in the corpus (n per 1,000 words)

Table 3.2: Distribution of pronouns in the corpus (n per 1,000 words)

	Blun.	Clif.	Mark.	Bare.	Spee.	Gibs.	Hunt.	Kirb.	Skea.	Flem.	Math.	Leig.	Davi.
1^{st} (sg.)	4.04	27.16	4.73	6.16	2.15	1.39	0.45	0	3.99	0	0	0	0
2^{nd}	3.56	38.59	37.17	12.54	15.91	0.99	7.37	0	9.32	0	0.22	2.83	0
1^{st} (pl.)	3.80	0.43	1.58	0	0.21	9.94	0.22	3.08	2.00	0.90	0.65	0	0

335). In most texts, the 2^{nd}-person pronoun is the most frequent pronoun used (e.g., Clifford; Markham; Baret; Speed; Hunter; Skeavington; Leighton-Hardman). It is in these texts, too, that it is particularly the 1^{st}-person plural pronoun that often plays a very marginal role. However, it may also be observed that in a number of other texts, the most frequent pronoun forms attested are 1^{st}-person plural pronouns (e.g., Gibson; Kirby; Fleming; Matheson). Only in the earliest text sample by Blundeville are 1^{st}-person singular pronoun forms proportionally the most frequent category, although frequency differences between forms are rather small in this particular text sample (cf. table 3.2).

A Kendall's-τ test, a non-parametric test for statistical (in)dependence between ordinal (or rank-transformed) variables which is more robust than Spearman's-ρ, may serve to indicate whether the data drawn from the 13 texts in the corpus may actually be interpreted as showing an overall trend (Hilpert & Gries, 2009). Via rank-order correlations, the sequential order of the standardised frequencies per pronoun (n per 1,000 words of text) is assessed. Figures may be interpreted such that in the absence of a trend, a Kendall's-τ value will be close to 0, whereas a value close to 1 indicates an increase in frequency, and a value approaching -1 conversely indicates a decrease in frequency over time (cf. Hilpert & Gries, 2009, p. 390). The data in table 3.3 provides correlation coefficients together with two-tailed p-test values for each of the three pronoun classes over the 13 data points in our corpus.[9]

Table 3.3: Kendall's τ on standardised scores of pronouns across 13 texts

	1^{st} (sg.)	2^{nd}	1^{st} (pl.)
Kendall's-τ	-0.714	-0.588	-0.275
$p_{two-tailed}$	0.001	0.005	0.2

Given the data in this table, the relatively high negative value for the 1^{st}-person singular pronoun suggests that there is indeed a very significant downward trend in the use of this pronoun (Kendall's-τ= -0.714, $p_{two-tailed}$= 0.001). Usage of the 2^{nd}-person pronoun also shows a very significant decrease over time in these 13 texts, although a lower τ (-0.588) suggests that this decrease is not as strong as that for the 1^{st}-person singular pronoun ($p_{two-tailed}$= 0.005). Even weaker is the τ-value for 1^{st}-person plural usage (Kendall's-τ= -0.275), but since the p-value for this pronoun class is non-significant, the data rather indicates no distinguishable trend should be assumed here

[9] As some of the later texts in our data show equal cell values, this creates ties (as in, a draw in ranking between two figures) which prevent the calculation of exact p-values. In addition, minor differences are obtained using different packages to calculate τ in R (i.e., `cor.test()` and `Kendall()` in package "Kendall"), which however do not change the conclusions drawn on the basis of the results obtained here. The figures in table 3.3 are based on the former function.

($p_{two-tailed}$= 0.2). In the following sections, the frequency and use of the three personal pronoun categories will be dealt with in separation.

1st-person pronouns (singular)

As could already be gleaned from figure 3.1, 1st-person pronouns are particularly well attested in the current selection of instructional texts up until roughly the middle of the eighteenth-century, with the text by Skeavington a notable late exception with a rate of 3.99 attestations per 1,000 words. The use of 1st-person pronouns in Clifford is particularly marked, but for a more elaborate discussion of this particular text the reader is referred to section 3.2.1. Table 3.4 provides figures for the raw frequencies, standardised frequencies as well as total word count per text.

Table 3.4: Distribution of 1st-person (sg.) pronouns in the corpus

Source	word count	*n*	*n* per 1,000
1565 Blundeville	4,209	17	4.04
1585 Clifford	4,639	126	27.16
1607 Markham	4,439	21	4.73
1618 Baret	4,386	27	6.16
1697 Speed	4,652	10	2.15
1721 Gibson	5,028	7	1.39
1796 Hunter	4,480	2	0.45
1823 Kirby	4,547	0	0
1840 Skeavington	4,507	18	3.99
1886 Fleming	4,437	0	0
1921 Matheson	4,599	0	0
1977 Leighton	4,588	0	0
2009 Davies	4,308	0	0

The examples provided in (1) may illustrate part of the range of discourse functions that can be found in the horse manuals under study. For example, the *'I wishe'*-phrases in (1-a) convey a demand for a physical act by, it is assumed, the reader/addressee. Note, however, that the target of the directive remains implicit in this utterance (see also chapter 4).

(1) a. 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, neyther would I wishe him to be fedde all the day longe, but rather from noonetyde, untill it be iii of the clock in the afternone, to be brydled, and to stande champying thereon, with his tayle tourned toward the maunger. Blundeville (1565)

b. Concerning the watering of a running horse there is a difference betwixt some mens opinions and mine, but because I have spoken thereof heretofore, I will but

touch their opinions briefly, and referre the taking or leaving of them, as (in their iudgement) they shall finde reason to perswade. Baret (1618)

c. Such is the method of fattening horses recommended by some authors of eminence, but I think a much better method is, after bleeding, or giving the horse a smart purge, to turn him out to grass, or, if in the winter, to give him plenty of good hay and corn, with now and then a warm mash, and a moderate share of exercise, which will keep his flesh from being loose and flabby, which must be the case with such as are fed in the hasty manner I have transcribed above. Hunter (1796)

d. I shall make some further observations, as it may occur in the daily employment in the stable, and shall proceed to notice things in that routine, which is daily to be pursued. Skeavington (c1840)

The second utterance, (1-b) by Baret (1618), engages with classical authorities and other (potentially contemporary) experts on the subject of horse care. The 1^{st}-person pronoun is used in a context here that explicitly stresses the difference between the opinion of the writer as opposed to those of others. The example in (1-c) by Hunter uses a 1^{st}-person pronoun in a similar fashion to distinguish the opinion of the author from those of the authorities; the writer creates a platform for his own presence in the text or discourse stage. Rhetorically, the 'move' in this utterance also sets up a contrastive reading, which heavily relies on the non-canonical word order and is motivated by requirements of information structure. One possible alternative would have been '[. . .] but to turn him out to grass after bleeding, or giving the horse a small purge, is a much better method I think'. However, such a rewording probably does not do justice the informational prominence of this 'better method' proposed by the author, which ends up in end-focus in (1-c).

The two 1^{st}-person pronouns in sentence (1-c) both seem to appear in metacomments, although of different qualities: whereas the former is used in a clause that serves to focus the reader's attention to a certain piece of information which is to come next (i.e., on discourse *content*), the second 'I' occurs in a postnominal reduced relative clause which, next to its grammatical function, also serves a discourse-structuring function by establishing a textual link back to a discourse topic (the 'hasty manner' [of feeding]) which was dealt with in the prior co-text. In the last example sentence provided here, (1-d), the use of a 1^{st}-person pronoun in a text-structuring context is even more explicit: the writer intervenes to provide clear signposts for the discourse content that the reader may expect in the ensuing discourse.[10]

Another strategy for an author to intrude in the flow of discourse is by way of

[10]Note also how the second member of the coordinated VP contains a deleted pronoun ("and Ø shall proceed"). Although certainly not restricted to this example (cf. (1-b)), it signals that simply relying on overt pronouns (and using their frequency as a metric for the characterisation of a text as involved or detached) may be overly simplistic. This is especially apparent in text domains that tend to go for complex (paratactic or hypotactic) coordination structures, as seen in some of the examples above.

self-reference using a full noun phrase. This seems a dispreferred strategy that is only mentioned here for the sake of completeness. Only two cases of this kind were found in the current corpus, both in the early twentieth-century sample by Matheson (1921): "In the succeeding paragraphs the author will deal with matters such as grooming, feeding, clipping, [...]" and "The author strongly recommends all horse owners to add a proportion of beans to their forage, [...]". Although it is stressed that this is a minority pattern which in all likelihood indicates idiosyncratic usage, it puts the lack of 1^{st}-person pronouns as signs of authorial presence in samples after the arrival of the twentieth-century in a somewhat different light.

With respect to the discourse-structuring function that is connected to the use of 1^{st}-person pronouns, particularly those highlighted in section 3.1.1, it may be observed that the 1^{st}-person pronouns illustrated in (1) are used in contexts that provide structure to the discourse *content*, but do not overly much contribute to a strictly hierarchical discourse structure or schematic representation of the organisation of the discourse (see especially chapter 6.1).

2^{nd}-person pronouns

To find a frequent use of 2^{nd}-person pronouns for purposes of instruction and a sense of direct involvement might not be surprising in a text domain such as the current one. However, with exception for the texts by Hunter; Skeavington; Leighton-Hardman, it is somewhat remarkable that the use of 2^{nd}-person pronouns is negligible after the advent of the 18^{th}-century. One of the key pointers for such a finding may be sought in the drift towards an impersonal, informational style of writing in professional registers (cf. Biber & Finegan, 1989, 1997) and our sampling from potentially more information-oriented registers in the Late Modern English period. Table 3.5 provides absolute and standardised frequencies for 2^{nd}-person pronouns in the current corpus.

As far as the distinction between 2^{nd}-person singular and plural pronoun use is concerned, we cannot establish with any degree of certainty that these pronouns exclusively refer to singular rather than plural addressees. However, from the cross-section of examples in (2), it may be assumed that the majority of these pronouns are intended as 2^{nd}-person singular pronouns. In addition, these surface forms are not generally interpreted as non-referential uses of the pronoun (cf. Stirling & Huddleston, 2002, pp. 1467-8).

(2) a. Thus you see that Camerarious woulde have a horse be servyd with provender thryse a daye: Blundeville (1565)

b. In the sommer season your running fountaine is the best, for it is the coolest, and in the winter your deepe Well water is best, for it is the warmest. Markham (1607)

Table 3.5: Distribution of 2^{nd}-person pronouns in the corpus

Source	word count	n	n per 1,000
1565 Blundeville	4,209	15	3.56
1585 Clifford	4,639	179	38.59
1607 Markham	4,439	165	37.17
1618 Baret	4,386	55	12.54
1697 Speed	4,652	74	15.91
1721 Gibson	5,028	5	0.99
1796 Hunter	4,480	33	7.37
1823 Kirby	4,547	0	0
1840 Skeavington	4,507	42	9.32
1886 Fleming	4,437	0	0
1921 Matheson	4,599	1	0.22
1977 Leighton	4,588	13	2.83
2009 Davies	4,308	0	0

c. When you have given the Horse this Scowring, rub him well all over with Whisps and a Curry-comb; Speed (1697)

d. But if your Horse be already fat, you ought to observe diligently what Quantity of Food, and what Degrees of Exercise are sufficient to keep up the same Order and Oeconomy in his Body, and likewise the same Degree of Activity and Spirit; Gibson (1721)

e. [...] if you do not allow yourself sufficient time, things cannot be done as they should. Skeavington (c1840)

f. Instinct will usually tell you if anything is wrong the moment you enter the box but a minute or two spent looking at the horse will not be time wasted; Leighton-Hardman (1977)

The first sentence, (2-a) drawn from Blundeville (1565), shows the use of a 2^{nd}-person pronoun in a concluding or evaluating context which involves a reference to a classical authority, Camerarius. In the category model of directives by Hyland (2002b), this usage is indexed as requiring a cognitive rather than a textual or physical act of the addressee.

Example (2-b), in turn, illustrates the practice of replacing an article by a possessive personal pronoun, which is also regularly attested in a number of Early Modern medical recipes studied by Marttila (2011, p. 143).[11] Although a number of cases of this use of the possessive pronoun may be found, such use is infrequent in the current corpus. In fact, the lion's share of possessive pronouns serve as determiners in the NP 'your horse(s)', and maybe somewhat more remarkably, with NPs referring to parts of the body 'your hand(s)/leg(s)/mind', etc. cf., Grlach (1992, pp. 748-749) finds that in

[11]Note that there is no indication in the text to suppose that this 'fountain' and 'well' should be owned by the reader and/or horse-owner.

Middle and Early Modern English recipes, the use of possessive determiners instead of articles may mark informality and/or reader-proximity after 1500, but that this usage may have become archaic or socially stigmatised after the 18^{th}-century. The lack of such possessives in our later text samples may coincide with this social stigma or, even more likely, may be connected to a register drift toward more formal and impersonal-informational linguistic characteristics.

Example (2-c) provides a direct address to the reader and horse owner, in which the writer obliges the addressee to carry out a physical act in the real world. Intuitively, this sentence illustrates the most prototypical form of 2^{nd}-person pronoun usage in texts with a procedural function.

The 2^{nd}-person pronouns in (2-d) and (2-e) are part of a conditional context. However, the exact nature of these conditionals illustrates some of the variety in discourse function and degree of obligation on the part of the reader in procedural texts in the current corpus: the conditional in (2-d) takes the form of a simple if-statement (*if situation X, then perform act Y*), whereas the negative statement in (2-e) is of a different nature: it conveys a moral obligation on the addressee to heed the stated warning (e.g., *'make sure you allow yourself sufficient time, or else ... '*).

Example (2-f) from Leighton-Hardman (1977) seems remarkably involved for a text sample that is generally low in personal pronouns overall (see table 3.5). In comparison to other samples taken from texts after the mid-19^{th}-century, however, the text by Leighton-Hardman still contains a considerable number of such pronouns.

1^{st}-person (plural)

The absolute counts, standardised counts (per 1,000 words) and the total word counts per text for the use of the 1^{st}-person plural pronoun are provided in table 3.6. Of the horse manuals which employ the 1^{st}-person plural pronoun, it is particularly the text by Gibson (1721) that stands out for its abundant use (n= 50, or 1.0 % of total word count). As was already observed in the discussion of figure 3.1 above, the use of 1^{st}-person plural pronouns appears to be marginal in the current corpus. However, the sharp decrease in usage over time seen for the 1^{st}- and 2^{nd}-person (singular) forms is largely absent with the 1^{st}-person plural.

Although absolute frequencies remain low throughout, the use of the 1^{st}-person plural towards the later end of the sample set might tie in with the observation that the use of 'inclusive we' is quite common in PDE formal registers with an informative discourse purpose, particularly academic writing (Hyland, 2001). Biber et al. (1999) provide figures from which it can be gathered that the use of 'we' in academic writing does not exceed the absolute use of this pronoun in other registers (i.e., conversation, fiction and news). At the same time, however, it is the most frequent pronoun used in

academic writing, and its use is roughly on par with the frequency with which it may be found in other written registers (which generally have considerably higher frequencies for other personal pronouns; Biber et al., 1999, p. 334).

Table 3.6: Distribution of 1^{st}-person (pl.) pronouns in the corpus

Source	word count	*n*	*n* per 1,000
1565 Blundeville	4,209	16	3.80
1585 Clifford	4,639	1	0.22
1607 Markham	4,439	7	1.58
1618 Baret	4,386	0	0
1697 Speed	4,652	1	0.21
1721 Gibson	5,028	50	9.94
1796 Hunter	4,480	1	0.22
1823 Kirby	4,547	14	3.08
1840 Skeavington	4,507	9	2.00
1886 Fleming	4,437	4	0.90
1921 Matheson	4,599	3	0.65
1977 Leighton	4,588	0	0
2009 Davies	4,308	0	0

Regarding attestations of 1^{st}-person plural pronouns in the current corpus, we may distinguish between two particular uses based on the referential quality of 'we' (see also particularly the 'secondary uses' of 1^{st}- and 2^{nd}-person pronouns in Payne & Huddleston, 2002, p. 1467ff). On the one hand, we find a use of the inclusive first person plural (or 'workshop we'; cf. Marttila, 2011), which is expected to refer to the writer as well as the reader in our sample (e.g., "This subject, therefore, requires some share of our attention." Kirby, 1823, see (3)). It may be noted that the use of 'we' to refer to entities in the discourse co-text (e.g., characters in a fictional dialogue) or to other third parties, rather than to the interlocutors in the discourse context (producer/receiver), as for example mentioned by Taavitsainen (1997), was not attested in the current sample. In addition, our findings may connect to the suggestion that the use of workshop-we in contemporary scientific writing is a(n) (early-)modern development, since Taavitsainen (2000, p. 196) does not find any attestations of the inclusive 1^{st}-person plural pronoun in the early English medical texts (1375-1550) she studies. The other use of the 1^{st}-person plural pronoun, 'we' used as the self-referential majestic plural (or 'royal we'), is also occasionally encountered in the current corpus (cf. examples in (4) below).

(3) a. Now if any man demaund if it have those faults why it is used so much in *Italy*, I answere, that their Barlye and our is of a contrary nature, and doth not offend so much, yet neither of them both to be esteemed for good provender, where oates are to be got. Markham (1607)

b. OATS are the best, and most suitable kind of grain for feeding horses at present known amongst us, for when they are kept till thoroughly dry, there is no danger

of those disorders attacking the horses that are fed with them, [...] Hunter (1796)

c. [...] but their wine is very weak, and is probably not so wholesome for horses as our ale. Kirby (1823)

d. With us, barley is apt to scour horses, and make their urine red, especially at its first being given. Kirby (1823)

e. [...] when the strain is greater, then the amount should be increased, else we shall have waste without replenishment, and premature wearing out. Fleming (1884)

The first example, (3-a), may show traces of scholastic thought style in the way it enters into a dialogic mode. The address seems as much directed to critical contemporaries, among which the reader may also potentially be gathered, as to the opinions of classical authorities, however, with Markham precipitating possible objection from his audience. Such stance-taking may be taken as an illustration of an author's awareness of persons in the immediate context of discourse, rather than a focus on those outside of it. This example neatly illustrates how, for the discourse purpose of instruction, the procedure ('how') goes hand in hand with explicitly arguing the 'why' in our text domain. In the earliest text samples, it seems readily acceptable for the 'why' to include the practice of trying to convince the reader of a certain procedure with anecdotes and best-practice examples, which in later texts is replaced by references to a new form of authority, that of empirical findings drawn from natural science.

Examples (3-b), (3-c) and (3-d) are quite illustrative for the use of the inclusive 1^{st}-person plural pronoun in our corpus, in implying or eliciting a geographic (regional or national) frame (e.g., *"Now there bee of our English writers which would have your horse to drinke verie much"* Markham, 1607). This frame is particularly striking in example (3-d), where the preposed PP serves to set up a contrastive reading which allows a clear distinction between 'us' (the English/British) and 'them' (in this case 'other countries', particularly those with warmer climates, where barley is used as a common fodder).

A late example of 'inclusive we' can be found in (3-e). The exact 'inclusive' quality of the pronoun in this example may be somewhat questioned, however, as the author and addressee only take part in a metaphorical sense. The closest alternative reading to the phrase in (3-e) may be an impersonal paraphrase such as *X obtains* or *there shall be X*, since the 'waste' that is spoken of refers to a type of waste internal to the horse's muscles, and not a physical substance that the horse-owning audience is able to deal with in any active form. In addition, it may be noted that the remaining three cases of 1^{st}-person plural pronouns in the late 19^{th}-century text sample by Fleming (cf. 3.6) are all used with an impersonal reading. Despite the impersonal nature of this usage, the 1^{st}-person plural pronoun does seem to include and pertain to primarily those involved in the management of the horse. Such cases are therefore included as having 'inclusive

we'-reading, albeit a marginal one.

Examples of the majestic plural in (4) can be found from the earliest up until the latest text sample that contains a 1^{st}-person plural pronoun (cf. table 3.6). The distinction between the majestic plural and 'inclusive we' is not as clear-cut in our earliest examples, however, possibly due to the use of verbs such as *talk* and *speak* which presuppose active interlocutors and could thus be taken to be inclusive rather than self-referential. Categorising such cases as 'royal we' rather than 'inclusive we' rests partly on the decision to rely not exclusively on the type of verb used, but rather to go by the semantic agent: although the author sets up a dialogic stage, the only interlocutor performing the action (talking/speaking) is the writer himself (cf. "Hetherto *we* have talked onelye of the horses meate, now lette *us* speake somwhat of his dryncke." Blundeville, 1565). See also the discussion of interactive versus egocentric, solipsistic use of 1^{st}- and 2^{nd}-person pronouns in Taavitsainen (1997, pp. 247-251).

(4) a. Agayne in the sprynge tyme, or at the fall of the leaffe, horses woulde be fedde for a certayn dayes with grasse, to scoure them, and to make them cleane within, whereof wee shall talke hereafter in his proper place. (Blundeville, 1565)

b. we judge it might not be unnecessary, before we proceed to the particular ordering of Horses, with respect to those things, to take some notice, in the first place, of their proper Food, and the Vices to which some are addicted in feeding; (Gibson, 1721)

c. The observations we are about to make will chiefly apply to the horse. (Kirby, 1823)

d. We recollect meeting in the travellers' room, at an inn in Birmingham, a respectable butcher, who kept some good Horses. [...] We believe his name was Allcock. (Skeavington, c1840)

e. We have known horse owners feed their animals five times a day, but four times is quite sufficient, and if the exigencies of circumstances demand it, horses will thrive quite well on three feeds per day, but this statement does not disturb the maxim already laid down. (Matheson, 1921)

Although the initial 1^{st}-person plural pronoun in example (4-b) by Gibson seems a clear case of the majestic plural, the interpretation of the second 'we' in this example is somewhat ambiguous. Although it could be interpreted as inclusive (both the reader as well as the writer will, in the flow of discourse, proceed to this topic in due course), it is clearly the author who is setting the agenda. Thus, rather than this 'proceeding' being a joint venture between interlocutors, the reading that the *I* (i.e., the semantic agent) will proceed, and the *you*-element that is implicitly included in the 'we' will follow (the semantic patient), seems more likely.

In example (4-c), there is little doubt who will present the observations presented in the ensuing discourse. This use of the majestic plural by Kirby (1823) seems exemplary

for informative-impersonal writing found in academic discourse, it seems.

Maybe somewhat surprisingly, the examples of the majestic plural in (4-d) and (4-e) appear inside anecdotal text fragments. Rather than the impersonal stance, which may be associated with Late Modern and contemporary English informative writing, these utterances signal a clear presence of the author in the discourse (i.e., "writer intrusion"; cf. Hyland, 2005, p. 190). In Early Modern English text fragments, the pronoun used in such anecdotal sections is almost exclusively the 1^{st}-person pronoun *singular*, with the plural form being reserved for metatextual comments as illustrated in (3) and (4) above.

A special case of Reader Involvement: Christopher Clifford's *Schoole of Horsmanship*

The text sample by Clifford was identified as an outlier in the discussion of figure 3.1 in section 3.2. Part of the reason for why it is treated separately here, and not in conjunction with other Early Modern horse manuals with a high pronoun frequency (e.g., Markham (1607)), is because the current sample does not appear as a monologic text. Rather, the chapter in Clifford (1585) from which the current text sample was drawn is written in dialogue form.[12]

This text form is not at all exceptional in instructional writing, and the mimetic dialogue form of writing in fact seems intimately connected to the rise of handbooks as a new genre in the fifteenth-century English vernacular (cf. Taavitsainen, 1997, 1999, 2000, 2004). Texts such as these can usually be placed on a continuum between the older tradition of scholastic question-answer texts and the most fictional mimetic dialogues, which next to instructing also had an entertaining discourse purpose. In the current sample, however, the fictional frame seems to primarily serve a functional purpose rather than provide any textual cues that might indicate entertainment. One of the key features of such mimetic dialogues is the framing of the text as an imaginary discussion between a novice and expert (or patient/physician, etc. Taavitsainen, 1997). The current text is no exception, with the author ('Clifford') reserving the role of expert for himself, and staging a stable boy named 'Kingdon' for the role of novice, supplying questions or positing statements to which the expert replies (cf. (5)).

(5) Kingdon.
I have seene diverse men would give their horses foure handfulls in the morning and eight at night.

[12] It should be noted, however, that although the monograph by Clifford is to a large extent written in dialogue form, some later chapters are written in monologic mode. Next to this text-internal variation, this text also illustrates the heterogeneity of the current corpus, containing e.g., dictionary/encyclopedia entries, professional (veterinarian) advice texts and academically-oriented instruction, as well as popular lay tracts such as monologic manuals and mimetic dialogue text forms.

> Clifford.
> I graunt that thou hast seene it, and so have I, and also have proved it, for I have made mine horses therewith not onlie take with laske, but it hath pased whole through them, even as they did eate it. And I have also so cloied or glutted diverse horses therewith, that they have utterly abhorred their provender: Therefore I would with you to give your horse his provender often, and by little at once. (Clifford, 1585)

The text by Clifford is being dealt with here separately because, in contrast to other texts in the corpus, the *I*s and *you*s here do not necessarily refer to the interlocutors in the discourse context, but may rather refer to the (fictional) characters in the discourse itself (cf. Fludernik, 1993, 1995). Although the distinction between the *I* of the 'narrator Clifford' and the 'author Clifford' may be negligible, the address terms used by this authorial presence to refer to his audience may be all the more noteworthy.

As was already observed above, the text by Clifford contains *th*-forms for the 2^{nd}-person pronoun; roughly one in three (29.8%, or 53 on a total of 179) of the 2^{nd}-person pronouns in the sample takes this form. As a text printed in the second half of the sixteenth-century (i.e., 1585), the use of such forms may be late, but certainly not exceptionally so (cf. Nevalainen, 2006, on the use of *thou* in texts in the Helsinki Corpus; p. 193-6).

In the current sample, many answers make use of either *th*- or *y*-forms throughout. The use of *you* and *thou* does not appear to be consistent, and switches occur even within answers (cf. example (5)). However, nor are these forms in completely free variation: the novice Kingdon, for example, consistently addresses Clifford with the more deferential 'you'. In addition, following a switch from *th*-forms, *you* is generally used consistently throughout the answer. Only one example in the current sample was found which displayed the reverse pattern, with a *th*-form following the deferential 2^{nd}-person pronoun (e.g., *"Clifford. That must I referre to your owne iudgement, for that some horses may bleede more than others by a great deale, therefore thou must take heede first, knowe well thy horses qualitie and strength, and afterward let him bloud accordingly."*; Clifford (1585)).

In light of the orientation towards his readership, the interpretation of such examples is far from straightforward. Taavitsainen (1997, p. 246) notes how even if the 2^{nd}-person pronoun may be used to address a fictional character in the text, the actual reader will likely be involved by way of "the invocation of a more generalised *you*, with a pretense of being applicable to the current reader by virtue of its gnomic truth value" (based on Fludernik, 1995). To what extent the actual reader is meant to take cues from the variation between *you*-forms and *thou*-forms is unclear based on the evidence in the

current sample.[13]

With respect to the referential nature of the single case of a 1^{st}-person plural pronoun, three interpretations are available: either as referring to the author/narrator (i.e., the majestic plural), as referring to the fictional characters in the dialogue (i.e., inclusive, but rather with respect to the narrator and textual-addressee, the 'text participants', and not including the interlocutors in the exophoric context) or as including the author and reader as discourse participants (i.e., Clifford the writer and his real-world addressee, in addition to any other potential discourse participants such as fictional characters).

Although the pronoun in question occurs rather towards the end of the example in (6), its interpretation benefits from the immediate discourse context. After the global topic of bloodletting is introduced, the narrator Clifford seems to stray off-topic. It is not Clifford himself who returns back to the global discourse topic, however. Rather, the possessive pronoun is contained inside a textual metacomment (or potentially, a metacomment inside a metacomment, see the square brackets in the original), and thus does not necessarily pertain to the scripted dialogue between these fictional characters. In fact, the metacomment occurs in the second turn by Kingdon, which redirects the flow of discourse back to the global discourse topic of bloodletting (*"This is by the waie (but to our purpose:)*", Clifford (1585)). This interruption may well be a production error or editorial mistake, but crucially, it illustrates how the 'authorial-Clifford' intrudes in the discourse *outside of* the text section allocated to the fictional 'expert-Clifford' (that is, where author and narrator naturally overlap). It seems therefore not altogether unlikely that the 1^{st}-person plural pronoun here is used to include the actual author and his real-world addressee, rather than any other combination of exo- or endophoric interlocutors.

(6) Kingdon.
But were it not best in the spring time to let anie horse bloud, for that I have heard some of the opinion, that it is good, for that the horses of Polonia, as they say, let themselves bloud once in the yeere, and that in the spring time?

Clifford.
I graunt they say so, but those that teach suche untruthes as these be, I may well compare them to the learned fooles which carrie their science in their sachels, and their wisedome in their lippes, speaking opinions, and what they have heard, which is intollerable in such men to recount what they have read, but to the good rider, souldier, keeper, or farrier that will have credits given to his wordes, hee must not recount what hee hath read, but what hee hath seene and done with his hands.

[13]This is further compounded by the fact that in his address 'To the Reader' in the preface to his manual, Clifford also switches from *thou* to *you*.

> Kingdon.
> This is by the waie (but to our purpose:) I pray you teach me in what vaine it is best to let my horse bloud, and what order is to be observed therein.
>
> Clifford.
> First, thou shalt let him bloud in the necke vaine, [...] (Clifford, 1585)

Further analysis of such pronouns in terms of discourse participants and referentiality is beyond the remit of the current chapter. Suffice it to remark that the fictional novice character in dialogic texts such as seen in Clifford (1585) to a considerable extent serves to supply the discourse topics for conversation. Such 'stooges' may thus be an implicitly essential device for prioritising and structuring the discourse content, as well as ensuring the (dis)continuity of discourse topics, in dialogic texts of this kind.

3.3 Discussion and conclusion: Pronoun use and personal involvement

As may have been clear from the previous section, surface forms such as personal pronouns offer a ready cue to revealing a communicative function of expressing personal affect and promoting reader involvement. In general, what the findings in the current chapter indicate is that the use of personal pronouns steadily decreases in the texts in our corpus, but that the trajectories are different for these pronouns according to person and number, as well as their specific use in context.

1^{st}-person singular pronouns, which may indicate a more author-centred rhetorical style (cf. Atkinson, 2001) are particularly prevalent in the first half of the corpus. Although the text fragments here do not provide enough material to definitively substantiate the claim that we are dealing with a clearly author-centred style in the text as a whole, what does appear to be the case is that these pronouns are particularly prevalent in metadiscursive comments, which may be taken as signs for authorial intervention (i.e., the presence of the author in the text). The decrease in explicit self-reference by way of the 1^{st}-person singular is particularly noticeable in the run-up to the Late Modern English era, with the late 18^{th}-century a seemingly pivotal time frame.

The use of 1^{st}-person plural pronouns may be mentioned in connection to this decrease, since in the guise of the majestic plural it is often regarded as a substitute form of self-reference for *I* in certain contemporary registers. Although figures for the 1^{st}-person plural pronoun have not been specified for the two functional uses of this form, i.e., the royal-we and workshop-we, it seems unlikely that the former form takes on the function of self-reference in our corpus. Figures show that although there is a somewhat

more noticeable use of *we* in Late Modern English texts, especially in comparison to the use of the other pronoun forms, the standardised scores do not indicate that this increase is a direct consequence of the abandoning of 1^{st}-person singular *I*. The proportional use of 1^{st}-person plural forms in the second half of the corpus is only a fraction (more than on one occasion less than 1 occurrence per 1,000 words) of the use of 1^{st}-person singular *I* in the first half of the corpus. In addition, usage of the we-lexeme also represents occurrences of workshop-we which generally involves both writer and reader. If anything, these figures thus suggest a general decrease in author-centring, both by way of 1^{st}-person singular *I*-forms and the 1^{st}-person majestic-plural *we*-forms, rather than a development in forms to express this discourse function. Changing register conventions (i.e., register shift) may be one important cause for this perceived difference in periodic preferences, although sampling bias cannot be ruled out altogether, nor can developments connected to a change in the make-up of the audience consuming this particular sub-register of popular lore.

Abating use of 2^{nd}-person pronouns, either singular or plural, seems to be in line with this development in the later half of the corpus. Such overt reference to the reader as discourse participant cannot be connected to a decrease of an author-centred style alone, and seems to indicate a general trend towards impersonality, or a more detached style, in Late Modern English horse manuals. It should be kept in mind, however, that idiosyncratic differences exist here too, with the mid-19^{th}-century manual by Skeavington (c1840) being far more personally involved, in terms of both 1^{st} and 2^{nd}-person pronoun use, than any other texts in the corpus published after the first quarter of the 18^{th}-century. If expressing proximity between interlocutors is still a discourse purpose in the detached style of popular lore observed in the second half of the corpus, it is realised by other means than via personal pronouns which are co–referential with participants in the immediate discourse context.

With reference to patterns of register diversification in the history of English (e.g., Biber & Finegan, 1997/2001), these figures suggest that if we were to consider the current grouping of texts as a single genre, it is more in line with the trajectory of professional registers (e.g., science, law, medicine) towards a more impersonal-informative style of writing than with the general trend towards colloquialisation and speech-like features which non-specialist, popular registers seem to undergo (Biber & Finegan, 1997/2001). That is, at least in terms of the use of pronouns as markers of personal involvement.

However, if explicit mention of writer and reader add to discourse structure and discourse coherence, as is suggested by e.g., Taavitsainen (2009) with particular reference to the use of metadiscursive comments, does the general lack of such overt markers as of the start of the 19^{th}-century make such later fragments less coherent? This does

not seem likely, although it might very well be that genre and register shifts also affect this apparent difference. As was already clear from table 3.1, the last text sample by Davies (2009) does not contain any 1^{st} or 2^{nd}-person pronouns. Nevertheless, this text does not appear as particularly incoherent, and it is likely that other grammatical and/or paralinguistic markers help to achieve coherence in such texts. Incidentally, the same goes for (sub-)registers that are characterised by a general absence of 1^{st}- and 2^{nd}-person pronoun use in multidimensional corpus studies (cf. Biber, 1988). What can be said about such text largely devoid of pronouns referring to reader and writer is that they generally work as less personally involved, or more detached, than texts which frequently resort to the use of 1^{st}- and 2^{nd}-person pronouns.

Other surface features, however, may equally indicate such form-to-function signals to express orientation towards the reader. For example, Fludernik (1993, p. 221) regards both 2^{nd}-person pronouns or imperatives as markers of "function of address" in narrative prose, and as was noted above, Hyland (2002b) studies three typical linguistic surface forms of directives (imperatives, modals of obligation and predicative adjective phrases with a *to*-complement clause) across contemporary academic prose genres and disciplines. That there are other devices that may be used to express such personal involvement seems likely, for example via involved/affective markers identified in section 3.1.1, but this cannot be assessed here using the limited focus on pronouns which are co-referentially linked to participants in the discourse. Diachronic multifactorial studies such as Taavitsainen (1993) or Biber and Finegan (1989, 1997) take into account even more features that are interpreted as possible indicators of personal/emotive affect, which offer a more comprehensive approach to personal affect.

However, in light of such form-to-function mappings, it is not usually reported to what extent the selected forms represent a proportion of their associated functional mapping (that is, the share of the discourse function expressed by a particular form). By taking into account pre-defined surface forms only (the 'true positives'), we may fail to account for surface forms which express this same discourse function but are not included in the feature selection set (i.e., the 'false negatives').[14] Although exceptions may be mentioned, see for example Jucker, Schneider, Taavitsainen, and Breustedt (2008) who concern themselves with surface pattern detection (precision and recall) in the context of historical speech act research, such interpretations of proportional representation of a discourse function seem extremely difficult to quantify, or define

[14]In connection with this, one may wonder to which extent the form-to-function procedure above includes 'false positives'; that is, the use of 1^{st}- and 2^{nd}-person pronouns which do not involve any orientation towards discourse participants. Although such a lack of (intended) involvement seems difficult to assess, it is a theoretical possibility at least: e.g., Fludernik (1993, p. 222) finds cases of 2^{nd}-person pronoun use with "no observable addressee function", and Taavitsainen (1994, p. 251) notes cases of the use of 2^{nd}-person *you* as a "projection to *I*, someone addressing the self in an egocentric focus".

formally. Similarly, Kohnen (2004) flags up the issue of charting surface forms as a proportion of the total expression of a certain discourse function as one of four methodological issues in corpus-based research in historical pragmatics.

Nevertheless, it by now seems accepted that future studies into interpreting the discourse function of surface features has to account for observed patterns as well as a systematic assessment of false negatives and false positives. Thus, a form-to-function mapping cannot be satisfactorily approached by interpreting forms contained in the feature selection set alone. For example, section 4.4.2 in the next chapter connects to this issue of the use of 1^{st}- and 2^{nd}-person pronouns in functional contexts associated with the realisation or absence of personal involvement.

In general, the next chapter will approach the issue of expressing personal involvement in the current selection of texts from another angle; one which adds to our understanding of the development towards avoidance of personally involved prose as seen in the current chapter.

3.4 Chapter summary

In this chapter, the use of 1^{st}-person singular and plural pronouns as well as 2^{nd}-person (singular or plural) pronouns is used as an indicator of a personally involved style of prose. Based on the frequencies observed in text samples in the current corpus, Early Modern texts seem to be more personally involved than texts from the Late Modern English period in general. In addition, the use of such markers of personal involvement may serve various overlapping functions, for example through overt authorial intervention for structuring the discourse, or establishing personal proximity through presenting a future act as a joint endeavour by both reader and writer.

Statistical corroboration of these frequencies suggests that the decrease in the use of markers of personal affect is significant for 1^{st}-person singular and 2^{nd}-person pronouns. A trend cannot be established satisfactorily for 1^{st}-person plural pronouns, with frequencies showing a slightly decreasing but non-significant pattern of development for the use of this form. These results indicate that although texts in the current corpus may be analysed as popular instructive-informational prose, at least as far as can be judged on this particular feature, its development is more in line with the evolution of professional registers rather than with non-specialist, popular forms of prose.

Idiosyncratic differences between text samples suggest that although general trends may be observed in the use of personal and possessive pronouns in the corpus, these conventions do not restrict variation in the adherence to such norms by writers in the current sub-genre of popular lore.

Chapter 4

From Function to Form: Addressee Orientation

4.1 Introduction

As a pendant to the previous chapter, another way to study the expression of personal involvement is by way of an onomasiological approach, or function-to-form mapping (cf. Jacobs & Jucker, 1995; Traugott, 2004; Brinton, 2015). Rather than an inventory of linguistic forms and their associated functions in discourse, the focus here is thus on one particular discourse function and the surface realisations by which it is expressed. In any study of this kind, the issue of a *tertium comparationis*, the element that remains fixed in the contrastive analysis of two or more objects, is of methodological importance (cf. Jacobs & Jucker, 1995). When assessing the function of certain forms in historical perspective in the previous chapter, for example, it is the grammatical form which remained fixed, with the contrastive analysis limiting itself to an interpretation of the discourse functions realised by these forms. Conversely, in function-to-form mappings one may identify and compare the different historical surface realisations of a particular function, and it is therefore this discourse function that should be kept stable.[1]

Defining the *tertium comparationis* deserves particular attention in the current case, since realisations that express personal involvement may take many forms. A wide variety of different surface forms and metrics, for example, have positive loadings on the dimension that Biber and others characterise as capturing personal involvement in both synchrony as well as diachrony (cf. Biber, 1988; Biber & Finegan, 1997; Biber,

[1] It is probably naive to assume that in both mappings there is no change in either form or function when it is designated as the stable factor in a comparative approach. Ultimately, opting for a form-to-function or function-to-form mapping is more a matter of methodological *perspective* rather than a methodological attempt to control for internal variation in what could be seen as the independent variable in an experimental design (cf. Jacobs & Jucker, 1995, p. 13, 19).

2001). Following this train of thought, personal involvement may even be described as a global textual or stylistic feature rather than a discourse *function*, with the current design taking on the characteristics of a *feature*-to-form mapping. A more workable design would involve restricting our analysis to one well-defined discourse function that is able to capture a large share of the forms that are associated with this stylistic feature of personal involvement.

An approach that lends itself for such an agenda may be adopted from a well-researched area in the context of historical function-to-form research involving the diachronic study of speech acts. Borrowing from Searle (1976) the notion of 'speech act' as the most basic unit of communication, i.e., the act that a speaker performs by making an utterance, what remains fixed in historical studies that take such an approach is usually either the speech act itself, or its illocutionary point (or purpose).[2]

Jacobs and Jucker (1995, pp. 19-20) note that speech act theory has been suggested as "the main methodological tool in historical pragmatic analysis". However, this type of diachronic research involves a number of delicate issues, not least of which being the major problem of identifying (dis)similarities in "the conditions governing the communication between author and audience" (Jacobs & Jucker, 1995, p. 21, citing Bergner 1992). Such conditions, and by extension, the pragmatic basis for communication, may be markedly different for a past context than in the contemporary setting with which researchers are familiar. Thus, the degree to which a particular speech act is conventionalised, and the contextual conditions in which it would be seen as felicitous, may be regarded as "conditioned by history" (Jacobs & Jucker, 1995, p. 21). Determining the adherence to conventions of speech act use, as well as the conventions regarding speech act realisation strategies, brings to the fore the problem of capturing a speaker's intention in expressing a certain speech act. Stetter (1991) holds an extreme position with regard to this point, claiming that there can be no historical dimension to speech act theory, "since it is impossible to grasp exactly what a speaker meant by his utterance" (Jacobs & Jucker, 1995, p. 19). Nevertheless, it seems vital to be as precise as possible about the circumstances of the utterance used to express a certain speech act,

[2] Jacobs and Jucker (1995, p. 19) propose the illocutionary force of a speech act to be the stable *tertium comparationis*. Based on Searle (1976), however, we argue that it is the point or purpose of an utterance, rather than its force, that we take as the fixed element here. For example, the difference between illocutionary *point* and illocutionary *force* as illustrated by Searle (1976) is that although requests and commands may share the same illocutionary point (which is to get the addressee to do something), the illocutionary force between them is decidedly different (Searle, 1976, p. 3; note, however, that both utterance types may be counted among the directive illocutionary acts).

Further support for this is that there is a marked tendency for expressions of indirect directive speech acts to acquire, through usage, similar meanings as direct directives (cf. Traugott, 1972). In a study of the etymologies of direct directive verbs in PDE, Lau has shown that such verbs derive from much less directive meanings associated with asking, requesting, desiring or showing (Lau, 2015). This suggests that illocutionary force may not straightforwardly be assumed to be particularly stable over time (see the previous footnote).

especially in a diachronic context (cf. Jacobs & Jucker, 1995, p. 21). In the end, what must be acknowledged is that lacking explicit information on usage, deducing historical conventions based on patterns of use is complicated, and involves an unknown margin of error (see §4.2.2 on related issues).

Although an important methodological concern for any sort of research in speech act theory that lacks access to informant judgements, interpreting an author's intentions based on the linguistic realisation of a speech act would take us too far into the realm of (historical) speech act research proper. We therefore limit our focus here to identifying a speech act that is readily available in the text domain under study and which offers an insight into the expression of personal involvement as a feature of a text's stylistic character. Whatever else may have changed in the context of the communicative act (shifts in the make-up of the audience, changing conventions of politeness, etc.), the purpose of these texts (i.e., to instruct how to take care of a horse) as well as the illocutionary point of the act (to get the reader to consider acting in the proposed fashion) are assumed to have remained largely the same in the history of English.

Given the instructional texts used in the current study, an illocutionary purpose that offers itself for our approach may be found in the class of directive speech acts. Directives constitute one of the five main classes of illocutionary speech acts in the taxonomy by Searle (1976), and are described in general terms as "attempts [...] by the speaker to get the hearer to do something" (Searle, 1976, p. 11). Huddleston (2002, p. 853), too, regards directives as a cover term for a range of subtypes, i.e., speech acts proper, and mentions examples such as instructions, advice, commands and requests.[3]

As a successful directive is dependent on a future action by the addressee (cf. §4.2), directive utterances involve a special locus of interaction between addressor and addressee in written text. The degree of personal involvement expressed in this interaction is of particular interest here. Although common disclaimers in the interpretation of speech acts must apply,[4] we thus take the proportion of directive speech acts realised along a scale of (im)personality as an indicator of personal involvement, which in itself

[3]Below, we will also refer to directives as either 'speech acts with directive (illocutionary) force' or 'directive speech acts'. The latter may be somewhat controversial, as directives are generally regarded not as a speech act proper, but rather as a speech act *class*, in the taxonomy proposed by Searle (1976). Nevertheless, as the distinction between subtypes of directives is based on linguistic form as well as syntactic and semantic properties, we maintain the label directive speech act to encompass as broad a range of potential linguistic realisations, including speech acts proper, grouped primarily for their function or purpose.

[4]A personal rather than a more impersonal realisation of a directive may for example be considered to be more persuasive in some contexts, but lacking information on a writer's decision to formulate a more personal directive as opposed to an impersonal realisation, it is impossible to ascertain the extent to which cultural conventions of (im)personality in writing have played into an author's decision in diachrony. In addition, there is no way of assessing the expected degree of success based on the linguistic realisation of a directive, and therefore, a writer's potentially deliberate efforts to opt for such a more 'persuasive' formulation.

may be an important feature of period-specific style in the text domain under study here.

Before we turn to observations regarding personal involvement and style, however, we will address some relevant observations on the use and grammatical form of directives in Present-Day English, followed by a brief overview of literature on directives in the history of English. This will serve as a backdrop for introducing the coding scheme for the utterances in our corpus, which is based on the linguistic realisation of the semantic agent of the action proposed in the directive (cf. section 4.2.4). That is, the agent is realised as either the addressee ('Direct Address'), as another entity but not the addressee (although this entity may be co-referential with the addressee; 'Indirect Address') or as not-realised in the directive utterance at all ('Suppressed Addressee'). The results of our corpus analysis will be discussed next, involving a profile of each text as well as the corpus in general in terms of the degree of personal involvement as provided by these linguistic realisation strategies for a writer's orientation towards his reader. A summary of the findings in this chapter in section 4.5 is preceded by an evaluation of the degree of personal involvement across both directive and non-directive utterances in the corpus.

4.2 Theoretical Foundations: Directive speech acts

Given that our primary aim in the current approach is to use the surface realisation strategies of directives as indicators of personal involvement in instructional writing, a grammatical description of the ways in which such utterances may be realised in English seems appropriate. Instead of referring to speech act research that follows up on the initial work by Austin and Searle, we therefore turn to contemporary grammars of English, i.e., *A Comprehensive Grammar of the English language* (Quirk, Greenbaum, Leech, & Svartvik, 1985), the *Longman Grammar of Spoken and Written English* (Biber et al., 1999) and the *Cambridge Grammar of the English language* (Huddleston & Pullum, 2002) for our main theoretical discussion of directives.

These three grammars all deal with directives to different degrees, with Quirk et al. (1985, pp. 827-833) for example devoting only a few pages to the interpretation and use of directive utterances in English. Since the majority of common examples and issues involved in this work are extensively explored by Huddleston (2002), we will defer to the later work in particular. Similarly, we will not refer to the *Longman Grammar of Spoken and Written English* extensively. Lacking a particular section unifying a treatment of utterances with directive force, references to regular surface forms of directives appear mainly in the context of the use of imperatives (pp. 219-222) and modals of obligation or necessity (pp. 493-496). A useful observation by Biber et al., however, concerns the

fact that imperatives in writing, especially when they occur in the registers of academic prose and news, are usually found as explicit instructions and with direct address to the reader (either in identifiable instructions or with a intertextual function guiding the reader through the text Biber et al., 1999, p. 222). In addition, Biber et al. note that modals of obligation or necessity are commonly used not to mark logical necessity, but rather to express personal obligation in Present-Day English (with the exception of modal *must*; Biber et al., 1999, p. 494).

The most comprehensive discussion of directives in a grammars of Present-Day English, to which we turn now, seems to be available in the *Cambridge Grammar of the English language* (Huddleston & Pullum, 2002). It is offered mainly in conjunction with a discussion of the illocutionary force of imperatives (Huddleston, 2002, pp. 924-944).

4.2.1 Directives in Present-Day English

Although imperatives may be the characteristic clause type for directive utterances, Huddleston includes other clause types such as interrogatives and declaratives as non-imperative directives in his discussion of this speech act class. Whereas the truth value of propositions is the prototypical concern of declarative clauses, directive speech acts are stated to not have truth values (Huddleston, 2002, pp. 929). Rather, what characterises the various clause types that may express a directive speech act is that they are, directly or indirectly, concerned with future compliance with the advocated action of the proposition: "A directive expresses a proposition representing a potential situation: realising or actualising that situation constitutes compliance with the directive" (Huddleston, 2002, pp. 929).

Although all directives thus 'promote compliance', they may do so with varying degrees of strength. The illocutionary force thus depends in part on the clause type of the directive, but is generally affected by a range of factors, e.g., the situational context in which a directive is uttered. Both Quirk et al. and Searle provide examples of such immediate situation factors, noting that differing levels of authority between speaker and hearer may affect the interpretation of the illocutionary force of an utterance, e.g., whether a directive is taken to be a suggestion or an order (cf. Searle, 1976, p. 5), and Quirk et al. (1985, p. 831) remark that the illocutionary force may also depend on the relative benefits that the action may have for each of the interlocutors. Such situational factors are also reflected in three important features that stand out in the categorisation and discussion of directive imperatives by Huddleston (2002), i.e., the relative authority of the addressor, the agentivity of the addressee and the 'wilfulness' of the directive. Regarding the notion of authority, for example, Huddleston (2002, p. 930) notes that the addressor may present him/herself as an (institutionalised) authority in some types of

directives (e.g., orders, commands, demands) in which non-compliance by the addressee is not generally presented as an option. Thus, whether a writer "presents [him]self as invoking the authority to require compliance" partly determines the construction or clause type with which a directive is realised (e.g., *Open the door!* vs. *Open the door, will you?*).

Such 'wilful' directives, in which an addressor presents him/herself as requiring (and expecting) compliance, may be contrasted with 'non-wilful' directives. In the latter category, compliance is suggested as being in the addressee's own best interest ("compliance is something I will, not for my benefit, but rather something I present as being in your interest" Huddleston, 2002, p. 930). Advice (*Don't put all your eggs in one basket*) and instruction (*Mix flour with cold water to a smooth consistency*) both belong to the category of non-wilful directives, as do recommendations, warnings, suggestions and expository directives.[5] Although instructions are non-wilful, and thus in the addressee's benefit, it is noted that compliance is presented as a necessary condition for achieving the relevant goal (or arriving in/at the intended end-state) – be it the use of an appliance, finding a destination or the preparation of a meal (Huddleston, 2002, p. 931).

Expository directives (e.g., *See the example provided in figure 1*) do not deviate from other directives structurally or in terms of interpretation. Rather, these directives receive their label from occurring in a particular text domain. They are particularly prone to occur in expository writing, "to engage the active participation of the addressee" (Huddleston, 2002, p. 931), and may be compared to the directives which seek to elicit a 'textual act' in the categorisation of Hyland (2002b). Huddleston (2002, p. 931) notes that such directives serve a similar discourse purpose to instructions, in that compliance serves "the purpose in hand: in this case, following the speaker's exposition".

In all of these non-wilful directives, the decision to comply with the suggested action is left entirely up to the addressee. This points to a third feature in the discussion of directive imperatives by Huddleston (2002), which is the assumption of the addressee's agentivity with respect to the activity involved. Although it might seem evident that the addressee of an imperative is the likely agent to carry out the future action put forward in the directive, this feature is shown to have ramifications for the surface forms of imperatives, as well as their interpretation.

According to Huddleston (2002, p. 932), the imperative construction itself "leads us to assign an agentive role to the understood subject-referent" (Huddleston, 2002,

[5] In the distinction between the non-wilful directives *suggestion* and *advice*, the accountability of the source is a minor feature that may be of relevance. In the latter, the source of the advice is expected to be able to justify the proposed course of action, whereas this need for justification is non-existent with suggestions (Huddleston, 2002, p. 930)

p. 932). Although imperatives usually lack a subject, it is generally understood to be 'you', the addressee (cf. Huddleston, 2002, p. 925). This assigning of an agentive role to the subject is even the case when the predicate contains a stative VP which in declaratives would be associated with a non-agentive reading. In the declarative *John is patient*, for example, the stative nature of the verb and the non-agentive role of the subject are markedly different from the imperative *Be patient*, in which the agent-addressee is expected to be able to exercise self-control (Huddleston, 2002, speaks of "agentive reinterpretation" in this context; cf. p. 932).

The fact that the agentive role is usually associated with the grammatical subject is further underscored by the observation that passive imperatives are relatively uncommon. "[I]n declaratives whose predicate assigns an agentive role to one of the arguments[,] the argument concerned is aligned with the subject of the active, not the passive" (Huddleston, 2002, p. 932). For example, in the active *John joined Ross* and passive *Ross was joined by John*, only the active utterance has an agentive subject. That this agentive role association with the subject carries over to imperatives is illustrated by comparing the active imperative *Join him* to its passive counterpart *Be joined by him* (which, outside of a very specific context, would seem rather odd). That passive imperatives need not be ungrammatical, however, is underscored by perfectly natural utterances such as *Be warned* or *Don't be intimidated* (examples taken from Huddleston, 2002, p. 933).

A final illustration of the association between agentivity and the subject-referent of an imperative can be found in wishes or examples drawn from advertising. In both cases, we are presented with a situation which is not normally considered to be under our control, but where an agentive role is nonetheless imparted onto the addressee, e.g., *Get well soon* and *Win a sports car for an extra £1* (rather: 'Spend an additional £1 on your ticket to give yourself a chance of winning a sports car'; cf. Huddleston, 2002, p. 933).[6]

Non-imperative directives: Interrogatives and declaratives

Just as with directive imperatives, it seems that these three features, i.e., the addressor's (projection of) relative authority, the (non-)wilfulness of the directive utterance and the assumption of an addressee's agentivity, are readily applicable to non-imperative directives. However, although interrogative utterances may have directive illocutionary

[6] In the next section, §4.3, we will come back to the issue of agentivity of the addressee and will introduce a differentiation based on the interactional nature of directives in combination with the linguistic realisation of the referent for the agent of the action. That is, 1) the addressee is assumed to be the agent of the action, and is realised as such, 2) the agent is linguistically realised but is not directly identifiable as the addressee, and 3) the semantic agent associated with the action is left unexpressed, although the utterance has directive force.

force, Huddleston (2002, p. 939) regards them as having *indirect* directive force at best. Such speech acts with indirect illocutionary force occur when

1. the illocutionary force of an utterance deviates from the illocutionary force usually associated with a clause type (e.g., declarative clause type with associated statement force), or when

2. the propositional content of an utterance is actually different from what a speaker wants to convey with a particular illocutionary force (see Huddleston, 2002, pp. 861-5).

This second situation might be clarified using the example question *Do you know what time it is?* While the propositional content of this interrogative is captured by the declarative statement 'You know what time it is', the indirect illocutionary force of the original interrogative might either be that of an interrogative with question force ('What time is it?') or of an imperative with directive force ('You go to bed!'; cf. Huddleston, 2002, pp. 861-862). Whereas clause types are mutually exclusive, illocutionary force is not: utterances may have more than one illocutionary meaning, depending on their context of use (Huddleston, 2002, p. 859).

One clear difference between imperatives and interrogatives is that in the latter, the agent-addressee is prototypically expressed overtly using the second-person pronoun 'you' as subject, e.g., *Could you open the lid?*. In the same way that imperatives can accommodate 3^{rd}-person subjects, however, interrogative directives can have 3^{rd}-person subjects, e.g., *Can everyone make sure they sign the sign-in sheet that is going around?*. Nevertheless, such interrogatives are judged to be not entirely equivalent to imperatives since they are not addressed to a particular addressee (Huddleston, 2002, p. 941).

Declarative directives, in turn, can occur with either direct or indirect illocutionary force. Direct force is carried by declaratives that contain an illocutionary verb (a verb that denotes an illocutionary act) used performatively, "i.e., to effect the performance of the illocutionary act it denotes" (Huddleston, 2002, p. 859). Such performative constructions, e.g., *I order / advise you to join them*, heavily rely on the performative use of (a limited set of) illocutionary verbs, making it primarily a lexical rather than a grammatical phenomenon (Huddleston, 2002, pp. 859-860).

This fairly restricted set of declaratives which express direct illocutionary force by way of a performative construction may be contrasted with the wide spectrum of directive declaratives with indirect illocutionary force. These may contain a broad range of cases, for example the expression of deontic necessity (*you must come now*), an addressee's future actions (*You will/are going to beg*) or involve an overt expression of a speaker's explicit wants or needs (Huddleston, 2002, p. 941), *I want / need /*

would like you to carry me home.[7]

Based on the above, several surface forms present themselves as expressions that may express directive illocutionary force for the discourse function of instruction (based on Huddleston, 2002):

- Imperatives (direct),
- Interrogatives (indirect), and
- Declaratives (direct/indirect), either:
  - direct force (e.g., with performative verbs),
  - indirect force (e.g., personal pronouns with verbs of desire)

That the range of non-imperative clauses functioning as directive speech acts may be wider than the canonical declaratives in Huddleston (2002) is further illustrated by the directives discussed in studies mentioned above such as Hyland (2002b) and Biber et al. (2004). Both may be thought of as corpus-linguistic form-to-function approaches, although at different levels of granularity: where Hyland mines his data for a set of pre-defined surface forms based on the literature (Hyland, 2002b, p. 222), the method chosen by Biber et al. (2004) is completely data-driven, by including lexical bundles (word-level 4-grams) and offering (functional) interpretations for those attestations with the highest frequencies.

In one particular subcategory of lexical bundles identified by Biber et al. (2004), that of "obligation/directive" bundles within the larger category of "stance bundles", it is particularly verbs which express an addressor's wants or needs ("desire verbs") which are readily attested (cf. Biber et al., 2004, pp. 391-2). Although such obligation/directive bundles primarily occur with second-person pronoun subjects (*You need to know . . .*), a second type of bundles found in this class features a first-person subject in combination with a verb of desire (and usually a second-person object; e.g., *I want you to know* Biber et al., 2004, p. 390). Often, such stance bundles can be analysed as either obligation/directive bundles or 'desire bundles', another subcategory of stance bundles (e.g., *I don't want to* be there when he gets home). Correct categorisation between these subcategories usually depends on contextual interpretation, but what typifies bundles that are analysed as directive rather than desire is that in the former, it is explicitly the speaker's intent that it is the *addressee* who is instrumental in bringing about a certain state or carrying out a certain action. Such bundles thus point to utterances that primarily serve a directive function (Biber et al., 2004, p. 390).

Such expressions involving a desire verb plus a pronoun that refers to a discourse interlocutor are thus associated both with the situational characteristic of personal

[7] Note that Searle (1976, p. 11) identifies 'want', 'wish' or 'desire' as the sincerity condition for the class of directives. A sincerity condition refers to "the psychological state expressed in the performance of the illocutionary act" by the speaker (Searle, 1976, p. 4).

stance (cf. psychological state of the speaker Searle, 1976) as well as directive function (Biber & Conrad, 2009, p. 68). In general, therefore, surface forms with first- or second-person pronouns are analysed as *personal* stance bundles, expressions that are overtly attributable to the speaker or writer, by Biber et al. (2004). These contrast with impersonal obligation/directive stance bundles, which refers to bundles which "express similar meanings without being attributed directly to the speaker/writer" (Biber et al., 2004, p. 389) An example of an impersonal bundle "with no personal pronoun at all, even though they still clearly direct the reader to carry out some action" provided in the context of obligation/directive stance is of the form 'it is important to note that' (Biber et al., 2004, p. 391). This extraposition construction is the only example of this kind of 'impersonal directive' in Biber et al. (2004), but is instrumental in introducing non-canonical word-order constructions in the class of utterances with directive force.

Next to imperatives and modals of obligation, Hyland (2002b) includes such constructions with "predicative adjective[s] expressing the writer's judgement of necessity/importance controlling a complement *to*-clause" as one of three types of surface forms of directives (e.g., 'it is important to observe ...'; cf. Appendix 1 in Hyland, 2002b, for a feature list of the directive surface forms studied). Both Hyland as well as Biber et al. (2004) regards such information structuring constructions as potential carriers of directive illocutionary force (albeit with indirect force, if we follow Huddleston, 2002).

With respect to pattern detection of such 'directive extrapositions', Hyland observes that

> "adjectival predicates with *necessary / important / essential* etc. seem to guide the reader fairly directly to the action stated in the extraposed *to*-clause. These can be compared with those [extraposition constructions] that carry a more evaluative stance, indicating only what the writer considers *interesting*, *surprising*, or *relevant*. Thus surface form alone is an unreliable indicator of directive force and every instance has to be examined in its sentential context to ensure its pragmatic effect"
>
> (Hyland, 2002b, p. 217)

Although the issue raised in the second part of this excerpt is mainly a concern for large-scale, automated form-to-function searches in the study of directives, it dovetails general issues surrounding the study of illocutionary meaning in the history of English (cf. Kohnen, 2001b, 2002, 2004, 2008b).

4.2.2 Directives in the history of English

Kohnen (2008b) identifies as a major problem in corpus-based diachronic speech act research the issue of not knowing, and therefore, not being able to access in a corpus "all the manifestations of a particular speech act in a past period" (Kohnen, 2008b, p. 296).

From tracing one particular manifestation (i.e., a linguistic form, such as imperatives) in historical data, it is impossible to assess whether an observed trend is also applicable to other manifestations of the same speech act. As a result, a 'linguistic form-to-speech act' mapping is judged an extremely [risky business] in diachronic speech act analysis (Kohnen, 2008b, p. 296). Next to this problem, Kohnen (2004) introduces three other methodological issues in diachronic speech act research:

1. Retrieving all relevant forms of a speech act (detecting 'hidden manifestations', i.e., false negatives)

2. Interpretation of cases due to problems with interpreting illocutionary meaning, in diachrony (e.g., detecting 'pragmatic false friends', i.e., false positives)

3. The issue of not knowing the extent to which corpus manifestations represent underlying directives in actual use (i.e., external or ecological validity)

4. Drawing conclusions based on a small number of attestations (lack of sufficient data, 'sparse data problem')

Whereas the issue noted in the quote by Hyland (2002b) above refers particularly to the second problem, that of the detection of false positives when overly relying on surface form, the first problem outlined here is particularly relevant for arriving at a comprehensive history of speech act usage. The third and fourth problem, in turn, although decidedly relevant for research of a historical pragmatic kind, can be regarded as generally inherent issues for the study of linguistic change of any kind.

The problem for historical pragmatic form-to-function approaches is particularly well-formulated by Kohnen (2008b), where the author observes that any form-to-function approach involving corpus-based speech act research will

> "always have to rely on an initial assumption about the formal specification of the speech act under investigation, and, however large the corpora tested, [researchers] cannot exclude the possibility that some other manifestation of the speech act are hidden somewhere in the corpus. Neither can they specify the frequency of these "unknown" manifestations."
>
> (Kohnen, 2008b, p. 295)

As a different approach, Kohnen suggests a more traditional philological methodology which almost likens a function-to-form mapping, by determining "all the relevant forms of a speech act [...] by means of a genre-based micro-analytic bottom-up methodology" (Kohnen, 2008b, p. 296).

As stressed elsewhere, for a detailed and systematic study of speech acts, Kohnen advises to base any analysis of this kind on a study of (diachronically) comparable text

domains (e.g., genres or text types) for which the functional profile remains more or less fixed (Kohnen, 2007, p. 141). The notion of a fixed functional profile relies on the assumption that the communicative context for which a speech act is used is subject to little variation.

In choosing sermons as objects of study in such a genre-based approach, in essence a genre of religious *instruction*, directives offer themselves readily as speech acts for analysis: "We can assume, for instance, that the genres associated with religious instruction require directive speech acts in every period of the history of English. Against this background, realistic comparisons seem to be possible" (Kohnen, 2007, p. 141). Although Rissanen observes that directive speech acts in Early Modern English writing are primarily found in works containing instruction (e.g., cookery books and medical-recipe collections Rissanen, 2000, pp. 277-280), Kohnen thus extend the range of genres in which directives may be attested, as historical corpora such as the *Helsinki Corpus* mainly contains such secular genres of instruction.

Although initial analyses concerning the use of directives involved religious prose exclusively, in later studies this is complemented by cross-checks with other genres and across corpora of English (e.g., the *Helsinki Corpus*, the *London-Lund Corpus*, the *BNC*, and 'dictionary corpora' such as the *Middle English Dictionary*). As such, a fuller and more representative sample of the use of directives in the history of English is achieved. The most comprehensive summaries of results into the use of directive speech acts in the history of English is provided in Kohnen (2007, 2008b), incorporating findings from a number of previous studies on the use of directives in diachrony by his hand (e.g., Kohnen, 2001b, 2002, 2004, 2006, 2007, 2008a, 2008b).

The general picture that emerges from the historical data on sermons is that there is a general decline in the use of readily identifiable directive speech acts in the history of the English. Where the Old English data shows the highest frequency of directives (8.5 per 1,000 words), this rate drops over the course of ME and EModE (respectively, 5.1 for the 15^{th}-century, and 4.8 and 2.8 for the 16^{th}- and 17^{th}-centuries), and sees a slight increase towards the end of the 20^{th}-century again (Kohnen, 2007, p. 153). Thus, sermons become less 'directive' initially, decreasing in the amount of explicit guidelines for the audience to observe, before becoming slightly more 'directive' again in the late twentieth-century (Kohnen, 2007, p. 152). The EModE is said to be pivotal in these developments, with the rise of indirect manifestations and a notable decrease of direct manifestations which already begins in ME. According to Kohnen, this trend reflects the "growing importance of considerations of politeness [...as] English becomes more less explicit, less direct and less face-threatening" in Early Modern English (Kohnen, 2004, p. 246).

In terms of the use of specific manifestations, the decline from OE to ME is largely

due to decline of the first-person imperative construction, e.g., *uton we*. In addition, Kohnen (2007, p. 155) notes a decreased use of modal constructions in the 15^{th}-century compared to the data from the 10^{th} and 11^{th}-centuries. For Early Modern English, the decline continues in all categories of directives, including a significant reduction in the use of constructions using modals. The only exception seems to be an initial rise in the frequency of second-person imperatives in Early Modern English, although these too later seem to follow the general decline seen with other directives. The picture that emerges is that "Early Modern sermons employ fewer directives but also tend to make directives more face-saving" (Kohnen, 2007, p. 155). The higher frequency that is seen in the use of directives at the end of the 20^{th}-century, in turn, is primarily instantiated by a rise in usage of modals, first-person imperatives (*let's* and *let us*) and a higher use of indirect constructions (particularly speaker-based constructions Kohnen, 2007, p. 156). Although the above trends are mainly based on frequencies of directives in sermons in Kohnen (2007), they show striking parallel to the combined data when the genres of private letters and prayers are included (cf. Kohnen, 2008b).

However, Kohnen (2008b) also shows that genres show marked differences in their reliance on certain manifestations of directives. For example, whereas sermons and private letters have relatively similar proportions of directives compared to the large share of directives in prayers (measured as attestations per 1,000 words), preferred manifestations across these genres differ considerably: sermons rely on a mixture of imperatives (1^{st} and 2^{nd}-person) and modals, whereas in prayers performatives and 2^{nd}-person imperatives predominate (the latter overwhelmingly; cf. Kohnen, 2008b, p. 305). Private letters make frequent use of performatives and 2^{nd}-person imperatives as well, but in this genre the use of indirect strategies to express a directive speech act seem to be relatively common (ranging from 14% up to 36% of attestations in the late twentieth-century sample Kohnen, 2008b, pp. 304-305).

Regarding the topic of corpus-based pattern matching searches, indirect directives are generally a problematic category due to their structural diversity. This poses some problems for future automated searches of directives in the genre of private letters, according to Kohnen (2008b, p. 305). Compounding this problem is the fact that indirect directives seem to increase both in frequency and variability in the history of (modern) English (Kohnen, 2008b, p. 309). Another problem next to the issue of detection is the "inferrability" of such indirect directive constructions. That is, determining the extent to which an utterance may still be interpreted as a directive when lacking an imperative, modal or performative verb (Kohnen, 2008b, p. 309). We will turn to this issue in section 4.2.4.

In terms of general retrievability, however, the genre-based, bottom-up approach advocated by Kohnen shows that the majority of manifestations of directives is covered by

the 'common types': imperatives, performatives and modals (Kohnen, 2008b, p. 309). Tracing these particular forms in form-to-function approaches will uncover most manifestations of directive speech acts (cf. Kohnen, 2008b, p. 302, where it is stated that the 'unpredictable manifestations' not covered by these three common types together account for about 10-12% of directives in the three genres studied).

Second-person imperatives turn out to be not only the most prototypical, but also the most prolific, form of such 'common types' of directives in the history of English (excepting the Old English data). Kohnen (2008b, p. 309) remarks that "the evidence of the present data suggests that second-person imperatives may be something like the unmarked manifestation of directives". In addition, whereas other common forms may show genre-specific patterns of occurrence, second-person imperatives occur with comparable high frequencies across genres. Findings such as these seem to lend further support for the perspective of Huddleston (2002) to regard imperatives as the unmarked clause type realisation of directive speech acts in English.

Nevertheless, the rise of second person imperatives in Early Modern English sermons is described as somewhat puzzling, as this category of directives may be assumed to be relatively direct and face-threatening. As was already noted in Kohnen (2004), however, a closer look at the semantic nature of the verbs used reveals a change in the use of such second-person imperatives. Kohnen notes that whereas such manifestations in Middle English usually refer to explicit, real-world acts which an addressee is to carry out in everyday life, in modern English religious prose, second-person imperatives far more often involve 'mental acts' that serve to decode the message or grasp the meaning of the text (cf. Hyland's (2002b) directives with intended cognitive or textual acts Kohnen, 2004, p. 244).[8]

4.2.3 A working taxonomy of directives

Kohnen identifies a wide range of linguistic forms that may serve to express directives in the history of English. Reflective of the assertion that it is difficult to classify directives consistently (cf. Kohnen, 2008b, p. 300), overviews of the respective surface patterns differ slightly in consecutive papers on directives in the history of English (cf. Kohnen, 2002, 2004, 2007, 2008b). The following list, containing four broad classes of directives, is based on Kohnen (2008b):

1. Performatives (with 1^{st}-person singular or plural subject) plus "an object referring to the addressee and the requested act" (Kohnen, 2008b, p. 298)

[8] In terms of politeness, the author notes that it is probably "far less face-threatening to guide people through a text using imperatives than employing imperatives in requests which affect their everyday lives" (Kohnen, 2004, p. 244), which would fit the general picture that religious instruction becomes more polite and less face-threatening over time.

2. Imperatives (broad definition)
    - 1^{st}-person:
        - Periphrastic *Let us/let's*-constructions
        - Subjunctives (usually with inverted word-order, "go we"; cf. 'hortative subjunctive Traugott, 1992, p. 185)
    - 2^{nd}-person: 'regular' imperatives
    - 3^{rd}-person: subjunctives or constructions with *let* plus 3^{rd}-person referent which includes the addressee
3. Modals
    - 'Regular' modal expressions: utterances with modal verbs and other lexical items expressing obligation, etc.
    - 'Other' modal expressions: e.g., impersonals and passives denoting modal meanings such as obligation, possibility, permission
4. Indirect directives
    - Speaker-based declaratives
    - Hearer-based interrogatives
    - Hearer-based conditionals
    - 'Other' indirect manifestations

Although the list of directives derived from Kohnen (2008b) is instructive for an overview of frequently occurring surface forms in the history of English, it needs to be kept in mind that our focus deviates from that of Kohnen: where his primary aim is to offer a comprehensive account of the history of the directive speech act and its (cross-genre) manifestations, our scope is much more restricted (see §4.2.4).

In addition, apart from a few minor issues of visual arrangement, one marked difference between the classification by Kohnen (2008b) and that based on Huddleston (2002) is that whereas the latter gathers all declarative modal expressions among the declarative indirect speech acts with directive illocutionary force, Kohnen generally regards utterances with modals as one of his four main classes of directive speech acts. However, the four sub-classes of indirect directives are partly sub-categorised based on 'point of view' (cf. Kohnen, 2002, p. 166) and one of these, speaker-based declaratives, represents declaratives with first-person pronouns plus verbs that express the attitude or volition of the speaker with regard to the intended act – a definition which does not necessarily exclude modal verbs (cf. Kohnen, 2008b, p. 300, example (8)). Although point of view (or perspective) is no doubt an important feature of directives or requests, we choose to follow Huddleston (2002) in regarding clause type as the main criterion

for a classification of directives and will come back to the issue of perspective in section 4.2.4.

It may also be observed that a classification based on clause type is more in line with a categorisation of directives in an earlier publication by Kohnen, in which two typical examples of modal declaratives are both regarded as indirect directives (i.e., 1^{st}-person pronoun with modal of volition or desire (*I want you to ...*) and utterances containing a 2^{nd}-person pronoun in combination with a modal of obligation or necessity, e.g., *You ought to ...*; cf. Kohnen, 2004, p. 238).

A second benefit of regarding clause type as key in the current classification is that the group of 'other' modal expressions mentioned by Kohnen (2008b), e.g., passives and impersonals which denote obligation, can straightforwardly be grouped with the indirect directive extraposition constructions mentioned by Hyland (2002b), Biber et al. (2004). As a result, the constructions in this new set may be seen as belonging to a category of declaratives with various non-canonical word-order patterns or information-packaging constructions (cf. Ward et al., 2002) which all share (a certain degree of) directive illocutionary force.

Apart from the classificational status of declaratives containing modals, the overviews of Huddleston (2002) and Kohnen (2008b) show a striking degree of overlap. Both authors unequivocally regard imperatives and performative declaratives as having direct illocutionary force, and interrogatives with directive force are part of the set of indirect directives in both taxonomies.[9] The global classification of frequent manifestations of directive illocutionary force used here thus looks as follows:

- Imperatives (direct)
- Interrogatives (indirect)
- Declaratives (direct/indirect)
  - Canonical word-order with direct force (e.g., performative constructions)
  - Canonical word-order with indirect force (e.g., personal pronouns with verbs of desire)
  - Non-canonical word-order with indirect force (e.g., extraposition constructions)

[9]Note, however, that whereas Kohnen (2008b) usually distinguishes between imperatives based on person (first, second or third), the label of imperatives below will be used to refer to second-person imperatives exclusively, unless otherwise stated. The reason for this is that first of all, first-person *let's/let us*-constructions do not feature as a frequent pattern in our corpus. This finding is corroborated by the fact that this type of imperative seems very infrequent in Early Modern English overall (although the genre of Handbooks actually represents the highest proportion in the *Helsinki Corpus*, it only represents 16 cases in mainly interactive, dialogue-type instructions which largely deviates from the nature of the texts in the present corpus; cf. Kohnen, 2008b, p. 307). Second, third-person imperatives are regarded here as involving a very specific type of anaphoric reference, and by extension, illocutionary obligation on the addressee (cf. §4.2.4).

Having outlined a general classification of surface forms that can incorporate frequently occurring realisations of directive speech acts from a synchronic as well as diachronic perspective, in the next section we turn to the discourse function associated with the orientation towards the addressee – and by extension, of personal involvement between discourse interlocutors. One issue mentioned above, the realisation of the addressee as agent of the required action expressed in the directive, will be the main focus in the next section.

4.2.4 Agentivity and interlocutors: Addressee orientation

As a primarily function-to-form mapping, the present approach involves the linguistic realisation of directive speech acts in the current corpus. As mentioned previously, three 'meta-categories' for linguistic realisations will be introduced here as an intermediary, above the level of specific grammatical constructions. Although generally conceptualised as a measure for the degree of personal involvement, the basis for classification in these super-categories is rather grounded in the semantics of agentivity and co-referentiality conferred by the linguistic realisation of the directive.

General outline of the addressee orientation categories

Intuitively, the classes of Direct Address, Indirect Address and Suppressed Addressee refer to the realisation of the addressee. That is, given an author's leeway in addressing an interlocutor or audience using a directive speech act, these labels signify whether the reader is addressed directly ('Direct Address'), whether it is done indirectly ('Indirect Address'), or otherwise via a construction that, whilst being instructional and directive ('imparting upon the interlocutor an obligation to act'), leaves the addressee implicit (covert realisation of the addressee; i.e., 'suppressed addressee'). Suggested surface realisations of these classes may illustrate these addressee orientation categories.

Direct Address, for example, involves utterances that either express a directive speech act using a second-person pronoun referring to the addressee, or via an imperative. Examples of such constructions include both the speaker-based and hearer-based indirect directives mentioned by Kohnen (2004), e.g., *I want you to water the horse* and *You should provide the horse water*, as well as ordinary imperatives, "Give horses plenty of water in summer".[10] In addition, performative constructions usually include a reference to the addressee as the grammatical object, and will in those cases be counted in this class (*I order you to be quiet!*). Although imperatives normally do not realise the

[10]However, not every utterance containing a 2^{nd}-person pronoun referring to the addressee is a directive, and although imperatives generally refer to 2^{nd}-person imperatives here, the status of 1^{st}- and 3^{rd}-person imperatives is somewhat more complex. See the discussion on imperatives below, section 4.2.4.

subject in the surface string, it is generally understood that the addressee in English imperatives is co-referential with the (omitted subject) second-person pronoun 'you' (Biber et al. (1999, p. 219), Huddleston (2002, p. 925), Quirk et al. (1985, p. 828)). In other words, even when omitted, "in imperatives the subject is always referentially tied to the addressee(s)" (Huddleston, 2002, p. 927). Both second-person imperatives with and without a realised subject have equal status in terms of Direct Address status, especially as Rissanen observes that the second-person subject of a directive imperative is more commonly expressed in Early Modern than Present-day English (Rissanen, 2000, p. 277).

Indirect Address consists of surface forms in which the orientation towards the addressee takes the form of a full noun phrase which is, or may be assumed to be, co-referential with the addressee. See for example chapter titles in Baret (1618), e.g., "How the Horseman should governe himselfe and his Horse" (chapter 7) and "A Horseman should be loving and gentle" (chapter 11). In running text, such noun phrases should be readily replaceable by a deictic marker such as a second-person pronoun without altering the propositional content or intended target for carrying out the future act. An example from Markham (1607):

(1) Now for mingling Pease, Beanes, Fitches and wheate branne together, it is a moste unwholsome provender [...], so that I would wish all that love their horses not to love this kinde of food.
(Markham, 1607)

Although example (1) contains a declarative clause with a 1^{st}-person pronoun plus a verb of desire (*I wish ...*), it is nevertheless identified here as an Indirect Address because the phrase "all that love their horses" is co-referentially linked to the intended addressee. That is, the agent of the directive is understood to be interchangeable with a deictic 2^{nd}-person pronoun 'you' as address to the reader, e.g., *I would wish (for) you not to love this kind of food.* Although the utterance in (1) is a declarative, Indirect Address readily occurs in other clause types. Imperative utterances containing 3^{rd}-person subjects, for example, can be gathered among possible surface realisations in the category of Indirect Address (see section 4.2.4 below).

Lastly, the category of "Suppressed Addressee" may contain a variety of surface forms. These include, for example, extrapositions with adverbs of necessity or importance (cf. (2-a)), agentless passives (cf. (2-b)), modals with an inanimate subject (2-c), and verbs of necessity or obligation with generic control and a semantic patient-subject, e.g., (2-d). Note that some (elements of) the surface forms provided here are not mutually exclusive; e.g., an utterance containing an agentless passive containing a semantic-patient subject.

(2) a. It is therefore necessary that such horses as are not regularly worked should receive

daily a moderate proportion of exercise, Kirby (1823)

b. manie thynges are to be considered, as the kyndes and substaunces of meate, and whiche be beste [...] (Blundeville, 1565)

c. food should consist of the two principal constituents which are required to sustain the body; (Fleming, 1884)

d. in general, growing horses need a higher percentage of protein than mature horses. (Duberstein & Johnson, 2009/2012)

Categories of addressee orientation and assumptions of agentivity

Although the previous section may leave the impression that the three categories introduced here are primarily based on linguistic form rather than (speech act) function, the examples provided here are only meant to illustrate some of the most common constructions. More in particular, underlying the classification into Addressee Orientation is the linguistic realisation of the agent of the required act of the directive, which is directly connected to the degree of proximity and personal involvement conveyed by an addressor. With regard to this discourse feature, it is not primarily the reference-to-self by an author that is the focus of personal involvement here, but rather, the level of directness with which the interlocutor is addressed.

Given that the defining feature of a directive is to impart upon an interlocutor an obligation to act, the essence of the Direct Address utterance is that the addressee is overtly identified as the agent of the future action. In the perspective taken here, the category of Direct Address is thus the most personally involved manner to convey such a directive.

The use of Indirect Address realisations marks a more circumscribed manner to refer to the addressee as the agent of the required act. Lacking a deictic marker, the addressee is not overtly identifiable as the agent of the required act. Nonetheless, an agent is linguistically realised and this agent denotes an entity which an interlocutor is able to identify as. In terms of referential interpretation, the use of such an Indirect Address to the reader may be compared somewhat to the pronoun 'one' in being more circumscribed in terms of reference to a specific person or particular individual (cf. Huddleston & Pullum, 2002, pp. 426-427, 1467-8). As with pronoun 'one', and utterances which elicit a non-referential reading of 'you' (for example in *What can you do?*), the referential quality of the co-referential entity in Indirect Address utterances relies to a considerable extent on the interpretation by, as well as resulting self-identification of, the addressee (cf. the use of 3^{rd}-person imperatives).

Usually, these constructions are found in context with a moral appeal included in the directive. In terms of politeness, it may be suggested that such moral appeals mitigate the face threat otherwise conveyed in a more involved manner using a Direct

Address category deictic 'you'. In running texts, these utterances often make reference to examples of good practice; see, e.g., (3-a),(3-b), or conversely, the avoidance of bad practice (cf. (3-c) and (3-d):

(3) a. A conscientious horse keeper will spare neither time nor labour [...] upon the animals placed under his care. (Matheson, 1921)
b. A good horseman constantly watches both the amount and quality of grass available on his paddocks and equates this with the condition of the horses. (Leighton-Hardman, 1977)
c. In fast travelling, every Horseman of common sense, will ease his hack up the hills; (Skeavington, c1840)
d. Nevertheless, no sane person would ever wittingly purchase forage damaged in the manner referred to. (Matheson, 1921)

The Suppressed Addressee category includes constructions that allow an addressor to realise directive utterances in the least personally involved, or most detached, manner. Not only is the addressee not linguistically realised in the surface string in such utterances, but given that there is no semantic agent available, there is also no way to infer an indirect co-referential connection between a discourse entity in the utterance that *is* the agent, and the addressee. The grammatical subject in such utterances may for example be a dummy subject, have a semantic patient-subjects (overriding the 'natural' association between grammatical subject and semantic agent) or contain an agentless passive seen above in example (2). What unifies these utterances, then, is that they are agentless utterances carrying directive illocutionary force.

The following overview provides a brief summary of the three Addressee Orientation categories with respect to agentivity:

Direct Address (DA) The agent of a future action proposed in a directive utterance is the addressee

Indirect Address (IA) The agent of a future action is realised as a full noun-phrase or 3^{rd}-person pronoun which is co-referential with the addressee

Suppressed Addressee (SA) Utterances for which the propositional content involves a required action or obligatory state, but which do not express the agent of this act or the agent to realise the trajectory towards this state ('agentless directives')

Addressee orientation and 'request perspective'

Although the Addressee Orientation categories may share some similarity with the notion of 'point of view' or 'perspective' in for example Blum-Kulka and Olshtain (1989), these concepts are not identical. Blum-Kulka and Olshtain distinguish between four categories of requests based on the perspective chosen by the addressor: 'hearer ori-

ented' (*Could you clean up your room?*), 'speaker oriented' (*Could I use your pen for a moment?*), 'hearer and speaker oriented' (*Can we move on now?*), and 'impersonal' requests (Blum-Kulka & Olshtain, 1989, p. 203). The former of these, speaker- and hearer-oriented perspectives, were already encountered in a summary of Kohnen's findings in section 4.2.2. Both these perspectives as well as the hearer and speaker-oriented category are regarded as Direct Address orientations here: the realisations of these request perspectives express a degree of personal involvement by overtly manifesting the discourse interlocutors (using deictic first- or second-person pronouns), and crucially, they refer to the addressee as the intended (partial) agent of the required act.

Requests that are classified as having 'impersonal request perspective', in contrast, are characterised by Blum-Kulka and Olshtain (1989, p. 203) as employing either neutral agents (e.g., 'one/they/people') or passivisation. The primary reason for the use of such constructions is that they soften the impact of a request's imposition on the addressee (Blum-Kulka & Olshtain, 1989, p. 203). Although there is strong agreement in both the addressee orientation and request perspective schemes that "any avoidance in naming the addressee as the principal performer of the act" (Blum-Kulka & Olshtain, 1989, p. 203) may soften the impact (i.e., impositive force) of a directive, the two example realisations of 'impersonal perspective' fall into different categories in the model proposed here. Whereas the 'neutral agents' realise an agent that may be co-referentially linked to the addressee and are therefore classified as Indirect Address, passivisation is a different case. As a non-canonical word-order strategy, passives can theoretically accommodate any of the three Addressee Orientation types. However, since passive directives prototypically leave the agent of the future act unexpressed, e.g., the agentless passive as seen in (2-b), they are usually associated with the class of Suppressed Addressee.[11]

Incidentally, it may be observed that in the example used to illustrate the category of 'impersonal perspective', i.e., *So it might not be a bad idea to get it cleaned up*, the phrase that Blum-Kulka and Olshtain (1989) designate as realising the 'impersonal perspective' is the extraposed element of an extraposition construction (cf. (2-a)).

Imperatives

With Huddleston (2002) identifying imperatives as the default clause type for directive speech acts, the classification of utterances containing an imperative verb as directives does not usually pose a problem. What is of relevance, however, is to distinguish between imperatives based on their (implied) subjects and the according classification

[11] That is, of course, unless the agent is realised as the complement of a PP in a long passive, e.g., *In spring, young horses are fed by you / the care taker of the horse*, in which case an utterance would be analysed as DA or IA, respectively.

in the respective Addressee Orientation categories.

The default or ordinary imperative with or without 2^{nd}-person subject is the most prototypical case of Direct Address. However, imperatives with 1^{st}-person subjects are readily attested in the history of English (cf. Kohnen, 2008b), particularly *uton*-constructions in Old English and the use of *Let's/Let us*, which is a common construction in Middle English and thereafter (Kohnen, 2007, p. 145).[12] Both cf., Kohnen (2004, p. 240) as well as cf., Huddleston (2002, p. 924) distinguish ordinary imperatives, including *let* used as a full verb, from *let*-imperatives or 'periphrastic' imperatives.

Such first-person inclusive *let*-imperatives, in which the meaning of 'allow' has been bleached (Huddleston, 2002, p. 924), are relevant in the context of Addressee Orientation. Since the reference of such imperatives includes both the writer as well as the addressee (cf. workshop-we in §3.2.1), it is regarded as remarkably personally involved and should therefore be counted as Direct Address (primarily because the addressee is not linked indirectly via inferrable co-reference such as in pronoun 'one' or non-referential 'you', but is directly linked by way of deixis). Only one case of the periphrastic 1^{st}-person inclusive imperative was found in our corpus (incidentally already encountered in section 3.2.1):

(4) Hetherto we have talked onelye of the horses meate, now lette us speake somwhat of his dryncke. (Blundeville, 1565)

As such, (4) is the only case of an imperative in the Direct Address category without a 2^{nd}-person subject. The only other attestation in our corpus that contains a use of full verb *let* not with a second-person but with a 3^{rd}-person imperative subject, to which we turn presently, is found in the sentence in (5):

(5) Let no man expect great performance, unless his horse be full of hard meat, and in CONDITION. (Skeavington, c1840)

This utterance could be interpreted as a case of a so-called "open *let*-imperative" (cf. Huddleston, 2002, pp. 925, 937). Given that the status of this 'open type' as a syntactically distinct construction is debatable, it is generally gathered among the ordinary imperative use of the full verb here.

In the current corpus, most instances of *let* thus occur as full verb, either in imperative form (6-a) or in (non-)finite declaratives (6-b). In fact, most instances of the use of the full verb *let* occur in 2^{nd}-person imperatives, either with the sense of 'allow' (e.g., *let him eat/stand/drink/rest*) or indeed with the specific sense of 'bloodletting' (6-c).

(6) a. [...] toss up his Litter and let him rest till Morning. (Speed, 1697)

[12] No cases of the (1^{st}-person) subjunctive were attested in our corpus.

b. Their main problem is usually how to maintain the horse's condition without letting him get too fresh. (Leighton-Hardman, 1977)

c. [...] and afterward let him bloud accordingly";[13] (Clifford, 1585)

3^{rd}-person imperative subjects, such as (5), were already associated with the category of Indirect Address above. The primary reason for generally including such imperative forms in the category of Indirect Address is that they establish a co-referential link between the addressee and the agent of the directive future act implied in the imperative. As Kohnen (2008b) notes, "[s]uch constructions can be classified as directives if the referent of the third-person subject includes the addressee. This is the case, for example, with general expressions referring to humankind" (Kohnen, 2008b, p. 299).

According to Huddleston, the range of possible subjects for 3^{rd}-person imperatives is relatively restricted, and usually needs to have personal denotation (Huddleston, 2002, p. 926). However, an important distinction should be made with respect to the status of the antecedent in directives with a third-person subject. Although in declaratives one may fairly straightforwardly distinguish between (optional) vocatives and (obligatory) subjects, e.g., *(John$_{[voc]}$,) I$_{[sub]}$ am off!*, this is not the case for all directive utterances. Particularly in imperatives, it can be difficult to distinguish between third-person subjects and vocative noun phrases, e.g., *The front row(,)$_{[voc/sub]}$ start queuing* (Quirk et al., 1985, p. 829).

Huddleston (2002) offers a useful insight in that the referential status of subjects and vocatives in imperatives usually seems quite distinct: whereas there is no direct relationship between the subject-referent and the addressee in declaratives, the subject of Present-Day English imperatives "is always referentially tied to the addressee" (Huddleston, 2002, p. 927). In utterances such as *Somebody open the door*, there is no other way to understand the 'somebody' than 'somebody *among you*'.

(7) a. Somebody at the front(,) write your name on the board. [voc. or subj.]

b. Somebody at the front write their name on the board. [subj.]

As Huddleston (2002, p. 927) argues, the possessive pronoun 'your' in (7-a) is not anaphoric but deictic, involving not the initial noun phrase but a (generic) addressee in the discourse context. In contrast, the possessive determiner 'their' in (7-b) is anaphoric, indicating that the initial NP should be read as (a third-person imperative) subject.

Since in our corpus we encounter Indirect Address mainly in declaratives, the distinction between vocative NPs and 3^{rd}-person subjects is not of vital importance here.[14]

[13] In the discourse context, the regular sense of 'let him bleed (out) accordingly' seems less likely than the reading, 'bloodlet him accordingly'.

[14] A complicating factor in this respect may be the practice of using 2^{nd}-person possessives as a

What is of relevance, however, is the issue of referentiality of 3^{rd}-person subjects for a correct interpretation of Indirect Address utterances and their likely antecedents. Huddleston notes that in case the subject of a declarative refers to the addressee(s), it does so "only coincidentally" (Huddleston, 2002, p. 927). This generally applies to most of our Indirect Address cases. Given that we are dealing with utterances with some degree of directive force, this force will then extend to utterances in which the reader can identify as the intended semantic agent of the action proposed in the directive (and as encoded by the subject-referent NP).

4.3 Method

Using the operationalisation of utterances described in section 3.2, i.e., (parts of) sentences between punctuation markers '.', ';', ':', '!' and '?', we determined the addressee orientation for every utterance in the corpus. That is, whether the main discourse purpose is either instructional or non-instructional (e.g., anecdotal, descriptive, explanatory). For each instructional utterance, it was then established how the obligation on the addressee is grammatically encoded: whether a writer addresses an expected interlocutor/audience directly (overt agent, Direct Address), whether this is done indirectly (addressee is not realised as the agent; Indirect Address), or via a construction that, whilst being instructional and directive ('imparting upon the interlocutor an obligation to act'), leaves the addressee implicit (covert realisation of the addressee; i.e., Suppressed Addressee). Non-instructional utterances are generally tagged as 'Non-instructional', and additional text elements that are not part of the main body of the sample text (titles and (sub-)headers, etc.) are tagged as 'Non-text/NA'.

Although the latter category of utterances is of little relevance for the current purpose, it is provided here as somewhat of a control, providing the number of utterances that are not tagged as running text in the present corpus. As these utterances contribute towards a sample's total word length they were included in the current procedure, although no further linguistic analysis was carried out on these utterances. Nevertheless, in the event of high frequencies for this category, both absolute as well as in proportional terms, a closer look at particularly the lay-out features of a text seemed warranted.

In addition, it must be kept in mind that the category containing Non-instructional utterances is a bin-category at a different level of granularity than the subdivision of directives in the respective Addressee Orientation classes. It is therefore only with great caution that the figures in this class should be compared to those for other classes.

---

general determiner in cases of non-possession, as was flagged up in the discussion of example (2-b) in section 3.2.1. Such cases did not show up in utterances identified here as Indirect Address, however.

One prominent issue is how to determine whether an utterance is a directive if there are few (verbal) markers of directive force (cf. the issue of "inferrability" of Kohnen, 2008b, as noted above). This is particularly salient for categorising utterances in the Suppressed Addressee category, as these utterances are usually agentless declarative clauses, and are therefore *indirect* directives only. Particularly for examples such as (2-c) and (2-d) above, it can be questioned to what extent such utterances should be interpreted as primarily descriptive (statements) or instructional (directives). It may be noted that the utterances provided in (2) seem to vary in terms of strength of the 'force of illocution', suggesting that directive force may be a matter of degrees.

In practice, it is often difficult to assess to what extent we are dealing with a declarative with statement force (e.g., a descriptive utterance) or a declarative with directive force. For example, is *Hay is necessarily fed in winter* in (8-a) a directive or a non-directive (descriptive) utterance? Although it seems rather descriptive, i.e., a declarative with statement force, there is a clear indication of necessity conveyed by the adverb *necessarily*. Although *Hay is necessarily fed in winter* is a declarative clause, it seems to impart an obligation on the care taker of the horse to make sure that the suggested procedure is observed.[15] This sense of obligation does not seem to be apparent in the bare *Hay is fed in winter* statement (unless it is provided by the surrounding discourse context). On the other hand, with a case like (8-c) the directive illocutionary force is clear, due to both the modal of obligation as well as the expression of necessity by way of an adverb. Intuitively, the declarative utterance in (8-c), with a modal but without an additional adverb marking necessity, is situated in between (8-a) and (8-b). The difference between such 'activating' declaratives and a descriptive declarative such as (8-d) thus seems a matter of degrees rather than a strict dichotomy.

(8) a. Hay is necessarily fed in winter.
b. Hay should be fed in winter.
c. Hay should necessarily be fed in winter.
d. Hay is fed in winter.

According to Huddleston (2002), indirect speech acts may vary in the degree to which they express indirect illocutionary force.[16] This degree is based on for example

[15] In case the context elicits the more descriptive reading that hay must be fed in winter because there is no access to it in other seasons. It seems that this descriptive reading is more readily encountered when the global discourse topic of a section is 'hay' (i.e., spelling out the different circumstances and conditions in the use of this type of fodder), whereas the directive reading is more likely when the global discourse topic is concerned with feeding (i.e., spelling out to the reader which type of fodder is best fed in which season). It is likely that this interpretation involves discourse-sensitive information-structural concepts such as *topic-comment* and focus which, unfortunately, cannot be discussed in this section in detail).

[16] Although probably superfluous, it may help to stress that the 'indirect' in our label 'Indirect Address' does not refer to indirect illocutionary force as used by for example Searle (1975) and Huddleston

the distance between the propositional content expressed (by the utterance as well as its clause type) and the propositional content intended (by the speaker; cf. Huddleston, 2002, p. 862). Additional non-propositional markers of indirect force may also add to the degree of indirect force conveyed (cf. Huddleston, 2002, p. 864).

If we allow for the fact that utterances can differ in degrees of indirectness, the declarative utterances in example (8) can be ranked or classified accordingly. It does not seem far-fetched to state that the stacking of such markers adds to the degree of indirectness conveyed, e.g., the difference between (8-a) and (8-c) seems clear. However, the difference between the marking of directiveness by way of an obligation modal (8-b) versus a necessity adverb (8-a) is more complex.[17] In addition, as the degree of illocutionary force may also depend on the local discourse context, hard and fast rules for the ranking of the various individual markers of indirect 'directiveness' (e.g., modals, necessity/obligation adverbs, clause type) do not seem available.

Noting that it is not our aim here to distinguish between degrees of indirectness for particular indirect directive speech acts, we leave this issue for the moment. What matters most for our purposes instead is to be able to distinguish between declaratives such as (8-a) as and (8-d). All other things being equal (e.g., information conveyed by the discourse context), the former will be classified as carrying directive force while the latter will carry statement force due to the availability of a marker of indirect directive force. Thus, if an utterance is interpreted as having a degree of directive illocutionary force, either through direct (e.g., imperative, performative verbs with directive function) or indirect markers of indirectness (e.g., adverbs and modals expressing necessity or obligation), it will be classed here as a directive utterance. Further classification into one of the three Addressee Orientation types will then ensue.

4.4 Results

4.4.1 Distribution of addressee orientation types in the corpus

As a first step in determining the level of personal involvement of the instructive manuals in the corpus, table 4.1 provides the absolute frequencies of utterances across the various Addressee Orientation categories, as well as utterance total and mean utterance length in words. Although these raw frequencies do not prove particularly insightful for a direct interpretation of a text's use of directive utterances, the last column containing category totals may be used to gauge the use of respective AO categories in the

(2002). Rather, it refers to an indirect orientation towards the addressee in the surface realisation.

[17] The question as to the extent to which such adverbs of obligation or necessity should be regarded as part of the propositional content or rather as non-propositional markers of indirect illocutionary force (Huddleston, 2002, p. 864), is beyond our scope here.

corpus. As is to be expected, a large share of utterances are categorised as belonging to the Non-instructional class, comprising 1,251 out of a total of 2,167 utterances (or 57.73%). As a class containing a broad variety of clause types and utterances without directive force, this is a particularly heterogenous cluster. For example, and as seen in the examples provided above, anecdotal (cf. (4-d)) or descriptive ((3-d)) utterances are well-represented in the corpus and may feature readily in this class. Such utterances, as well as any other utterance for which it could not be established with absolute certainty that there was a degree of directive illocutionary force involved, were added to this class. In general, erring on the side of caution seemed a better strategy than being too permissive with respect to the inclusion of utterances in the set of directives.

The total number of utterances categorised as Non-text seem particularly affected by the contributions of Speed (1697) and Davies (2009). Together, these samples account for nearly half of the total for this category (41.45%, or 34 out of a total of 82 utterances). An inspection of text samples clearly shows that, in comparison to the other samples selected here, these particular authors provide frequent sub-headers for paragraphs and other text fragments. This suggests a high degree of visual discourse structuring by means of textual lay-out features.

Table 4.1: Distribution of Addressee Orientation categories in the corpus (n utterances per text)

Source	DA	IA	SA	Non-instr.	Non-text	Utt. (total n)	Utt. length (words)
1565 Blundeville	14	3	34	73	3	127	33.99
1585 Clifford	37	2	4	39	7	89	52.96
1607 Markham	47	1	12	36	4	100	45.31
1618 Baret	32	5	9	53	3	102	43.92
1697 Speed	112	2	10	19	18	161	29.74
1721 Gibson	1	7	33	151	3	195	26.70
1796 Hunter	39	10	26	50	5	130	35.28
1823 Kirby	3	3	55	135	4	200	23.58
1840 Skeavington	24	19	42	97	5	187	24.87
1886 Fleming	0	2	51	134	6	193	23.74
1921 Matheson	2	15	48	123	5	193	24.57
1977 Leighton	13	6	69	126	3	217	21.60
2009 Davies	1	0	41	215	16	273	16.13
Cat. total	325	75	434	1251	82	2167	30.95

Taken together, the three classes of directive utterances account for roughly two-fifths of the data (325 + 75 + 434 = 834, or a combined total of 38.49%). Within this category, it is particularly the Direct Address (n= 325) and Suppressed Addressee (n= 434) categories which are well-represented; respectively 40.94 and 54.68% of total directive utterances. The Indirect Address category, in turn, does not appear as a very frequent linguistic realisation for conveying to the addressee that a certain future act

is desirable. With a total of 75, or 9.45% of total utterances, it is the least frequent category overall. Nevertheless, all authors seem to employ this strategy at least once or twice (with exception for Davies, 2009), and it is thus not particularly characteristic of certain authors or time periods. Nevertheless, the samples by Hunter (1796), Matheson (1921) and particularly Skeavington (c1840) seem to employ this type of Addressee Orientation comparatively frequently (n= 10, 15 and 19, respectively; but see below for a discussion of these figures as within-sample percentages).

Given the specific nature of the text domain, and in light of a selection based on certain discourse topics, the number of directive utterances in the corpus might not seem particularly high. However, it serves as a useful reminder that not all utterances in texts drawn from manuals are necessarily instructive or procedural *per se*. Figures such as these, with high rates for Non-instructive utterances in texts explicitly written as handbooks or manuals, indeed suggest that the 'how' often necessarily goes hand in hand with the 'why' in instructive writing. An exception to this general rule may be found in the text by Speed. The majority of utterances in this text are indeed categorised as Direct Address (112 out of 161, or 69.57%), which dovetails the fact that a large share of our sample's selected discourse topics occur in procedural or recipe-style sections in this particular text.

Table 4.2 provides an overview of the percentage of utterances per text as distributed across the Addressee Orientation categories. These percentages shed more light on the proportional use of utterance categories within individual texts, and form the basis for figure 4.1. From this figure, a general picture emerges of trends in the reliance on the various utterance types in the present corpus.[18]

As was already seen before in the description of absolute frequencies, Non-instructional utterances take up the largest share of utterance in the current corpus. This is followed by the category of Direct Address, which seems particularly well-represented in the Early Modern English period. Some notable deviations of this pattern may be mentioned, however, for example the manual by Blundeville (1565) in the first half of the graph, and, although rather late, Gibson (1721).

Despite the fact that a considerable amount of utterances (26.77%) are categorised as Suppressed Addressee, the majority of utterances is non-directive (cf. 57.48% in the Non-instruction category). As it turns out, anecdotal or apocryphal references to classical authorities in the scholastic vein, e.g., *as Russius/Camerarius/Vegetius says, ...*, seem to be particularly frequent in such utterances that are tagged as Non-instructional

[18]Given the current sampling method and number of data points, the above figures naturally come with a caveat. It needs to be recalled that the current sampling of texts was based on consistency of discourse content, and not with linguistic comparability in mind, nor the aim of ensuring external validity. Any suggested trends or patterns in the data are thus to be taken lightly in terms of inferences regarding diachronic developments.

Table 4.2: Distribution of Addressee Orientation categories in the corpus (% of utterances per text)

Source	DA	IA	SA	Non-instr.	Non-text
1565 Blundeville	11.02	2.36	26.77	57.48	2.36
1585 Clifford	41.57	2.25	4.49	43.82	7.87
1607 Markham	47.00	1.00	12.00	36.00	4.00
1618 Baret	31.37	4.90	8.82	51.96	2.94
1697 Speed	69.57	1.24	6.21	11.80	11.18
1721 Gibson	0.51	3.59	16.92	77.44	1.54
1796 Hunter	30.00	7.69	20.00	38.46	3.85
1823 Kirby	1.50	1.50	27.50	67.50	2.00
1840 Skeavington	12.83	10.16	22.46	51.87	2.67
1886 Fleming	0	1.03	26.42	69.43	3.11
1921 Matheson	1.04	7.77	24.87	63.73	2.59
1977 Leighton	5.99	2.76	31.80	58.06	1.38
2009 Davies	0.37	0	15.02	78.75	5.86

in this particular text. In the case of Gibson (1721), the near absence of Direct Address (1.54%, or 3 cases) is counterbalanced by an exceptionally high reliance on utterances categorised as Non-instructional. In fact, for texts written roughly between the mid-17^{th}- and mid-19^{th}-centuries (i.e., Speed, Gibson, Hunter, Kirby and Fleming), the figure suggests an almost complementary distribution between the Direct Address and Non-instructional utterance categories. In later texts, as of roughly the start of the 19^{th}-century, this distributional effect seems to weaken somewhat. Based on figure 4.1, one possible cause for this might be found in the use of the Suppressed Addressee category, which seems to gradually increase in usage. Particularly over the course of the Late Modern English period, the use of this category of directives is considerable, representing by-and-large within the range of 20-30% of utterances for respective authors. This use indeed seems to come particularly at the cost of forms which realise directives in a manner categorised here as Direct Address, as numbers for this class are seen to peter out (although Hunter, 1796, makes considerable use of this strategy still, with 30% of utterances of this type).

The last text, by Davies (2009), seems to not conform to the general pattern seen for the Suppressed Addressee category. In general, this text sample employs a fairly low use of directives, but the drop for the use of Suppressed Addressee directives (to 15.02%, on a total of 273 utterances) is particularly noticeable. Although it is primarily the category of Non-instructional utterances that in terms of frequency seems to benefit from the non-directive nature of the text by Davies, this text also shows a comparatively higher number of Non-text utterances compared to its contemporaries (i.e., 5.86%, versus 2.59% and 1.38% for the immediately preceding texts by Matheson and Leighton-Hardman, respectively). It may be noted that the other two authors who seem to

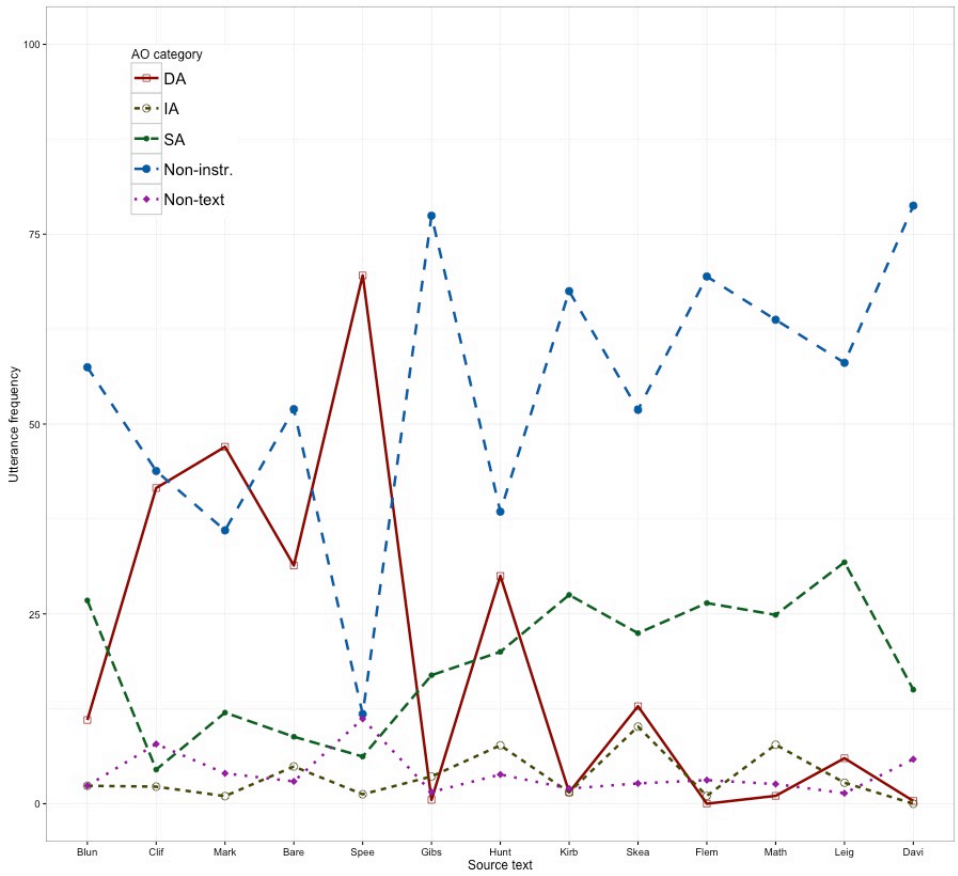


Figure 4.1: Distribution of Addressee Orientation categories in the corpus (as % of sample utterance total)

employ Non-text utterances to a considerable degree (see the data for the samples by Speed (1697) and Clifford (1585)), also seem to deviate from the general upward trend seen in the use of Suppressed Addressee directives as of roughly the late 16^{th}-century (with the text by Clifford, 1585, itself being the lowest point from which this slope takes off). However tentative, what this would suggest is that a frequent use of (sub-)headers is antagonistic, at least in terms of style of writing, type-setting and discourse organisation more broadly, to the use of directives with Suppressed Addressee orientation (cf. the text by Clifford, for example).

Although the line indicating the use of the Indirect Address category is fairly marginal throughout, three somewhat higher values can be detected for the authors already highlighted above: Hunter, Skeavington and Matheson. For both Skeavington and Matheson, the use of Indirect Addressee orientation seems to come at the expense not of other linguistic strategies to realise a directive; that is, the Direct or Indirect Address categories. As the percentages for these latter two categories remain relatively stable,

or are seen to even increase in the case of Skeavington), the use of Indirect Address seems to come rather at the cost of Non-instructional utterances, as is evidenced by minor troughs for this category with these respective authors. Partly due the heterogenous nature of the Non-instructional category, it is difficult to provide a tentative reason for why an increased use of directives with Indirect Addressee orientation would go together with an decrease in non-instructive utterances (cf. Hunter, 1796). A more thorough investigation of the relationship between comparatively frequent use of Indirect Address and lower rates for Non-instructional utterances seems warranted in light of a stylistic interpretation of these findings. Given the examples found for the use of Indirect Address in our corpus, such utterances often seem to make a strong moral appeal to the reader, which might be interpreted as a severely face-threatening act. Why such appeals tally with lower rates of Non-instructive utterances is somewhat puzzling, and it would be relevant to know whether such Indirect Address are in complimentary distribution with a particular type of utterance in the Non-instructive class. One possibility may be that texts with particularly low rates of Indirect Address might realise such moral appeals, when appropriate, using utterances that are not directly analysable as directives, for example using best practice-anecdotal declaratives. Additional scrutiny of such potentially competing realisations for moral appeals seems warranted, although the present coding of utterances is unfortunately ill-suited for such further study.

What we can investigate using the current coding scheme, however, is how the use of these Addressee Orientation categories informs the profiles of particular texts in terms of their general directiveness. Figure 4.2 provides the same percentages as those in table 4.2, specified for the three Addressee Orientation categories proper. The height of the bar chart reflects, for every sample, the cumulative percentage of utterances that are categorised as having a directive speech act function. Within each bar, different shades reflect the various Addressee Orientation types.

Corroborating trends noted in the previous figure, what is immediately striking in the current chart in figure 4.2 is the considerable proportion of directives which involve Direct Address in Early Modern English texts. However, this category does not seem to be used with great frequency in later texts. Rather, what predominates in Late Modern English texts is the category of Suppressed Addressee, making up the majority of directives used in these texts (cf. the lightest bars in figure 4.2). Although it is not the case that Suppressed Addressee is unused in the early manuals, it makes up only a small share of the total percentage of directives in Early Modern English texts (Blundeville being an exception here, with twice the amount of SA utterances compared to DA utterances; cf. table 4.2).

Although in Late Modern English texts the use of SA seems to come at the cost of

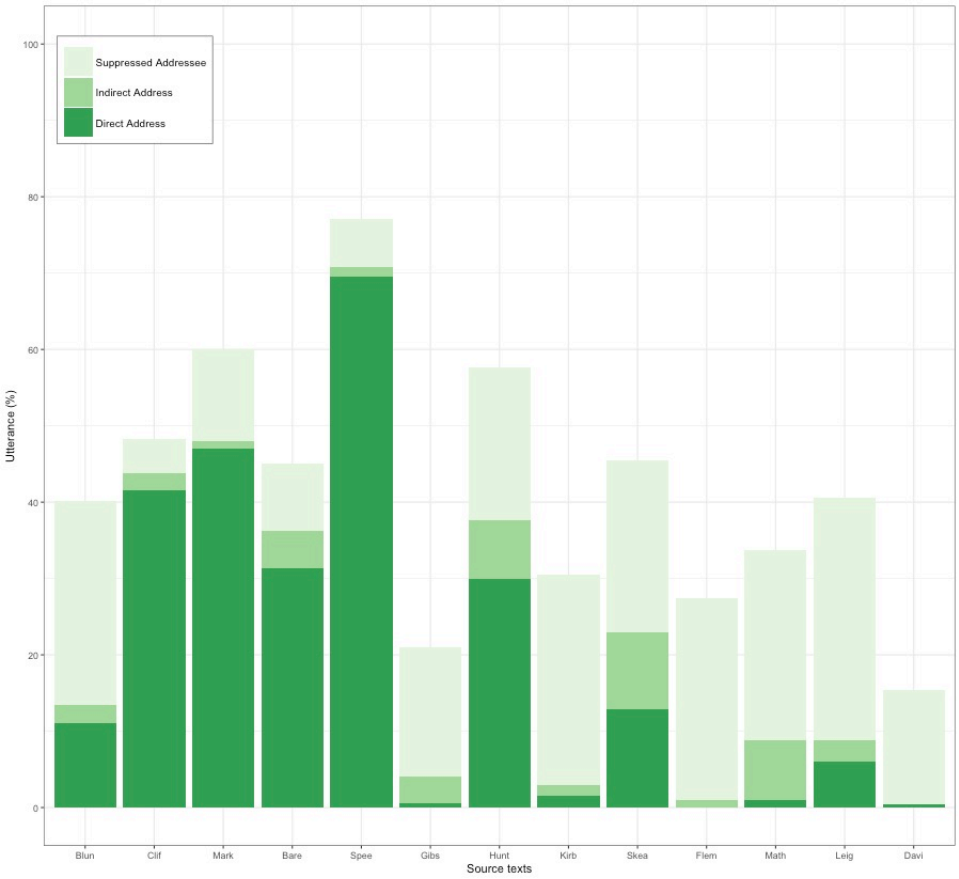


Figure 4.2: Distribution of directive Addressee Orientation categories in the corpus (as % of sample utterance total)

DA, this cannot be the complete picture: not only is the percentage of DA utterances relatively marginal in later texts, but the bars generally appear lower as well. Here, in turn, the sample by Hunter does not conform to this pattern; with 30% of utterances directly involving discourse interlocutors versus 20% of total utterances having a Suppressed Addressee orientation. The general picture that seems to emerge from this chart is that the earlier manuals make more use of Direct Address directives overall, and also seem to employ more directives as a share of the number of utterances in the sample (i.e., higher bars in general).

What the current figure adds to an interpretation of the use of Indirect Address as seen in 4.2 is how these figures tally with an overall profile of a manual's instructiveness. Texts that employ the Indirect Address category, particularly those in Late Modern English (e.g., Hunter and Skeavington) also seem to be more instructive/directive than their contemporaries generally (cf. the height of the bar chart). Although this observation does not seem to hold particularly for the early 20^{th}-century-text by Matheson,

and may therefore be due to idiosyncratic variation, it nevertheless seems to be the case that both Hunter as well as Skeavington seem not only to employ a certain degree of Indirect Address directives (7.69 and 7.77%, respectively), but also appear as those texts in the later period which still use Direct Address to realise directives to considerable extent (that is, respectively 30% and 12.83%).

Since both figures 4.2 and 4.1 are based on proportional frequencies, not merely absolute frequencies across categories in these texts, it is all the more striking that higher usage of Indirect Address seems to tally with a more frequent use of directives in general. Based on these figures, it seems that EModE manuals appear as more directive/instructive than those texts written after the beginning of the of Late Modern English period. In addition, these Early Modern English texts also express directives in a more personally involved manner – that is, assuming that SA may indeed be taken as an indirect way to express personal involvement. Later texts seem to avoid reference to either the reader or writer in the realisation of directives, as evidenced by a general absence of Direct Address.

The data underlying these figures thus suggest that Early Modern English equine manuals are more personally involved in general. However, one issue is how to interpret the degree of personal involvement seen in directives in relation to the degree of proximity conveyed in the remaining, non-instructive portions of the sample. This will be our main focus in the next section.

4.4.2 Comparing personal involvement in directive and non-directive utterances

Borrowing from the previous form-to-function approach (chapter 3) the occurrence of 1^{st}-person and 2^{nd}-person pronouns, we can compare the use of such markers of personal involvement in both directive as well as in non-directive utterances. Given that the use of 1^{st}- and 2^{nd}-person pronouns is taken to be a marker for proximity between writer and addressee, it may be possible to assess whether the use of such markers of personal involvement are distributed equally across directive and non-directive utterances. If this is indeed the case, it may indicate that texts that are high in (personal) directives in general can also be characterised as being more personally involved. On the other hand, if there is a difference in the use of such markers across directive and non-directive utterances, texts that appear as high in (personally-involved) directives need not necessarily appear as more personally involved in general. If the latter case holds, this may indicate that the degree of personal involvement conveyed by the realisation of a directive utterance is not a good proxy for the overall style of a text or text sample (in terms of the proximity between writer and reader).

Table 4.3 provides figures for the number of utterances that contain one or more

surface forms of pronouns, either referring to the writer or addressee, across directive and non-directive utterances in the corpus.[19] The columns in table 4.3 are derived by collapsing the data in table 4.1 in such a way that all AO categories indicating directives (i.e., DA, IA and SA) are grouped in one column, and non-instructional as well as Non-text utterances are grouped in another.

Table 4.3: Distribution of 1^{st}- (sg. & pl.) and 2^{nd}-person pronouns in directive and non-directive utterances (n)

	Directive	Non-directive
Contains Pronoun(s)	263	149
Pronounless	571	1184
Utterance total	834	1333

The frequencies in table 4.3 indicate that the distribution of pronoun-containing utterances is different for directive utterances than for non-directive utterances. Whereas for every utterance containing a pronoun there are two pronounless utterances in utterances containing a directive ($263/571 \approx 0.46$), the ratio of utterances containing a pronoun compared to pronounless utterances in non-directives is roughly 1:10 ($149/1184 \approx 0.13$). Expressed in percentages, pronoun-containing utterances represent 31.5% of directive utterances, whereas this figure lies at 11.2% for non-directive utterances. Proportionally, there are thus more pronoun-containing directives than pronoun-containing non-directives.[20]

A chi-squared test on this table shows that there is indeed an association between the occurrence of 1^{st}- and 2^{nd}-person pronouns and whether or not an utterance can be considered as directive ($\chi^2 = 138.07$, df=1, $p < 0.001$, with ϕ effect size $= 0.2524203$).[21] In other words, 1^{st}- and 2^{nd}-person pronouns are not uniformly distributed across directive and non-directive utterances in the sample set. χ^2 residuals show that utter-

[19] Utterances are counted once, even when they contain more than one pronoun, and irrespective of whether one or more referents for personal deixis are used (e.g., *I, you, we*).

[20] Note, however, that we cannot claim that pronouns are generally found more often in directive utterances than in non-directive utterances, since utterances are counted once even if they contain more than one pronoun.

[21] We take note of the fact that the use of the chi-squared test in corpus linguistics is far from uncontroversial (cf. Baroni & Evert, 2009; Bestgen, 2014; Gries, 2010; Lijffijt et al., 2016). For example, the assumption that all data points are independent of each other, e.g., the use of one AO category in a certain utterance not affecting the category used for the consecutive utterance, is an important methodological requirement for the use of this test. Particularly as data points for one sample are provided by one and the same author and may therefore be assumed to be related, the current data may violate this assumption (cf. Baroni & Evert, 2009). Although we are aware of such issues, more advanced statistical procedures that do justice to the design must be left for further research. The other requirement for χ^2-tests, that +/- 80% of expected frequencies in cells need to contain a minimum of 5 counts (cf. Gries, 2014), seems to be met in the current data.

ances containing pronouns are preferred in directives (+8.293702) and dispreferred in non-directives (-6.560189), and conversely, pronounless utterances prefer non-directive environments (+3.178531) over directive environments (-4.018449). This suggests that the degree of personal involvement as indicated by 1^{st}- and 2^{nd}-person pronouns is unlikely to be found outside directive contexts in instructive writing. In addition, text samples that do not contain a great quantity of directives will be unlikely to be personally involved elsewhere (i.e., in non-directive sections of text). As we have data on the specific categories of directives, however, it is possible to investigate whether there are specific Addressee Orientation categories that contribute to this effect in particular (see table 4.4).

Table 4.4: Distribution of 1^{st}- (sg. & pl.) and 2^{nd}-person pronouns in utterance categories in the corpus (n)

	DA	IA	SA	Non-instr.	Non-text
Contains pronoun(s)	220	10	33	147	2
Pronounless	105	65	401	1104	80
Utterance total	325	75	434	1251	82

Table 4.5: Distribution of 1^{st}- (sg. & pl.) and 2^{nd}-person pronouns in utterance categories in the corpus (*residuals*)

	DA	IA	SA	Non-instr.	Non-text
Contains pronoun(s)	20.1266615	-1.1279582	-5.4508553	-5.8905708	-3.4419163
Pronounless	-9.7517327	0.5465162	2.6410383	2.8540885	1.6676709

The chi-squared statistic shows that there is an association between the occurrence of 1^{st}- and 2^{nd}-person pronouns and the Addressee Orientation utterance categories (χ^2 = 595.91, df=4, p < 0.001, and Cramer's V effect size = 0.5243975). Judging by these residuals (cf. table 4.5), there is indeed a difference between the categories of directives. In fact, the only AO category that has a positive residual for pronoun use is the DA category (residual ≈ 20.13), which indicates that pronouns are preferred in directives with a Direct Address orientation. All other categories have negative residuals for the row that indicates utterances containing first- or second-person pronouns. In the same vein, positive residuals for the row labeled 'Pronounless' indicate that pronouns are dispreferred in all other contexts (except for the negative residual for the Direct Address).

These figures suggest that particularly the directive utterances that are tagged as

Direct Address add to the degree of personal involvement. Based on the data in tables 4.3 and 4.4, we can therefore conclude that it is particularly texts with high degrees of DA category directives (cf. the stacked bar chart in figure 4.2) that are most personally involved, and primarily by way of such 'involved' directives. Text samples with large amounts of all other utterance categories (cf. table 4.2), conversely, will cause a text to appear as more detached.

Three additional features may be noted in reference to table 4.4 containing absolute frequencies across the five AO categories, however. First of all, compared to the number of Direct Address utterances ($n = 220$), there is a fair number of non-instructional utterances that contain pronouns ($n = 147$). Naturally, these figures dwindle in comparison to the number of pronounless non-instructional utterances ($n= 1104$), but it is a reminder that not all utterances that contain pronouns are directives. Second, the cell indicating pronounless Direct Address utterances ($n = 105$) may be noted. This figure primarily represents subjectless 2^{nd}-person imperatives. Nevertheless, it must be borne in mind that in our tagging scheme, 2^{nd}-person imperatives may also count towards frequencies in the pronoun-containing cell if they contain 1^{st}- and 2^{nd}-person pronoun forms. This seems a prevalent pattern in practice, particularly with utterances containing second-person possessives, e.g., "wash his Mouth if he Foam, and gently cherish him with your Hand and Voice" (Speed, 1697). Although both forms of Direct Address thus seem to signal a proximal or interpersonal relationship between writer and reader, it is to be kept in mind therefore that the difference between pronounless and pronoun-containing frequencies is not a one-to-one mapping of subjectless imperatives versus all other forms of pronoun-containing Direct Address utterances.

Lastly, and in connection to the previous point, it may appear as somewhat of a truism that we find pronouns to be particularly relevant in utterances identified as Direct Address, since the classification of utterances in this category seemed to partly depend on exactly this criterion (cf. section 4.3). It is recalled here, however, that we first classified utterances according to whether it could be analysed as a directive speech act utterance (irrespective of the availability of pronouns), and only classified such directives according to other criteria in a second step. Whether or not a directive contained 1^{st}- and 2^{nd}-person pronouns was thus only a second-order criterion. In addition, table 4.4 also illustrates that there are a number of directive utterances in the IA and SA categories which do contain 1^{st}- and 2^{nd}-person pronouns (10 and 33 utterances, respectively). These figures represent directives which contain 1^{st}- and 2^{nd}-person pronouns, but importantly, the agent of the future act proposed in the directive is not realised using such a pronoun. Thus, we underscore that it is primarily the criteria of directive speech act and its agent realisation on which our classification of utterances is based.

4.5 Chapter summary

In the present chapter, directives are explored for their ability to gauge how a writer expresses proximity towards the reader. As personal involvement can be regarded as contributing to the stylistic character of a text, this chapter has tried to assess to what extent the discourse feature of personal involvement can be leveraged using directive speech acts, with a crucial role for the linguistic realisation of the agent of a desirable future act. Our current findings show that utterances that address the reader directly are particularly characteristic of the early texts in our corpus. On the other hand, suppressing a reference to the addressee in directive utterances is particularly frequent in texts in the Late Modern English period. The use of a third category, in which the reader is addressed indirectly, may be an infrequent and marked method to engage with the audience. In addition, findings suggest that such stylistic profiles may be connected with the use of visual, non-linguistic means of organising discourse.

In conjunction with the previous chapter, which presented an overview of the frequency of use of 1^{st}- and 2^{nd}-person pronouns, the findings in the current chapter provide two complementary perspectives on the degree of personal affect in instructional writing. Although differences in terms of the expressed level of personal involvement in the text samples in the current corpus might possibly be interpreted as an artefact of idiosyncratic variation rather than a periodic style of writing, the findings in part I of the current dissertation are also in line with general stylistic trends in prose writing in Modern English. The current instructive-informative text samples are seen to move away from personally involved advice and towards a more detached style of writing. Diachronic differences in the orientation towards the addressee in this text domain of secular writing are thus in line with developments in specialised and professional rather than popular registers of English.

Although borrowing from the historical pragmatics literature an onomasiological perspective at the outset of this chapter, it was not a primary aim here to provide an extensive description of the mappings of function-to-form. Albeit somewhat in the background of the current results, grouping directives in the current data set according to the various Addressee Orientation categories nevertheless revealed interesting patterns in the specific surface realisations of directive utterances. The data set thus seems to offer plenty of fruitful avenues for further research. We may highlight, for example, the use of constructions with *let* (primarily, *let*-imperatives), common realisation strategies that leave the agent of the future act in the directive unexpressed (i.e., Suppressed Addressee category realisations such as extrapositions, agentless passives, etc.), the frequency and use of modal verbs of necessity and obligation, and the specific discourse contexts that seem to elicit or license the use of Indirect Address directives.

Other aspects that may be addressed in future explorations may take some methodological issues into consideration. The differences in number of utterances and mean utterance length between samples might be one such avenue. Given comparable text sample lengths, for example, the current data seems to suggest that utterances become shorter in terms of absolute word length over time, while the number of utterances seems to increase. The question is how this relates to the linguistic realisation of directive utterances. Intuitively, commands (e.g., imperatives) may seem shorter in terms of absolute word length than descriptive utterances. On the other hand, grammatical constructions associated with the Suppressed Addressee orientation may appear as longer (and requiring more deliberate composition and possible re-editing) than such direct, interpersonal directive utterance realisations. The contradiction that early texts in the corpus are characterised by longer utterances in general, as well as a seemingly shorter type of directives, warrants a closer look at differences between mean length of directive utterances versus mean length of non-directive utterances – both in individual texts as well as across the corpus as a whole.

Part II

Bundles of Parts-of-Speech

Chapter 5

Exploring Periodic Prose Styles using Low-level Features of Grammar

5.1 Introduction

Large-scale corpus-based diachronic analyses (e.g. Biber & Finegan, 1989, 1997) have outlined historical trends in styles of writing, identifying clusters of linguistic features that co-occur in samples of text. Similarly, past and ongoing work at the University of Helsinki has provided valuable insights on linguistic developments in registers of English professional and specialist writing, particularly through the *Scientific Thought-styles*-project on the evolution of medical writing in English (see e.g., Pahta & Taavitsainen, 2014; Taavitsainen & Pahta, 2004). Both strands of research tap into the intuitive notion that, by exploiting a number of pre-defined linguistic features, we can chart diachronic changes in styles of prose. For example, Biber and Finegan set out to find groupings of co-occurring features to characterise (sociolinguistic) style and stylistic variation in genres of English (Biber & Finegan, 1989, p. 488), and an explicit aim in the *Scientific Thought-styles*-project is to describe stylistic developments in English medical prose.[1]

Insights on stylistic change gleaned from such large-scale studies can be connected to diachronic studies that trace particular grammatical features with a strong rhetorical footprint. For example, Lenker (2010) has noted that, in addition to grammatical and information structural change, the trigger for changes in adverbial connection in the history of English might well have involved conscious stylistic and rhetorical considerations by a group of influential natural philosophers. Examples presented by the author

[1] See http://www.helsinki.fi/varieng/domains/scientific%20thought.html [Accessed 17/02/2015].

illustrate how deliberate stylistic efforts in the 18^{th}-century have partly given shape to why, from a modern point of view, Chaucer's *Treatise on the Astrolabe* seems so markedly different from Adam Smith's prose, and what differentiates both from most of what has been written since (cf. Lenker, 2010, p. 233ff).

Halliday (2004), in turn, presents a study on the evolution of scientific language which provides qualitative insight on the empirical co-occurrence patterns found in multidimensional studies. From a functional-interpretative point of view, and referring to work by Mathesius (1928/64), Halliday points to certain changes in English that coincide with the introduction of printing and which involve quite substantial information ordering changes in the Late Middle and Early Modern English period, e.g., the rise of "thematic equatives" such as clefts and pseudo-clefts, changes in the use of the passive as well as changes in the thematic function of grammatical subjects (Halliday, 2004, p. 122). Such linguistic developments tap into changes in word order as well as discourse planning, and are difficult to make straightforwardly insightful using the macro-approach of a multi-feature/multi-dimensional framework.

Linking back to trends on which the *Thought-styles*-project chooses to focus, particularly the development from a medieval scholastic prose style to a new style of writing in the vernacular with the advent of the Early Modern age, Halliday argues that

> "in renaissance English there evolved a syndrome of features which collectively reoriented the grammar away from preoccupation with the experiential towards greater concern with the textual – from the clause as representation of a process towards the clause as organization of a message."
>
> (Halliday, 2003 [1990], p. 89)

Finding a method to prove or disprove such a claim based on a set of linguistic features is complicated, however. Common pitfalls include the illustration of a researcher's assumptions by expedient sampling of texts (in feature-based studies), or conversely, expedient sampling of linguistic features in large-scale macro-studies. After criticism of the field of stylistics by Fish (1982/1973, 1982/1979), this problem recalls the logical dilemma with which stylisticians are faced, termed the "Fish fork" by Stubbs (2005, p. 6):

> "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, p. 6)

This seems analogous to the trap of circularity in corpus-based research of selecting

texts on the basis of certain linguistic, rather than extralinguistic, properties (cf. Kopaczyk, 2013, p. 48). Scrutinising a corpus based on the former selection criteria will undoubtedly uncover these linguistic features of interest, but the value of such findings might be rather unsurprising (cf. Butler, 2004; Kopaczyk, 2013; Sinclair, 2004; Stewart, 2006).

Although comprehensive description, as described by Stubbs (2005, p. 6), ultimately cannot be achieved here, what may aid is to rely on methods that provide the features that are most distinctive of trends indicated by both strands of research in terms of proportional frequencies. In this way, the current approach is in line with Leech et al., who claim that "[f]rom a textual point of view, style is strongly associated with frequency" (Leech et al., 2012, p. 70).

5.1.1 Authorial fingerprints and contextual constraints

Idiosyncratic variation shows trends (habits), but is also constrained by the boundaries of the system: both in terms of what is permitted language-internally (e.g., in the syntax), as well as what has been learned (implicitly or explicitly) to be appropriate in terms of (con)textual conventions (a language external factor). For example, Stankiewicz alludes to syntactic norms that shape frequent forms produced in the context of production (Stankiewicz, 1991, p. 21). Thus, although syntax allows a range of constructional variation (Halliday, 1985/2004; Stankiewicz, 1991), it is more restricted than the vast choice in available lexical items.

Kopaczyk (2013) points to processes of acculturation and genre or text-type standardisation, which offer contextual cues for the composer of a message in determining what is appropriate in a given communicative situation. That is, a language user wanting to "fulfil certain communicative goals [will adjust] the linguistic production to the functional requirements of the context" (Kopaczyk, 2013, p. 45). However, such constraints may be provided by the wider demands of a certain genre, or by local constraints of a certain community of practice (see for example the volume by Kopaczyk & Jucker, 2013, for explorations of the notion of community of practice in the history of English). In turn, however, these constraints themselves are under the influence of social and cultural changes as well.

Genre is such a conditioning factor, affecting " 'processes of linguistic convergence and divergence in an environment of various coexisting norms, whether local, regional or national' (Meurman-Solin 2001: 241). What follows is that, in view of standardization as a process of extracting prestigious and efficient variants from an array of existing and competing forms, genre influences the choice of such forms for standardization purposes" (Kopaczyk, 2013, p. 41).

Taken together, these trends and boundaries may reflect predominant temporal

grammatical/stylistic templates. Neutralising the influence of genre as much as possible, we seek to maximise the degree to which trends seen in the data can be ascribed to the contextual influence of periodic norms or conventions.[2] Particularly in the research field of stylometry, or computation stylistics, the use of recurrent idiosyncratic patterns of language is exploited. Authorship attribution, the most prolific area of stylometry, seeks to apply such techniques in order to distinguish between authors – particularly with reference to problems of disputed authorship.

Rudman notes that nontraditional authorship attribution studies, i.e., studies that do not use external indicators or explicit mentions of a text's providence, are "based on the hypothesis that every writer has a unique and verifiable style" (Rudman, 2006, p. 611). Referring to Holmes (1998) and Koppel and Schler (2003), Tyrrk writes that the "common denominator" in various approaches to computational authorship attribution is that "an author's style can be understood as the totality of all the conscious and subconscious choices he or she makes during the process of writing" (Tyrrk, 2013, p. 186). In fact, with reference to the process of writing, Holmes (1985) mentions that Bailey (1979b) likens the process of writing as a Markovian model, given that "authors move through a continually branching network of choices, *this* word and not these others; *this* filling of a metrical position and not others that are possible" (Holmes, 1985).

However, conscious control of such patterns requires a high level of metalinguistic awareness, which is deemed rather unlikely, although not impossible (cf. Blackwell, 2000, and see below). Particularly in a forensic context, where applications of such stylistic (or stylometric) tendencies are highly relevant, the assumption is exploited that the use of function words and other, even more grammatical (or: less lexical, e.g., content words) surface forms escape producer's conscious control.

Bailey (1979a) specifies that the linguistic features on which stylometric analyses should be based "should be salient, structural, frequent and easily quantifiable, and *relatively immune from conscious control.*" (emphasis mine, Holmes, 1985, p. 330, citing Bailey 1979a). Sentence length, for example, is deemed an insufficient indicator of an author's stylistic habits as it is a likely variable that may be controlled in composition (Holmes, 1985, pp. 331-332).

It is noted that Baayen, van Halteren, and Tweedie remark that,

> "since the likelihood of storage in memory increases with frequency of use, and since awareness builds on memory, it is in the highest frequency ranges that conscious and deliberate wording and syntactic phrasing may be expected, leading to variation that is a function of, for example, narrative development rather than of an author's unconscious habitual use of language. Taken jointly, these considerations, which

[2] Genre-internal change can, of course, not be controlled for using such a methodology.

> pertain primarily to word usage, but which may also carry over to the highest frequency [of syntactic constructions], suggest that the lowest frequency ranges might provide a clue to authorship that is less contaminated by conscious rhetorical manipulation and thematic structuring that probably affect the higher-frequency units of analysis."
>
> (Baayen et al., 1996, p. 127)

Especially because adherence to genre conventions and periodic norms is in a sense also a deliberate effort, these high frequency items should prove of interest for an insight into such norms. What is more, the authors draw an explicit link between frequency of use of conventionalised text domain-specific strategies in the structuring of discourse (i.e., 'thematic structuring' and development of the narrative).

5.1.2 Exploiting linguistic form for the classification of texts

According to Lebart, Salem, and Berry, stylometric studies make use of statistical models based on quantifiable text linguistic form (cf. figure 1.1 Lebart et al., 1998, p. 10). Thus, stylometry should be seen as the study of quantifiable style markers with the aim of identifying unique features in an author's product of writing (and to differentiate it from the writings of other authors), or to distinguish between different texts written by the same author (Stewart, 2006, p. 769). The existence of a 'stylistic fingerprint', a unique blueprint of an author's idiosyncratic style of writing, is a central assumption on which the field is based (e.g., Holmes & Kardos, 2003; Juola & Baayen, 2005; Stewart, 2006).

Although the study of stylometry and disputed authorship has a tradition going back to at least the middle of the 19^{th}-century (cf. Holmes, 1998; Holmes & Kardos, 2003), two relevant strands that may be identified in the field of contemporary stylometry are literary-oriented studies aimed at identifying (dis)similarities between texts (or authors, registers, genres, etc.), and computer science studies which have a similar aim but regard authorship attribution primarily as an applied classification problem (cf. Eder, 2015, p. 169). The latter type of studies, for which an early example is the hallmark study by Mosteller and Wallace (1964) on the authorship of the *Federalist Papers*, aims at improving methodologies for textual classification and in most cases devotes little to no attention to the interpretative value provided by the feature set on which discrimination is based. Particularly since recent advances in machine learning, computational approaches to authorship attribution have attracted renewed attention (cf. Holmes, 1998; Juola & Baayen, 2005).

On the literary side, early work by Milic (1967) and Cluett (1976) may be mentioned. In addition, e.g., Burrows (1987, 1992) has advanced an influential approach which exploits the use of features of grammar rather than vocabulary richness or lexical

content for the identification of authorial styles. In particular, the use of frequent function words (e.g., *and, or, the, a, that*) is seen as a method that taps into author-specific patterns of language use which are largely free from conscious rhetorical manipulation (cf. Baayen et al., 1996). According to Holmes and Kardos (2003, p. 7), the "Burrows method" has proven to be a first port-of-call for many problems in authorship attribution.

Burrows compares human-made artefacts like literary texts to natural phenomena. Both are "repositories of meaning" (Burrows, 1992, p. 167), but whereas the latter are so 'by accident', the former are designed for that purpose. What both repositories have in common, however, is the fact that the meaning they contain can be thought of as meaningful "patterns amidst circumambient noise" (Burrows, 1992, p. 167). Discerning what is noise and what is meaningful, however, depends on the interpretation of the analyst – both for artefacts as well as natural stores of data. Importantly,

> "[i]n literary studies, as elsewhere, we examine and compare patterns. They may be sets of taxonomic differentiae, uniting certain texts in a genre or sub-genre or as moments in the evolution of a tradition or a school. They may be signs of individual authorship or of the larger cultural forces that inscribe themselves in and through the writings of individual authors"
>
> (Burrows, 1992, p. 168)

The search for templates of this kind seems an important theme for studies of literary stylometrics. For example, referring to the 'surface grammar' of cf., Fries (1952), Burrows notes that "[a] probabilistic grammar of English, in which rules of that kind were enunciated [e.g., "what is almost always, or usually, or not often, or only seldomly the case"], would much enhance our understanding of what is to be expected, grammatically, in writings of a given era, genre, or author." (Burrows, 1992, p. 172)

In connection, Burrows (1992, p. 174) calls for a focus on testing stylometric methods on cases where authorship attribution is less problematic to improve our understanding of general probabilistic patterns of the language. Although a parallel of such a proposal may be found in the use of a training set in computational unsupervised machine learning techniques on texts of uncontested provenance, what Burrows seems to have in mind is rather to identify patterns that inform our understanding of conventionalised styles of writing as evidenced by individual texts.

Similar calls have appeared elsewhere. In the context of metrics of vocabulary richness for authorship attribution, Holmes remarks that appeals by Tallentire (1972) for "linguistic templates i.e. norms for each literary period, genre and language so that we may isolate oddities peculiar to particular authors" (Holmes, 1985, p. 338) have largely remained ignored.

In a similar vein, Stamou calls for investigations of texts and text domains beyond the genres usually employed, "and not strictly belonging to famous authors[,] to examine the magnitude of stylistic change that one may detect" (Stamou, 2008, p. 196). Burrows (1987) states that the contemporary stylometric paradigm cannot be tested "as closely as it deserves" until stylistic methods are developed that are capable of comparing "the languages of different genres [and] different historical periods" Burrows (1987, p. 61). "Until we are better equipped to compare the stylistic properties of authors, periods, and genres, their relative importance cannot be assessed" (Burrows, 1987, p. 61). In addition, Forsyth notes that contemporary research into stylometry has barely started addressing this problem: "It seems fair to observe [...] that workers in stylometry have tended to rush into real problems without testing their tools on unproblematic cases" (Forsyth, 1995, p. 45).

Comments such as these make it seem worthwhile to investigate the features that may arise when investigating latent structures in the data. Conversely, although it is often noted that extra-linguistic factors such as genre, text type and periodic norms affect an author's style of writing, in most computational studies this phenomenon is usually regarded as a confounding factor rather than an object of study in and of itself.

For example, both Gamon (2004), Hirst and Feiguina (2007) present distinguishing between writings by the Brontë sisters as a notorious challenge in computational authorship attribution particularly because these writings involve authors from the same era and gender, with similar educational and socio-economic backgrounds, writing in the same genre and influencing each other's writing (cf. Hirst & Feiguina, 2007, p. 410). Text type, too, may "hide or change" the distinctive fingerprint of an author, as suggested by Tyrrk (2013, p. 201).[3] Referring to work by Biber (1995), Baayen et al. observe that, "[i]t is well-known that not only differences between authors, but also differences in register or text type are reflected in the relative frequencies of linguistic variables, many of which are syntactic in nature" (Baayen et al., 1996, pp. 121). Before embarking on practical questions of authorship, therefore, studies need to have at least a basic understanding of the range of variation to be expected in a certain register (Baayen et al., 1996, pp. 121-122).

Strong evidence for the conditioning influence of text domain is found in the lead-up to an authorship attribution task by Baayen et al. (1996). Based on a corpus of English prose across various registers of fiction and non-fiction, the authors note that the use of the most frequent function words, as well as the syntactic constructions with which these words are associated, is fairly consistent across a text (Baayen et al., 1996, p. 122). However, a principal component analysis (PCA) carried out on this data also

[3]It should be noted, however, that the perceived effect of such extralinguistic factors is also partially dependent on the analyst's choice of style-marker, with some markers being more affected by a certain factor than others.

shows that the two most important dimensions (i.e., principal components) are able to distinguish between the writings of one author writing in different registers (in this case, crime fiction and literary criticism). In turn, the differences between these two texts by the same author are shown to be greater than the differences between texts by different authors writing in the same register (Baayen et al., 1996, p. 122).[4] This leads the authors to conclude that "register variation masks author-specific variation" (Baayen et al., 1996, p. 129) when using most frequent words as a marker of style. Although the authors cannot rule out that the same results would have been obtained given either another random sample of texts from the same registers, or texts drawn from a random sample of other registers (cf. Baayen et al., 1996, p. 122), such empirical results underscore the importance of the conditioning effect of text domain or temporal norms, and substantiate calls by Burrows and others.

While noting that cross-register authorship attribution will benefit from "more systematic investigation of author-specific variation against the background of register variation" (Baayen et al., 1996, p. 129), the authors opt to control for this degree of inter-register variation by studying texts from a single register only (Baayen et al., 1996, p. 122).

However, although the above lends some support for our assumption that differences between texts in our current sample are reflective of another extralinguistic factor (i.e. periodic style) rather than idiosyncratic style of writing, author-specific variation cannot be ruled out as an important variable in the current experimental design. Although we note that some controls are available, for example in the form of texts written by the same author in different registers (or at least, sub-registers; e.g., the writings of Markham on a range of topics, and notably in poetry as well as prose), as well as texts written in the same period and register but by different authors, this has to be left for further research.[5] By studying texts from a single text domain, our current selection of texts is in concord with the approach proposed by Baayen et al. (1996). In addition, the findings by Baayen et al. (1996) suggest that, at least given certain markers of style, the influence of extra-linguistic factors on stylistic variation may be quite robust. In the next section, we will draw mainly from the computational stylistics literature for a

[4]Note that a third principal component axis, however, seems to capture exactly the shared features of these two texts, as opposed to texts by other authors writing in similar or different registers (cf. Baayen et al., 1996, p. 122).

[5]If such an agenda is to be carried out appropriately, it will have to control for other sources of variation in addition. In this context, we may mention an area of stylometric study known as stylochronometry (cf. Oakes, 2014; Stamou, 2008), which has as its main aim the assigning of texts written by the same author in the correct order of composition. An important confounding factor for such studies is the degree to which an author's style of writing is subject to change over time (i.e., intra-author variation). This phenomenon may be compared to the well-known distinction in sociolinguistic research between real-time change (differences across time for one cohort of a speech community) and apparent-time change (differences across age groups in a speech community at one point in time).

brief review of relevant features for the current approach.

5.1.3 Feature selection in authorship attribution and genre classification

Stewart (2006, p. 770) notes two approaches to the selection of features in such stylometric research: either selecting features that are deemed significant by traditional linguistic or literary study (cf. the procedure initially advanced by Milic, 1967), or selecting features, even if they seem arbitrary, that provide statistically significant differences between texts. Central in the second approach, which is particularly prevalent in quantitative approaches, is that "the *analysis* rather than the *analyst* determine[s] which word types and distributions [are] significant" (Stewart, 2006, p. 771; emphasis in original).

However, many different classifications may also be made based on the size of the style marker selected (e.g., "meta-word", word, sub-word, other; cf. Rudman, 2006, p. 614), the specific metric (frequency, bundle (n-gram), distribution rate, average, length, ratio, etc.), as well as the computational and statistical techniques employed. In fact, it may even be true, as Rudman (2006, p. 611) has warned, that every study in (non-traditional) authorship attribution uses a different methodology, and that experimental results cannot be straightforwardly compared as a consequence.

Two different feature set foci of stylometric studies may be distinguished. Although the list of features below is far from comprehensive, it will provide a general idea of the features belonging to each set:[6]

Lexical and surface string features E.g., word length, sentence length, collocations, (key) word frequency, n-word sequences, type-token ratio (lexical density), vocabulary richness, punctuation

Grammar- or syntax-based features E.g., function words, colligations, clusters or averages of tagged or parsed labels

The use of lexical and surface string features has a considerable tradition. In fact, up until the late 20^{th}-century, the main features used in stylometric research were word-based methods, particularly measures of frequent words or vocabulary richness (see for important overviews Holmes (1985) and Baayen et al. (1996)).

[6] Although there are stylometric studies that use content rather than form, for example by differentiating texts on the basis of discourse topics, these will not be discussed here. See however Stewart (2006, pp. 772-773) and Lebart et al. (1998, p. 7) for an overview. Also, Zhao and Zobel make a three-way distinction between lexical, grammatical and language-model methods in stylometry (e.g., Markov chains Zhao & Zobel, 2005, p. 174). As the latter type of method is not directly relevant for our current purposes, we have not incorporate it in our overview.

One of the first studies to use syntax-based features in an authorship attribution task is the previously mentioned experiment carried out by Baayen et al. (1996). Noting that the established usefulness of function words in the influential stylometric approach by Burrows (1992, 1993) taps into an author's idiosyncratic use of grammar, the authors argue that this phenomenon can be exploited even more productively. Whereas function words offer only indirect access to syntax, a domain assumed to be largely outside a writer's conscious control, this level might be accessed more parsimoniously using 'syntactic rewrite rules': combinations of annotation labels from parsed data (Baayen et al., 1996). According to the authors, "by focusing on syntax we are tapping into the more abstract, largely unconscious and hence most revealing habits of our authors" (Baayen et al., 1996, p. 128). The study's results show that although function words provide an economical method for discrimination that requires little text preprocessing, syntax-based methods offer higher discriminatory power and are more robust across texts than relying merely on words in the surface string (Baayen et al., 1996, p. 129).

A study that builds on the work by Baayen et al. (1996) in using syntax-based features for text categorisation is for example Argamon, Koppel, and Avneri (1998), who use a part-of-speech tagger to tap into textual style for a synchronic genre classification task. Where in the past genre classification studies mainly focussed on discourse content, the authors propose to apply a methodology based on syntactic features, arguing that the classification problem in authorship attribution and genre classification is a similar type of task. Since the generation of deep-structure syntactic labelling is time-consuming, the authors make use of function words as well as an automated POS tagger for the generation of POS trigrams as a "simple and robust" alternative method (cf. Argamon et al., 1998). According to the authors, "[POS] trigrams are large enough to encode useful syntactic information, while small enough to be computationally managable" (Argamon et al., 1998).

Koppel and Schler (2003), Koppel, Akiva, and Dagan (2003), Koppel, Argamon, and Shimoni (2003) continue the use of such shallow and computationally inexpensive features in the context of genre classification and author gender categorisation. Similarly, Santini (2004) uses POS trigrams on the basis of the CLAWS-5 tagger to annotate texts from the BNC for an experiment in genre classification. Interestingly, the author uses POS unigrams and bigrams as a cross-check for the appropriateness of trigrams for this task, and concludes that although bigrams with some settings have slightly higher accuracy rates than trigrams, trigrams in general outperform both uni- as well as bigrams (cf. Santini, 2004).

In an authorship attribution experiment using relatively short text fragments, i.e., 20-sentence segments of roughly 500-600 words, Gamon (2004) combines and compares both shallow grammatical features (function word frequencies and POS-tag trigrams)

as well as deeper levels of syntactic annotation (i.e., "context free grammar production frequencies and features derived from semantic graphs"; Gamon, 2004, p. 612). Although his results show that content-independent, deeper syntactic features in combination show a better result than n-gram metrics, POS trigram frequencies outperform all other feature sets if considered in isolation (Gamon, 2004, pp. 613-614).

Hirst and Feiguina (2007) use a partial syntactic parse (or set of rewrite rules) that is built upon part-of-speech labelling (Hirst & Feiguina, 2007, p. 408), and use bigrams (sequences of two successive elements) of these syntactic labels as input for their authorship attribution task. The authors note that such syntax-based features prove a successful method for attribution, even for the difficult task of distinguishing between writings of the Brontë sisters, and given the challenge of a small sample size (even for text samples just over 200 words; cf. Hirst & Feiguina, 2007, p. 415). However, it is also shown that the smaller the size of the fragment, the more the method benefits from additional features such as POS frequencies (Hirst & Feiguina, 2007, p. 415).

Graham, Hirst, and Marthi (2005) also make use of frequencies of part-of-speech tags for segmenting documents. Whereas POS-tag metrics are shown to be beneficial for discrimination, other style markers such as vocabulary-based features, character bigrams and low-level features such as sentence length and frequency of punctuation markers, are not.

These latter features were used by Stamatatos, Fakotakis, and Kokkinakis (2001), who finds that a combination of low-level features (sentence length, punctuation frequencies) and syntax-based features (e.g., counts of noun-phrases and verb phrases) is most beneficial for authorship attribution. In isolation, however, features of grammar outperform lexical or low-level features in distinguishing texts in a Modern Greek newspaper corpus.[7]

In a comparative study of stylometric methods and techniques, Eder (2015) concludes that the success of syntax-based features such as automated POS-tag trigrams are powerful markers of style. Although the lexically based most frequent word (MFW)-method is the most successful style-marker for authorship attribution (i.e., distinguishing between idiosyncratic styles) in a direct comparison of style-markers and classification techniques, it is closely followed by POS-tag trigrams. That is, using grammar-based POS-tags is in most cases a more successful method than relying on character based markers of style, particularly characters bundles, and irrespective of stylometric classifier (i.e., Burrow's Delta, support vector machines (SVM) and k-NN classification Eder, 2015).[8]

[7]Stamatatos et al. use a fully automated chunker (rather than semi-automatic parser, cf. Baayen et al., 1996), and use non-overlapping sentence chunks (e.g., NPs, VPs, AdvPs and PPs; Stamatatos et al., 2001, p. 197).

[8]Note that using the 200 most frequent words per author as a style-marker provides even better

Although Baayen et al. (1996) have argued that function words and POS-tagged texts tap into the use of grammar, many studies since have regard such indicators as shallow or "quasi-syntactic" (Koppel, Argamon, & Shimoni, 2003, p. 404) stylometric features at best. In addition, some studies such as those by Gamon (2004), report findings that suggest that the deepest syntactic feature sets might be most successful for a document discrimination task. However, such features are completely content-independent, and therefore do not offer promising features for an analysis that is aimed at providing successive interpretation of text material. Maybe not entirely surprising, therefore, methods proposed by Burrows (1992, 1993, 2002), are based on features which may be used for successive interpretation of literary texts. In addition, it may be recalled that our approach is to investigate trends in groupings of texts rather than to pursue a formal classification task. Given that Hirst and Feiguina and others regard POS tags are a feature "that straddles (or blurs) the line between the lexical and the syntactic" (Hirst & Feiguina, 2007, p. 411), this feature may be particularly useful for our current purposes. Added to this is that the use of POS tags, particularly in trigram sequences as shown above, is a robust method for detecting latent structures in the data.

Although the use of such POS trigrams is not yet common practice in studies on the history of English, the field is also not entirely devoid of approaches that use local bundles for linguistic analysis (see the next section, §5.1.4). In most cases, however, these studies concern the use of lexical rather than underlying grammatical bundles of linguistic material.

5.1.4 Local bundles and the history of English

With respect to a bundle approach, in this case of lexical bundles, Biber et al. (2004) have remarked that although traditional linguistic theory has not recognised such chunks as particularly meaningful, many bundles have "well-defined structural correlates" (Biber et al., 2004, p. 399). Evidence seems to be mounting for the importance of chunks to account for linguistic phenomena indeed. Erman and Warren (2000), Jackendoff (1997) and Sag, Baldwin, Bond, Copestake, and Flickinger (2002) all underscore the importance of multi-word expressions (MWEs), and in addition, other branches of linguistics have recognised the relevance of n-grams, for example the psycholinguistic studies by Brook O'Donnell, Rmer, and Ellis (2013), Ellis (2003) on the apparent use of chunks in language acquisition, and Bresnan (2014) for frequency effects

results for distinguishing between texts than using POS-trigrams. However, as a method based purely on lexical information, MFW may make for the most effective stylometric marker, but adds little to our understanding of what makes the texts in our corpus dissimilar in terms of grammatical properties. Since our aim is to contribute particularly to differences with respect to grammar(s) of prose, POS-trigrams thus seem a very attractive alternative.

in Present-Day English syntactic-change-in-progress. Similar methods are employed by Stubbs and Barth (2003), who employ "lexical chains" (i.e., 2-5 word n-grams) as a method for text-type discrimination in corpora of contemporary English. Similarly, in chapter 4 we encountered an interpretation of the use of recurrent lexical bundles, i.e., word-level n-grams, in the context of university teaching in findings obtained by Biber et al. (2004).

With particular reference to diachrony, several studies have applied methods similar to those used in the computational studies mentioned above for improving an understanding of trends in (the history of) English. For example, Culpeper and Kyt (2002) and Culpeper and Kyt (2010) use lexical bundles for the interpretation of recurrent functional patterns in English historical text. While the earlier study is presented as a pilot project, Culpeper and Kyt (2010) devote a chapter of their monograph on the use of lexical bundles in Early Modern English speech-based texts. Their results indicate that the use of formulaic lexical bundles in text types of prose are frequently used at transitions between direct and indirect speech (Culpeper & Kyt, 2010, p. 140). As the authors note, such formulaic phrases may have provided cues for the audience in a period before the establishment of clear graphological conventions for marking a shift to reported speech (Culpeper & Kyt, 2010, p. 140).

Another frequent use of lexical bundles in their corpus can be found at the start of 'units', either as launchers of utterances in play texts, or with a more discourse-organisational purpose at the start of clauses or phrases with an information-expanding function (Culpeper & Kyt, 2010, p. 140).Information-expanding bundles that illustrate this use are lexical clusters such as "as well as ...", "as much as", "the sort of" (Culpeper & Kyt, 2010, p. 131). In this context, Culpeper and Kyt refer to an observation by Biber et al. (2004) that some lexical bundles may be seen as introducing a structural 'frame' which is followed by an open 'slot'. Basically all the lexical bundles in the corpus of Culpeper and Kyt (2010) are said to comply with this pattern (Culpeper & Kyt, 2010, p. 140). In such a scenario,

> "[t]he frame functions as a discourse anchor for the 'new' information in the slot, telling the listener/reader how to interpret that information with respect to stance, discourse organization, or referential status."
>
> (Biber et al., 2004, p. 399)

Such findings support the view that some lexical bundles can be identified as formulaic expressions that have acquired a conventionalised discourse structuring capacity in Early Modern English (the specifics of which may be conditioned by genre or text type).

Kopaczyk (2013) focuses particularly on trajectories of conventionalisation and for-

mulaicity in her corpus of Scottish legal and administrative documents – a text domain which may be expected to contain a high degree of formulaic expressions. In line with conclusions that may be drawn on the basis of the findings by Culpeper and Kyt (2010), Kopaczyk remarks that the study of lexical bundles offers a particularly suitable tool for studying standardised expressions "on the level of the phrase and in the overall discourse structure" (Kopaczyk, 2013, p. 49-50). Such investigations of lexical bundles avoid the "atomistic approach" of studying the function of individual linguistic features which is "liable to miss much of what is going on" in the immediate context of discourse (Culpeper & Kyt, 2010, p. 140).

Culpeper and Kyt (2010, p. 103) recall work by Brinton (1996) and Traugott and Dasher (2002) as early studies that have investigated conventionalised multi-word expressions in historical perspective. To our knowledge, however, the only studies with a diachronic focus that apply bundles of non-lexical material for purposes of linguistic interpretation are the aforementioned studies by Tyrrk (2013) and Ernestus, van Mulken, and Baayen (2006). Although the study by Tyrrk (2013) mentioned earlier in this chapter was primarily framed as an authorship attribution problem, his use of POS frequencies to study Early Modern English medical texts generates new perspectives for the study of nominalisation strategies in the history of English. Similarly, the results based on tag trigrams of Ernestus et al. (2006) readily call for further linguistic analysis of genre and authorship differences observed in Old French manuscripts.

With respect to methodology in diachronic research, it is relevant to note that Culpeper and Kyt make use of a spelling regulariser (VARD) to arrive at comparable lexemes for their calculation of lexical bundle frequencies (Culpeper & Kyt, 2010, p. 112). In addition, their decision to use three-word lexical bundles is based on both methodological considerations (e.g., a manageable number of bundles to be analysed) as well as providing enough material which is suitable for interpretation (Culpeper & Kyt, 2010, p. 106). The present approach, which will be outlined below, builds on such considerations.

5.1.5 A meso-level, data-driven approach to charting periodic styles

A global approach such as undertaken in variationist multidimensional studies provides useful insights into register-specific conventional patterns of prose writing, for example using the frequency of personal pronouns as an indicator of the personal nature of a discourse as seen in chapter 3. However, as Biber et al. (2004, p. 400) remark, high frequency patterns in a corpus do not appear by coincidence, nor can they explain other (structural) features; they represent descriptive facts which need to be accounted for, and thus require linguistic interpretation. Such insights may be complemented by an inspection of local linguistic bundles, e.g., grammatical chunks or text domain-specific

formulaic phrases, to further our understanding of the extent to which frequently co-occurring patterns convey stylistic predispositions in textual composition.

In a certain way, our purpose puts the methodology employed in most authorship attribution and text classification studies on its head. We do not want to discriminate between texts but rather, using the features described above, investigate what it is about the prose writing of particular periods that gives it the sense of either shared or divergent styles. In this sense, ours is much more a text classification task than an authorship attribution task, although the methodology behind both approaches is comparable (cf. Argamon et al., 1998). However, note that we do not assign groups of texts *a priori*: first we want to know *if* Early Modern texts cluster together, and only then should we be concerned with the features which may explain these clusters.

A problem for our general aims in this chapter is that in the application of stylometric techniques, few studies link findings gained from statistical or computational analysis back to (patterns in) the texts in their sample set. Exceptions are for example Hoover (2003), who interprets his statistical findings in terms of the textual patterns found in *Nineteen Eighty-Four* and *The Inheritors* (note that the latter is subject of an analysis by Halliday (1981) which has been very influential in linguistic approaches to literary style (cf. Fish, 1982/1973; Hoover, 2003)). Tyrrk (2013) relates POS frequencies to the profiles of texts in a corpus of Early Modern English medical texts, noting how figures tally with differences between authors, sub-genres of medical writing and adherence to formulaic style conventions (cf. Tyrrk, 2013). Ernestus et al. (2006), in turn, do not go into a deep analysis of the tag-trigrams with high discriminatory power in their corpus of Old French texts for reasons of space. However, in their discussion of findings, they do mention how one frequently occurring pattern appears to be indeed more typical of verse rather than prose, and provide a sample from an Old French text (Ernestus et al., 2006, pp. 76-77).

With this in mind, we want to use a method that allows a reasonable discrimination between texts, as well as one which provides clear insights into the use of grammar by particular (groups of) authors. Given our current purposes, therefore, we do not want to lose all explanatory information in pursuit of discriminatory power. Since it is our aim to add concrete findings that may help gauge periodic styles, our approach will have to involve features that have substantial discriminatory value while still providing explanatory detail and insights (this is in contrast to for example Baayen et al., 1996, who note that consistent patterns of variation between texts is of lesser importance; p. 129).

Using bundles of parts-of-speech, as for example in Ernestus et al. (2006), may aid in such an objective. In Halliday (2003 [1990]) and in Halliday (2004), information-structural changes in the history of English are mentioned that are associated with

particular linguistic features which may induce perceptions of period-specific stylistic differences (cf. Adamson, 2000; Biber & Finegan, 1997/2001). In addition, if we want to compare to what extent authors within a certain periodic style rely on postmodified encodings for head noun modification as opposed to a pre-modifying or conditional strategy, for example, consecutive POS elements can capture such recurrent patterns. E.g. "Hay that is [too coarse]" (NN1-CJT-VBZ) versus ". Coarse hay [is]" (PUN-ADJ-NN1) versus "When the hay [is too coarse]" (AVQ-DET-NN1; tags as per the CLAWS-5 tagset). Note that such patterns that do not involve syntactic recategorisation, but rather reflect textual norms relevant to the (sub)register, could be equally as revealing when we want to account for differences between periodic profiles. Importantly, the use of n-grams allows us to assess whether texts show interpretable trends without making any *a priori* (genre) assumptions regarding linguistic features expected to occur in the data.

By charting features that may indicate period-specific conventions relating to tendencies in prose grammar (cf. Perret, 1988), the current chapter should be seen as pursuing a middle ground which connects both micro- and macro-approaches to surface variation in diachrony. For the retrieval of such features, and to guide our interpretation, we rely on objective statistical techniques. However, we focus not on individual tokens of lexis or grammatical constructions, but rather on such recurrent patterns of grammar (i.e., local bundles, chunks or n-grams). This data-driven method can thus be situated at the meso-level between multi-feature quantificational studies and historical studies which take into account a small set of features in great detail.

Next to exploration via a principal component analysis (PCA)[9] and discriminant analysis, both Ernestus et al. (2006) as well as Baayen (2008) use the Old French data set for a statistical procedure termed correspondence analysis (CA), which provides a useful visualisation of statistical frequency (count) data. We will introduce the rationale behind this technique in our next section.

5.2 Method

5.2.1 Corpus sampling and sample size

As a consequence of the selection of subject material on (global) discourse topics, it was unavoidable that the total amount of text per author per global topic is in some cases relatively brief. On top of that, the n-gram generation necessitated an equal text

[9]See also e.g., Binongo and Smith (1999), Binongo (2003) for application of PCA in stylometric research.

size cut-off point, ultimately yielding texts of exactly 4,000 tokens.[10]

It needs mentioning that the selection of 4,000 tokens was not random over the complete set of tokens available per author. Rather, these tokens constitute the first 4,000 tokens, and thus includes openings on a global topic (chapters, paragraphs, sections, etc.) but not necessarily equal amounts of closing sections due to the sample cut-off point and varying section sizes per topic. This was found to be an acceptable sacrifice, since it would maximally result in one incomplete clause per text sample. A possible alternative, e.g., cropping at the last punctuation marker before the 4,000-token cut-off point would have been more problematic, since it would have resulted in inconsistent frequency counts in the cells per text (which may even get amplified as a function of differences in sentence/clause length). In addition, there was no counterbalancing for the section length of passages on the respective discourse topics. Thus, it is possible to have 2,000 words on hay and 2,000 on watering in one sample, versus respectively 2,500 and 1,500 in another, for example.

5.2.2 Spelling regularisation using VARD

After manually entering the data and some minor cleaning, these texts were standardised for spelling using the Variant Detector VARD 2 (cf. Baron & Rayson, 2008).[11] The current study 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). Where appropriate, additional R packages are mentioned. The string processing, for example, made use of the R library `gsubfn` (Grothendieck, n.d.).

5.2.3 POS tagging & trigram generation

After submitting these texts to VARD, an automated tagger was used to enrich the data with Part-of-Speech tags. For the current study, we used the online CLAWS4 tagger in combination with the CLAWS-5 tagset (60+ tags).[12] The limited number

[10] This concerns POS tokens, not words, since the cut-off was introduced after POS-generation and subsequent lexical stripping.

[11] This involved lower casing, deletion of punctuation markers around roman and arabic numerals, deletion of illegible sections, expansion of abbreviations and special characters, e.g., "y^{t}","&c.", and variations thereof, to "that", "etcetera".

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

of tags in this particular tagset is advantageous for two reasons: it avoids fine-grained tagging errors that more elaborate tagsets are prone to, in addition to providing a potential significant increase in frequency counts.[13]

The same data was also tagged using the CLAWS-7 and the Penn-Treebank NLP tagger available through Python and the Apache OpenNLP tagger in R, package "OpenNLP". After evaluation of results, the CLAWS-5 tagset seemed most robust of the three automated taggers used. Using a more fine-grained tagset such as the CLAWS-7 and collapsing tags into more general classes, a method chosen by Tyrrk (2013), would have been an alternative method for avoiding fine-grained tagging errors as well as increasing frequency count totals.

Naturally, spot checks were carried out to inspect the accuracy of the automatically obtained POS-tags. For a number of texts, a cross-check was available in the form of manually annotated POS-files in the Penn-Helsinki corpus. Although minor inconsistencies were found (note that 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, foreign word)[14], these differences did not seem important enough to question the validity of the POS tags obtained using the current methodology. For example, an obvious spelling variant that was missed in VARD and required post-editing in some of the Early Modern texts (e.g., Blundeville, 1565; Markham, 1607; Morgan, 1609; Baret, 1618) was the lexeme 'bee' (21 instances), which had been consistently tagged as 'bee_NN1' rather than 'be_VBI'.

After enriching the data with these tags, some post-processing was carried out concerning punctuation: markers for quotations and bracketing were removed from the data (these were marginal classes by any measure). All other punctuation markers are part of a single category with tag PUN, as per the standard CLAWS-5 tagset. The remaining tagset had exactly 60 possible POS categories (see appendix A for an overview of the tagset, and appendix B for the most frequent POS categories in the texts in our corpus).

Example (1), taken from Baret (1618), illustrates the various steps of the current 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?

[13] See for example Glynn and Robinson (2014, p. 451), who report that the statistical technique presented in §5.2.4 is sensitive to distortion with cell frequencies of less than 8 counts. Cells with values below 10 should generally be avoided as a rule of thumb, and reported if they do occur (Glynn & Robinson, 2014, p. 464).

[14] See https://www.ling.upenn.edu/hist-corpora/annotation/labels.htm#pos_tags.

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. [POS string extraction]
`AV0 CJS PNP VHZ VBN AT0 NN1 TO0 VVI AT0 AJ0 NN1 PRP AT0 NN1 PUN CJC TO0 VHI PNP VVB CJC AV0 AT0 NN1 PUN CJC AV0 TO0 VVI NN0 PUN CJC DT0 AJ0 PUN PRP AT0 NN1 TO0 VVI DPS NN1 PUN DT0 DT0 PNP VDB VVB PRP PUN CJC AVQ PUN`

The data was subsequently stripped of lexical content, leaving only a string of POS tags per file. These strings served as the basis for the calculation of POS frequency averages as well as the generation of n-grams (i.e., POS grams).[15] The size of the n-grams was set to three, based both on a number of diachronic studies which report promising results using this setting (using either lexical bundles or bundles of grammatical tags; cf. Culpeper & Kyt, 2010; Ernestus et al., 2006), as well as empirical research on contemporary English (cf. Aarts & Granger, 1998; Stubbs & Barth, 2003; Gries, Newman, & Shaoul, 2011). In a comparative experiment of n-gram length for the study of subregisters in PDE, Gries et al. (2011, p. 10) conclude that trigrams strike an optimal balance between linguistic interpretability, statistical power and computational costs.

The trigrams generated on the basis of the string above in (1-d), up until the first noun tag, are provided in (2).[16] As can be seen, the created bundles are overlapping. 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, irrespective of lexical content). With Stubbs and Barth, we do not claim that such recurrent sequences are linguistic units in themselves, but do contend that they "provide evidence which helps the analyst to identify linguistic units" (Stubbs &

[15] N-gram generation was carried out in R using the `RWeka` library (Hornik et al., 2014).

[16] The NN1 tag is merely chosen as an end point for purposes of illustration. Note that there are no stop signs for the generation of trigrams until the end of each text sample is reached.

Barth, 2003, p. 69).

Using the R packages mentioned above, the various sequences that are generated can be grouped and tabulated by frequency. A cumulative list of the POS-trigram frequencies found across the 13 text samples in our corpus serves as the basis for a Correspondence Analysis (CA; cf. Benzcri, 1973; Greenacre, 1984; Greenacre, 2007; Murtagh, 2005), the specifics of which will be discussed in the next section.

(2) [trigram generation]
```
AV0 CJS PNP -- CJS PNP VHZ -- PNP VHZ VBN -- VHZ VBN AT0 -- VBN AT0 NN1
```

5.2.4 Correspondence analysis – rationale

Correspondence analysis has a wide range of applications, from corpus linguistics (e.g., Ernestus et al., 2006; Tummers, Speelman, & Geeraerts, 2012, 2014) and forensic linguistics ((cf. Bcue-Bertaut, Kostov, Morin, & Naro, 2014) to studies in chemistry, ecology, epidemiology, marketing and tourism (cf. Beh & Lombardo, 2014, for an overview). It is an exploratory multivariate scaling or ordination technique and as such is related to for example principal component analysis, 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 CA. 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 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.

As with other ordination techniques, the purpose of correspondence analysis is to reduce the dimensionality of a data matrix (e.g., every row or column might be considered as one dimension, resulting in a high-dimensional space) back to a number of latent dimensions. For the particular mathematics behind the reduction of the original number of dimensions to a few latent dimensions, see e.g., Greenacre (1984), Greenacre (2007), Murtagh (2005). Greenacre argues that while the mathematical and computational principles underlying correspondence analysis are fairly simple, the technique should be regarded as a geometric technique rather than a statistical one (Greenacre, 1984, p. 11). This highlights one of the main advantages of the method, which is to facilitate inspection of the often 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).

Nenadic and Greenacre (2007, p. 3) note that "[t]he total variance in the data matrix is measured by the inertia [...], which resembles a chi-squared statistic but is calculated on the basis of the relative [rather than absolute] observed and expected frequencies" (see also Greenacre, 2007). Since every latent dimension contributes to inertia, the output of the solution provides the principal inertia (or eigenvalue) for each dimension, as well as a percentage for how much it contributes towards total inertia. Eigenvalue rates can be compared to the variance of the principal components in PCA. Dimensions with high principal inertias account for most of the difference between intra-row and intra-column distances, and are therefore primary candidates for the axes of the low-dimensional plot.

Since the aim of correspondence analysis is to reduce a number of high-dimensional points in the data to a low-dimensional display which is visually interpretable, high percentages on the first few dimensions is sign of a good solution (cf. Greenacre, 1984, p. 67). Conversely, if a large percentage of the total inertia is distributed over many dimensions, this means that the points in the data set are not being represented well in the solution presented (Bendixen, 1996, p. 26).

There are several ways to approach the issue of selecting the number of dimensions, both visual (e.g., via a scree plot) or by using statistical means (see for an overview Beh & Lombardo, 2014, pp. 145-147). Two fairly simple methods are suggested by Bendixen (1996) and Beh and Lombardo (2014). Although noting that many practical applications of correspondence analysis ignore a strict threshold, Beh and Lombardo (2014, p. 145) advises to select the number of dimensions that together account for at least 70% of the total inertia (also termed the retention of the solution). Bendixen (1996, p. 26) suggests to compare the percentage of inertia accounted for by each dimension to the expected average contribution of either the rows and columns in the model (whichever yields the higher value), and to include dimensions with higher percentages than this expected average contribution (see the results below for examples).

Based on the data in the rows and columns in the contingency table, two distance (or Burt) 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 (sample texts, in the current case), while another contains the distances of columns by columns (i.e., 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 Burt matrix.

The plot that is drawn on the basis of this data, the symmetric or French plot, includes both the distance matrix for rows as well as that for columns superimposed onto each other ("analogous to the superposition of principal component scores and the loadings on these PCs in the biplot" in a principal components analysis; cf. Baayen, 2008, p. 129). As the singular value decomposition of the problem explains the complexity contained in the data in two ways (once via the rows, and once via the columns), these axes can be presented as overlapping, and the profiles for each row and column are drawn in the same plot.[17] Although we will return to the issue of interpretation of data points in the symmetric plot below, outputting the data in such a low-dimensional way allows the visual detection of clusters of associated rows and columns. Such CA plots, in addition to the underlying frequencies, distances, correlations and contributions of the rows and columns on the dimensions/axes, serves as the basis for our qualitative interpretation of the data.

Due to their ability to isolate confounding variables in the presence of multiple explanatory variables, as is often the case in corpus linguistics (cf. Tummers et al., 2012, 2014), statistically more advanced multivariate methods such as multiple correspondence analysis (MCA) are also available. Since the design of the current study is not to isolate certain variables, but rather, intends to explore the cumulative pattern revealed by the 'variables' in the rows (POS trigrams) and columns (text samples), it seems appropriate to restrict ourselves to what is called a simple correspondence analysis (CA). The current correspondence analyses are computed in the statistical environment R using the libraries `ca` (cf. Nenadic & Greenacre, 2007; Nenadic & Greenacre, 2014) and, as a cross-check, `languageR` (Baayen, 2014).

Reflections on the methodology

Two particular methodological concerns regarding the procedure above may be outlined here. The first concerns the use of a single punctuation label PUN for a range of punctuation markers, both sentence-medial (, : ;) as well as sentence-final (. ? !), and the second issue relates to the use of (semi-)automatic spelling regularisation in combination with POS tagging.

First, and as already mentioned, a concern for low cell frequencies in successive quantitative analyses was an important argument for using one PUN tag for a range of markers with varying status. Other considerations were also taken into account,

[17] Cf. "[c]orrespondence analysis may be defined as the principal component analysis of either the row profiles or the column profiles. In each case the profiles are weighted by their respective masses and the metric is defined by the respective chi-square distance. Both these problems are symmetric versions of the same underlying singular value decomposition" (Greenacre, 1988, p. 42).

however, and the merits of three alternative solutions for dealing with punctuation markers will briefly be discussed here.

As may be expected, the danger of low cell frequencies is highest with a tagging scheme that has separate tags for each punctuation marker (cf. the CLAWS-7 tagset). Although tags and POS trigrams generated on such a tagset are more specific, and probably add discriminatory power (i.e., better able to discriminate between authors or text samples), the end-result of statistical analysis might answer a slightly different research question.[18] In addition, a tagset that differentiates between multiple punctuation markers while at the same time reducing the variety in lexis and condensing the grammar into 60 odd POS tags, does not seem to do justice to the complexity of the data. For example, considering the fact that all personal pronouns and modal auxiliary verbs, irrespective of inflection and contraction, are reduced to one tag each (e.g., PNP and VM0; cf. appendix A), the creation of a system of multiple punctuation markers was deemed inappropriate.

Another suggestion, as proposed by Joanna Kopaczyk (p.c.), was to delete the PUN tags before running the trigram generation, and compare this data to the results obtained with running a single PUN label. Although an interesting proposal, it was judged here that this would create artificial results (e.g., grammatical clusters that do not actually occur in the investigated text in the order presented), as well as complicating the interpretability of results. That is, after lexical stripping and trigram generation, it becomes nearly impossible to distinguish POS trigrams that occur sentence-internally from those that occur across a (deleted) punctuation marker in the original text.[19]

A third alternative, and most in line with the tagset used in the Penn-Helsinki corpora, is to make a distinction between sentence-medial and sentence-final punctuation. Although this poses an appealing consideration, it was judged to largely fail to take into account the evolution in status and usage of separate punctuation markers. For example, the colon sign seems much more of a sentence-final marker in the earlier samples in the current corpus, whereas in PDE texts its function as a sentence-medial marker seems more or less established. A correct application of such a procedure would neces-

[18] Conversely, Jockers (2013) does use punctuation in his multivariate analyses of literary style, albeit a slightly different family of statistical techniques from the one used here. Specifically, in one study "any word or mark of punctuation with a mean relative frequency across the corpus of above 0.03 percent was selected" Jockers (2013, p. 69). The decision to use the 0.03 cut-off point was arbitrary, according to the author, but this "number was selected in order to ensure that the features were not context-specific words but function words that are used at a high rate." (Jockers, 2013, p. 69fn; the author argues that the rationale behind this choice can be found in Zipf's law). The PUN markers above this threshold in the study by Jockers are the apostrophe, comma, exclamation mark, hyphen, period/full stop, quotation mark, and semi-colon (NB: out of 42 features selected for the texts; all other features are function words Jockers, 2013, p. 69).

[19] But see below. A feasible variant of this procedure that does not rely on PUN tag is to run the analysis with only those POS trigram (types) in the corpus that do not contain PUN tags (i.e., exclusion trigrams after trigram generation).

sitate an analysis of every text sample for the status of each specific punctuation marker, as well as authorial consistency in keeping to this standard, which would drastically shift the focus of our current aims. Although the study of punctuation merits further investigation, particularly for its potential as a marker of periodic style (cf. Burnley, 1986; Parkes, 1992; Smith, 2012), this was not judged to be a primary focus for the current procedure. Relying on one global PUN label and consulting the original texts when particular combinations of three successive tags occur to investigate the variety in different punctuation markers used, seemed the safest methodology.

An additional point of concern for the current procedure, and with potentially even greater methodological ramifications, is the accumulation of error that is inherent in the use of several (semi-)automated sub-routines in succession.[20] Although every effort has been made to minimise the number of human errors in the phase of transcription and keying in of the text samples from the available material, it cannot be guaranteed that the .txt files are without error. However, these should be regarded as human errors, and their effect is probably negligible in comparison to the danger posed by more systematic errors that might have occurred in successive steps. First in the degree of correct 'translations' from the cleaned .txt files to the VARD spelling regulariser, and successively in the number of correctly assigned POS tags on the basis of the modernised spelling in the VARD-ed text files. (Since n-gram generation is fully automated it is not subject to such translational errors, but its degree of error depends completely on errors in previous steps.)

Given the analysis presented hereafter, for example, it cannot in theory be ruled out that the axis in the dimensional reduction that is presented does not represent a temporal axis, but rather the degree of tagging error; with on the one hand the early texts which represent data points that are relatively high in error, and on the other hand the most recent texts for which the tagger is most accurate. Although this is a valid point for concern, we have tried to pre-empt it by carrying out regular cross-checks of the VARD-ed and POS tagged text files.

In addition, we would like to stress here that in some sense it may in fact be seen as a windfall that the problem outlined here constitutes a systematic error, rather than a random, hand made error for which the consistency cannot be established. Next to being highly labour intensive, manual tagging with slight inconsistencies in adherence to a tagging protocol might create differences across text samples that cannot easily be corrected afterwards. Conversely, any inconsistencies in VARD-ing and POS tagging which was detected in the current procedure, prior to lexical stripping, was easily corrected using text replacement functions that can run through the entire corpus in

[20] Thanks are due to Prof. Andreas Jucker for hinting at this issue in a paper presented at ISLE-3, Zurich, Switzerland in 2014.

one operation.[21]

In addition, a manual tagging protocol might have led to an idiosyncratic tagset that complicates the comparison of the current corpus to other texts. The use of an automated tagger at least ensures there is a reference corpus of texts tagged with the same tagset labels. Tyrrk (2013, p. 191) notes that in a pilot study by Hiltunen and Tyrrk (2012) using a similar methodology, it was established 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). For a more elaborate consideration of the degree of error allowed in POS tagging, see Mair, Hundt, Leech, and Smith (2002, with reference to the degree of error in relation to the research question one is trying to answer) and particularly Rayson, Archer, Baron, Culpeper, and Smith (2007) for the accuracy of POS tagging in combination with the use of VARD as applied on Early Modern English texts.

5.3 Results

The present section will present a variety of statistical visualisations of the data. We will start by describing some distributional characteristics of the data set in general terms, and will continue with two correspondence analyses that are carried out on the texts and POS trigrams in the corpus. The first correspondence analysis is based on the 309 most frequent trigrams which cover 50% of the trigram tokens in the current data set. Next, a correspondence analysis is carried out on trigrams that occur with 100 or more observations in the corpus (the 58 most frequent trigrams, covering roughly 24% of observations). After a comparison of these plots and highlighting of the particular position of texts and nearby tags, we will continue with a number of additional analyses. The first involves an association plot of the 10 most frequent POS trigrams in our corpus, followed by a summary of the findings of additional correspondence analyses which can be found in appendix C. Before finishing with a summary of the findings contained in this chapter, we will look at a hierarchical clustering analysis of the texts and POS trigrams.

[21] Our stance seems in agreement with a position held by Gamon (2004) who, in response to criticism on the use of automated language analysis (in this case, a POS tagger), notes 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). We have to take an additional precautionary step on top of the assertion by Gamon (2004), however, in making sure that the established correlations between linguistic features are carefully scrutinised before embarking on further analysis and interpretation of our data.

5.3.1 Basic statistics: Trigram permutations, hapaxes and raw frequencies

With the tag set used for the current experiment, the number of possible POS trigram types amounts to 216,000 (60^3). In this light, it is somewhat 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 (in a non-mathematical sense) 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.[22] Of these tag combinations, the number of hapax legomena (tag combinations that are used only once) amounts 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 used in the corpus. Surprisingly, these figures roughly correspond to the typical 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).

Figure 5.1 provides a plot of the frequency as a function of the POS trigram rank, with subfigures for the distribution as sorted by frequency of occurrence. As can be seen, the distribution of POS trigrams is somewhat suggestive of a Zipfian distribution (as seen in rank-frequency distributions of words in natural language corpora), with the frequency of use of a trigram inversely proportional to its rank order. That the distribution does not follow this pattern perfectly can be gauged from the fact that the second ranked trigram (i.e., NN1-PUN-CJC) occurs with a frequency of 677 in the corpus, whereas it would have had to be half the frequency of the first ranked trigram (i.e., PRP-AT0-NN1 with a frequency of 756) in a true Zipfian distribution. In turn, the second ranked trigram should have occurred twice as often as the third ranked POS trigram (AJ0-NN1-PUN), but instead the latter occurs 547 times in the corpus.

When the same data is plotted on a log-log graph (figure 5.2), the approximation of a straight line indeed suggests that the distribution of the data obeys a power law (although this can only be established conclusively using further statistical analysis; cf. Clauset, Shalizi, & Newman, 2009, for a methodology). Since it is not essential for the analyses in this chapter to establish whether the requirements of a Zipfian distribution are met, we will leave this matter for the moment, all the while noting, however, that

[22] For sake of comparison, Hirst and Feiguina use a rewrite rule on POS tags to arrive at syntactic *bi*grams and find 2,999 types of bigrams used on a total of 15,876 ($= 126^2$) possible combinations (i.e. 18.89%), using 115 out of a total of 126 different syntactic labels in a corpus of half a million words of writings by the Charlotte and Anne Brontë (cf. Hirst & Feiguina, 2007, p. 410).

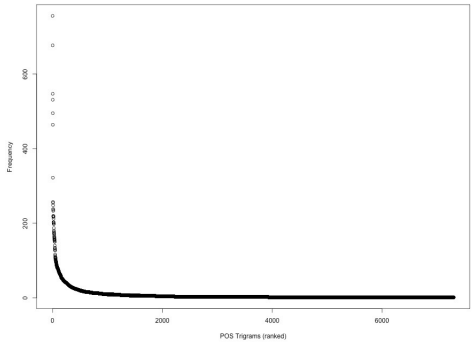


Figure 5.1: Distribution of POS trigrams

it is remarkable that POS trigrams in our corpus (a very condensed derivative of the actual lexical content) appear to be distributed somewhat in a Zipfian manner.

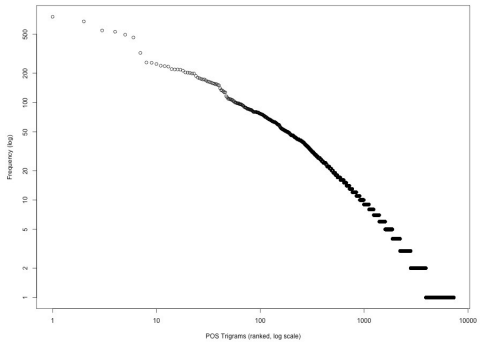


Figure 5.2: Distribution of POS trigrams (logscale)

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. As a result of n-gram generation, the number trigrams is $n - 2$ the number of tags in the sample. This is due to the fact that the first and last element of every 4,000 POS-sample cannot feature as a central element of a tag trigram (as there is no prior element for the first, and no following element for the last POS tag, no trigrams can be generated on the basis of these two tags).

In addition, just over half of the total number of tokens is covered by the 309 most frequent trigram types in the current data set, capturing 26,005 observations. As a result, note that less than one percent (309/216,000*100 = 0.1430556) of all possible combinations of three consecutive POS tags covers half of the observations in the data. For computational efficiency, the correspondence analysis and plots provided

in the next section are based on this half of the data, rather than the full data set containing the lower ranked trigrams and their frequencies (e.g., hapaxes). As the focus of the current chapter is to capture particularly those POS bundles that are indicative of periodic styles of writing, relying on trigrams that make up half of the observations in the full data set seems an adequate strategy (and recall the concern voiced by Glynn & Robinson, 2014, regarding low frequency cell counts). It is assumed here that it is particularly such frequently recurring patterns, rather than the large number of combinations appearing only once in a certain text or group of texts, that will cumulatively induce the notion of a particular style of writing.

5.3.2 Correspondence analysis: 50% of tokens

Before we consider the graphical information provided by the correspondence plots, the solution to the dimensional reduction must be inspected. We will restrict ourselves to reporting only the most essential information in the correspondence analysis for the interpretation of periodic clusters.

Table 5.1: Scree plot for correspondence analysis (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	4.0	88.2	*
9	0.017303	3.5	91.7	*
10	0.015024	3.0	94.7	*
11	0.013245	2.7	97.4	*
12	0.012698	2.6	100.0	*
Total inertia	0.493153	100.0		

Dimensionality of the solution

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 5.1). Using the method for dimension selection proposed by Bendixen (1996, p. 26), the expected average inertia is 100/(14-1) $\approx$ 7.69% for the rows (text samples), and 100/(309-1) $\approx$ 0.32% for the columns (tags). Only the first three dimensions have percentages above the highest of these values, and will accordingly be selected for the current solution. These dimensions explain respectively 34.8%, 13.9% and 11.9% of the inertia, to a total retention of 60.5%.

In contrast, the cumulative threshold of 70%, as proposed by Beh and Lombardo (2014, p. 145), is only reached when the fifth dimension is included in the solution; with the fourth and fifth dimension accounting for 7.2% and 6.6%, the total retention in the 5-dimensional solution lies at 74.4%. However, such a model seems a sub-optimal solution for a visual inspection of the data.

In addition, Tummers et al. (2014, p. 492) draw attention to the fact that high differences in percentages between dimensions suggest an asymmetry in the solution. Although there is some asymmetry in the three dimensions selected here (e.g., the first dimension accounts for nearly three times as much inertia as the third dimension), these differences do not seem large enough to cause concern based on the requirements set out in Tummers et al. (2014).

Before we investigate the visual output, it is necessary to inspect the rows and columns in the output in terms of the contribution to the three principal dimensions included. Restricting our observation to the text samples, the quality scores of the sources by Blundeville (0.397), Gibson (0.455), Hunter (0.129) and Skeavington (0.106) are questionable. (Glynn & Robinson, 2014, p. 469) note that "[a] quality score of less than 500 (50%) suggests that the position of the data point in question does not necessarily accurately represent the relation of that feature to the others". Although this observation does not invalidate these text samples outright, it must be kept in mind when interpreting their representation in the discussion of the plots drawn below.

Inspection of the symmetric biplot

The symmetric plot that is drawn on the basis of this correspondence analysis is provided in figure 5.3. In this plot, the POS trigram data points are indicated by triangles, and rows (text samples) indicated by dots. In addition, it may be observed that the shading in these plots indicates the relative contribution of data points to the dimensions, with darker shades indicating a higher contribution.

As is clear from the figure, a correspondence analysis based on the 309 most frequent tag trigrams generates a two-dimensional plot that is rather cluttered. In particular, the labels indicating the sample texts are obfuscated by the cloud of tags representing the data points of the POS trigram. For the remainder of this section, starting with figure 5.4, we will draw the plots without POS trigram labels.[23]

The labels of the axes show the factors 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.

[23] Note that despite not displaying the POS trigram labels in following, the positioning of text samples still relies on the cumulative frequencies of the underlying POS trigram found in these texts.

The two dimensions plotted here thus account for 48.6% of the inertia in the solution. Although this degree of retention in the first two dimensions of the model is not bad, it also indicates that more than half of the total inertia is explained by the remaining dimensions in the correspondence analysis.[24] Nevertheless, biplots may only assist in guiding an exploration of the often rich and complex collection of data, and we can only selectively discuss some of the information conveyed by the plots below.

A word of caution is necessary, however, in the interpretation of the standard graphical output for a simple CA (cf. Beh & Lombardo, 2014; Bendixen, 1996; Greenacre, 2007). The symmetric plot is based on the positioning of the axes representing the principal inertias of two separate distance tables (one for the rows and the other for the columns) onto each other. As a result, the row-to-row distances and column-to-column distances can be assessed at once. However, the symmetric plot does not offer a meaningful perspective for the interpretation of the distances between row (text) data points on the one hand and the column (trigram) data points on the other Greenacre (2007, pp. 72, 267). For a correct interpretation of the row-to-column distances, we need the axes to be drawn on the basis of the principal coordinates of one Burt matrix (e.g., the rows), and draw the data points of the other distance matrix (columns) as standard coordinates, as is done in what are termed asymmetric plots. Nevertheless, although the distances between trigrams and texts should be interpreted with caution, overlapping angles with respect to both axes may indicate similar trajectories for texts and trigrams.

Interpretation of axes and clusters

What can be gleaned from the plot with sample texts only (cf. figure 5.4) is that the data points do not appear to be plotted completely at random: 16^{th}- and 17^{th}-century texts cluster in the upper left quadrant, with the 18^{th}- and 19^{th}-century manuals in the centre near the origin, towards Davies' contemporary manual on the far right. However, it is also obvious that this slope is far from perfect: the 16^{th}-century manual by Blundeville (1565) does not appear towards the left but rather near the centre of the plot, for example, and the mid-19^{th}-century text by Skeavington is similarly displaced. One would have expected it to appear closer towards the right-hand side of the plot, among his contemporaries Kirby and Fleming. Other peculiarities include the relative position of Gibson compared to the other 18^{th}-century manual by Hunter.

The 17^{th}-century manuals all appear with comparable 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

[24] On the other hand, Lebart et al. (1998, p. 57) argue that even with very low inertias, lower than the current figures, a representation that accounts for the data relatively well may be valid still.

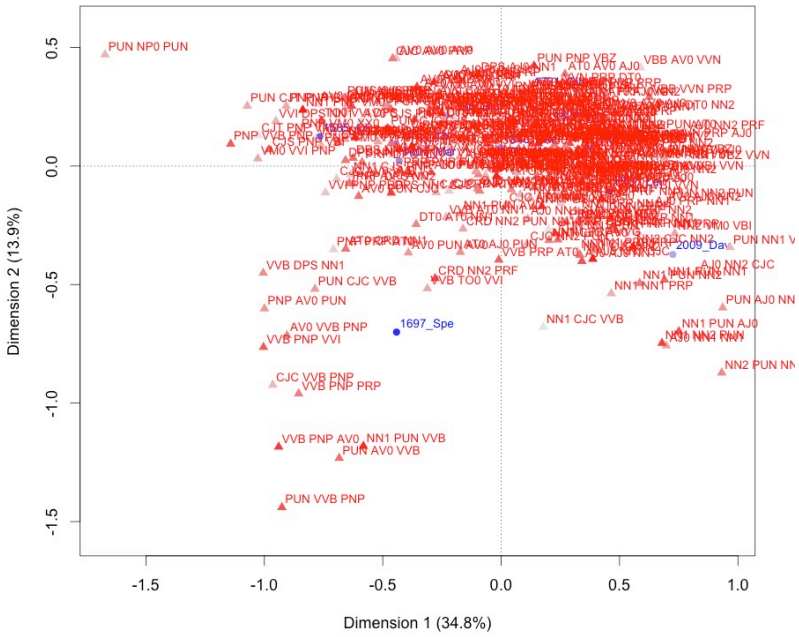


Figure 5.3: Symmetric plot of trigrams and sources (50% of tokens)

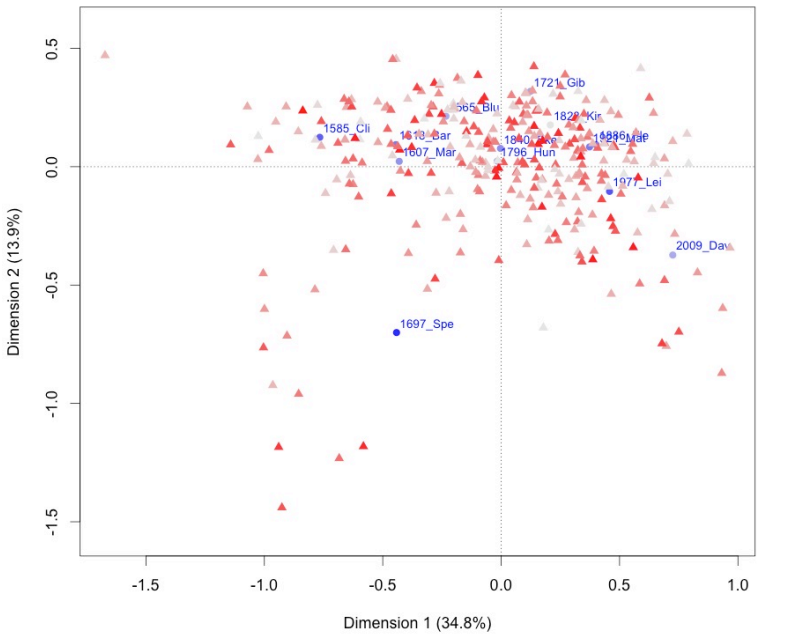


Figure 5.4: Symmetric plot of trigrams and sources (50% of tokens, no POS labels)

coordinate on dimension 2 in the corpus as a whole. We will discuss the position of the sample by Speed particularly in section 5.3.3, as well as in appendix C.

Dimension 1 The first dimension is associated with this particular ordering of text samples, and is also the dimension which accounts for most of the inertia in the data set as a whole (in both the current correspondence analysis as well as the one presented hereafter, in section 5.3.3). It suggests that this factor is related to chronology, although a great deal of additional information is 'hidden' in the 10 dimensions not captured in the current biplot (cf. the scree plot above).

An interpretation of the trigrams and their location in the plot, in a sense the other side of the coin, requires slightly more information than the relative position of texts. A cursory glance at 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. However, 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 5.2 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.

Table 5.2: Highest POS trigram correlations on Dimension 1 (in permille)

#	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

These ten trigrams provide an additional guide to an interpretation of factor 1. The quality 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, in turn, means that a given POS trigram is strongly associated with a particular factor (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 earlier or rather the later 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-a)). In addition, it serves to illustrate the frequent use of personal pronouns in Early Modern English texts in general, as seen in for example chapter 3.

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

b. " [...] 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)

The second example, (3-b), shows the trigram that is listed as the fourth combination with a high correlation and a negative coordinate on Dimension 1 (#12: DPS-NN1-PUN). In addition, it may be noted that this combination is preceded by a preposition, resulting in a 4-gram of the form PRP-DPS-NN1-PUN which also includes the first trigram combination mentioned here (#23).

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*:

(4) "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)

(5) "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).[25] The trigram contains a tag for a (subordinating) conjunction which, based on the plot,

[25] Note that the tagger here does not distinguish between the subjunctive and an infinitive form of 'be'. The difference in the frequency for this trigram in the early and later periods may thus also reflect a decline in the use of the subjunctive.

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 correspondence analysis below. One example is provided in

(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, we find combinations such as a past participle form of a lexical verb (VVN), a preposition other than 'of' (PRP) and an article (AT0; #31):

(8) 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 indicative of manuals in the later part of our corpus. In fact, the use of past participles of lexical verbs in three of the five POS trigrams on the positive scale may indicate prose styles which frequently employ the use of the passive voice. As is often observed, the passive voice is a common feature of contemporary informative prose, particularly scientific writing (e.g., Biber et al., 1999; Halliday, 2004; Huddleston, 1971; Swales, 1990).

The second trigram, VBZ-VVN-PRP (#133), represents an *-s* form of the verb 'be' (either *is* or the contracted form -'s), a past participle form of a lexical verb and a preposition other than 'of'. This type of trigram is found in sentences such as

(9) 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 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 (10):

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

The fourth POS trigram highlighted in this context again contains a past participle form of a lexical verb, a preposition other than 'of' and an unmarked adjective (i.e. not comparative or superlative; #206: VVN-PRP-AJ0). Two examples, (11-a) from Matheson and (11-b) 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, a singular noun and an 'of'-preposition (cf. also the phrases "corresponding increase of work" in (11-b) and "sufficient proportion of" in (10), to name just a few others).

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

Dimension 2 Providing an interpretation of the second factor, 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 5.3 lists three POS trigrams with the highest correlation on this factor.

For tags in the positive region, it is particularly the use of modals and pronouns which stands out in the plot. Table 5.3 contains two of such POS trigrams, i.e., #212: a modal verb, negative marker and a 'be' infinitive (VM0-XX0-VBI) and the combination

Table 5.3: Highest POS trigram correlations on Dimension 2 (in permille)

#	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

of a possessive determiner (e.g., 'your', 'his'), general adjective and a singular noun (#201: DPS-AJ0-NN1). The third POS trigram with a high correlation on dimension 2 has the form of a preposition, general determiner (e.g., *these, some*) and a singular noun (#41: PRP-DT0-NN1). Although this might not strike the contemporary reader as a particularly Early Modern combination, a decrease in frequency for this particular POS trigram can be detected in the present corpus after the onset of the 20^{th}-century. E.g.,

(13) "[...] 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)

In turn, all three 'negative' tags contain a punctuation marker, and it might therefore be suggested 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, & Huddleston, 2002). Several arguments seem to go against such a conclusion, however.

The first cue is that the plot indicates that all three negative tags can be found near the sample by Speed. These tags may well represent a strong association with this particular text, and if the text by Speed indeed turns out to be an outlier, these POS trigrams may similarly have to be regarded as outliers in the distance matrix over columns. Interpreting the axis in terms of the contribution of these tags may prove somewhat superfluous, therefore, particularly when dimension 2 mainly expresses a distinction between the text by Speed on the one hand, and all the other texts on the other (see §5.3.3 below).

Another problem with regarding the vertical axis as a dimension of punctuation is that although the first three POS trigrams with positive signs are all devoid of punctuation, this is certainly not the case for all trigrams in the positive domain of this axis. In fact, although not provided in table 5.3, the three POS trigrams above are followed

by three other trigrams with high correlations on this dimension that do contain a PUN tag: AJ0-PUN-PNP (#227: quality= 475, correlation= 351, coordinate= +363), AJ0-PUN-CJC (#38: qua= 489, cor= 300, coo=+374) and AV0-PUN-PNP (#38: qua= 413, cor= 263, coo=+345). Since such a finding obviously does not necessarily discount the hypothesis that the negative domain on dimension 2 is related to a heavy use of punctuation, these findings do require a closer inspection of the specific distribution and status of PUN-containing trigrams in relation to this dimension before any such conclusions could be drawn.

Another argument that would go against the straightforward interpretation of dimension 2 as a measure of punctuation is the fact that the two latest texts, those by Davies (2009) and Leighton-Hardman (1977), show up in the negative domain of the vertical axis. This suggests that these texts are associated with a high frequency of punctuation, or at least POS trigram sequences containing such PUN tags, but this is not what is suggested by the overall POS frequencies. Instead, these texts appear to be relatively low in terms of average use of punctuation (see appendix B).

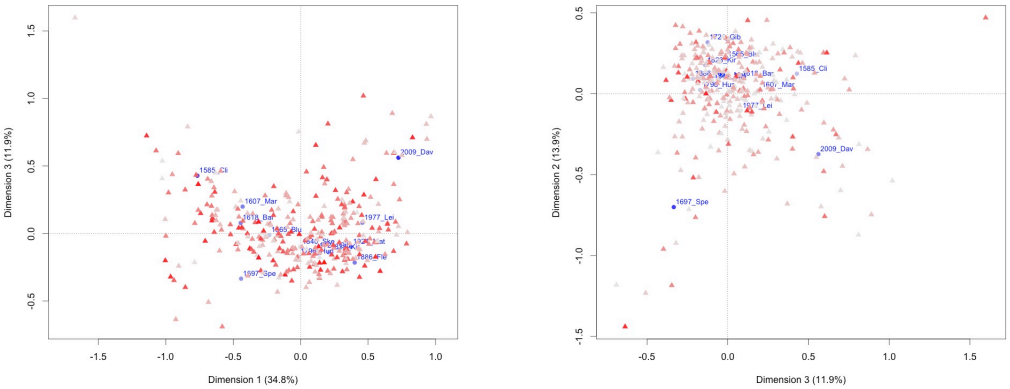
A general problem for the trigrams listed in the positive domain of dimension 2 is that these show a fairly low quality (i.e., 500 points and lower; cf. table 5.3). This indicates that there is a high probability that their position in the correspondence plot is not being reflected very accurately in the current solution. Some caution is therefore warranted in relying on these data points for an interpretation of the plot. For now, we will therefore consider dimension 2 as primarily a continuum that marks the distinction between an outlier (Speed, 1697) and the other texts in the corpus. We will return to this in section 5.3.3, where the number of tags in the plot is more easily interpretable.

We argue that the slope that can be detected in the plot of dimensions 1 and 2, in combination with the underlying sample texts and POS frequencies on which it is based, seem to suggest a diachronic trend away from the use of pronouns (both personal and possessive) and towards a reliance on the passive voice and potentially larger nominal clusters with both pre- and post-modification (or pre- and post head dependents; cf. Payne & Huddleston, 2002), see examples (8-a) through (12). Although further exploration of the data is warranted, it seems that an inspection of the first two dimension already provides results that dovetail findings on developments in genres and registers of English by for example Biber and Finegan (1997) and Halliday (2004). However, we will also briefly need to turn towards the third significant axis in the dimensional reduction of the current solution.

Dimension 3 The contribution of a third principal axis also turned out to be significant, with its eigenvalue accounting for 11.9% of total inertia. The plots in figure 5.5 provide a perspective on the data taking into consideration this additional dimension.

The figure on the left, figure 5.5a, provides a view of the cloud of data points 'from the top': while the horizontal axis still represents the first dimension, the vertical axis representing the third dimension now adds depth to the perspective obtained in figure 5.3.

Conversely, figure 5.5b shows how the three-dimensional space containing the cloud of data points in the symmetric plot is approached 'from the right' (that is, if the assume that the plot in figures 5.3 and 5.4 provides a frontal perspective on the data). The second dimension is represented here as the vertical axis, as in figure 5.3, and the third dimension is plotted here as the horizontal axis (cf. the *vertical* axis in figure 5.5a).[26]



(a) Symmetric plot, 50%, Dims 1 & 3 (b) Symmetric plot, 50%, Dims 2 & 3

Figure 5.5: Symmetric plot of the third dimension (50% of tokens)

With dimension 1 capturing the temporal difference between texts, interpreting the third dimension on the basis of just the positions of these texts is complicated. Instead, we may focus on how the third dimension helps in identifying clusters of texts in the plot.

The manuals that mark the outer bounds on either side of the origin in figure 5.5a are the texts by Speed and Davies. What is particularly remarkable about this third dimension, however, is that both a number of early texts (i.e., those by Clifford, Markham and Baret), as well as the two latest manuals by Leighton-Hardman and Davies, occur with positive coordinates on dimension 3. Nevertheless, looking at figure 5.5b, we can see how there is a division between the earliest texts occurring above the horizontal axis, and the PDE texts below it. This suggests that although these five

[26]It is noted that we deviate here from the practice to reserve the horizontal axis in the CA plot for the dimension with the highest principal inertia. Maintaining the second dimension as the vertical axis, as in the first symmetric plot presented here, and have the third dimension as the horizontal axis was assumed to make the visual inspection of the points in successive graphs somewhat less laborious.

manuals are separated from the majority of 17^{th}-, 18^{th}- and 19^{th}-century samples, they might have to be regarded as two different clusters on the basis of dimension 2.

This picture is reinforced when we look at these points from the top, using figure 5.5a: the earliest texts (Clifford, Markham and Baret, and marginally Blundeville) seem to represent a cluster in the upper left-hand side of the plot, whereas the PDE manuals are positioned on the top-right (with positive coordinates on dimension 1). The bottom-left quadrant is reserved for one author, Speed, whose text sample always occupies a quadrant in each of these respective plots in relative isolation.

Lastly, the picture conveyed by both figures 5.5a and 5.5b is that the texts published between the start of the 18^{th}-century and up until the beginning of the 20^{th}-century appear largely as one cluster in the plot. Although on the basis of the first two dimension (cf. 5.4) it was not possible to establish whether the manuals by Hunter (1796) and Skeavington (c1840) form their own cluster or not (as they appear so close to the origin), the perspective provided by the third dimension puts these texts in line with the Late Modern English cluster. The distance of these texts to the center of the cluster of Late Modern English texts is perceived to be the shortest by far.[27]

Using a visual inspection of figures 5.4, 5.5a and 5.5b, there is some evidence to suggest that there are four sectors (clusters) in the dataset:

1. An Early Modern-cluster: Blundeville (1565), Clifford (1585), Markham (1607) and Baret (1618)
2. The late 17^{th}-century manual by Speed (1697)
3. A cluster of Late Modern English texts: Gibson (1721), Hunter (1796), Kirby (1823), Skeavington (c1840), Fleming (1884) and Matheson (1921)
4. The Present-Day English manuals by Leighton-Hardman (1977) and Davies (2009)

Although it will reduce some of the richness of the data, reducing the number of POS trigrams will facilitate the interpretation of data points in plots such as the above. In the next section,we will present the results of a correspondence analysis and a (less cluttered) plot of only those POS trigrams that occur with considerable frequency in the corpus (i.e. 100 attestations or more).

5.3.3 Correspondence analysis: POS trigrams with 100^{+} counts

The decision to include only those POS trigrams for which the cumulative cell frequency in the corpus lies at 100 observations or more is somewhat of an arbitrary one (but then again, see Ernestus et al., 2006, , who only consider the top 35 trigrams). Nevertheless, these 58 POS trigram types still cover approximately a quarter of the tokens in the data

[27] But recall that these two texts were identified above as having the lowest quality scores across the text samples in the solution to the dimensional reduction problem in this correspondence analysis.

(23.50598%, to be exact). 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 based on these 58 most frequent trigrams offers a better visual inspection than the plots in section 5.3.2. On the other hand, the positioning of texts above is in some sense more reliable (not in terms of statistical robustness, but rather in terms of external validity), because the CA on which it is based takes into account a greater proportion of the POS patterns found in these text samples. The lower coverage of the tokens in the dataset is thus offset by a more easily interpretable plot. Where appropriate, however, we will link our current findings up with those in the previous correspondence analysis.

Dimensionality of the solution

As before, the scree plot of dimensions in the solution indicates that the first three dimensions offer a good retention of the inertia in the data (cf. table 5.4). Both by the criteria put forward by i.e., a cumulative percentage over 70%, Beh and Lombardo (2014) as well as those by Bendixen (1996).[28] Compared to the previous correspondence analysis, the first three dimensions account for more of the inertia in the model (e.g., 74.1% versus 60.5% previously). This is not entirely surprising: as the number of relevant tags is reduced, there is less variation in the data and fewer relevant dimensions will be necessary to explain the inertia in the model. Given that the dimensionality of the current problem is far smaller than that in the previous CA, it is only natural to assume that selecting a similar number of dimensions as before will account for more of the variation seen in the data. As can be seen, the inertia in the model is reduced from 0.493153 to 0.295120. As we have the same number of text samples accounting for the inertia over rows, the (poor) quality of the sources by Blundeville, Gibson, Hunter and Skeavington is the same as before.

Inspection & interpretation of the symmetric biplot

The symmetric plot of the correspondence analysis, figure 5.6, reveals a basic pattern for the text samples as seen in figure 5.4. Both plots point to a similar general structure, revealing clusters of texts and tag trigrams that are stable over both subsets of the data (i.e., roughly 25 % and 50% of tokens). However, some of the POS trigram labels are now more easily identifiable.

[28] Although the number of columns are reduced to 58, and the expected inertia over columns therefore changes to 100/(58-1) $\approx$ 1.75%, the expected inertia over rows is still maintained at roughly 7.69%. The percentages of the first three dimensions all exceed this mark.

Table 5.4: Scree plot for correspondence analysis (100^+ observations)

Dimension	value	%	cum%	scree plot
1	0.132283	44.8	44.8	**********
2	0.056618	19.2	64.0	*****
3	0.029757	10.1	74.1	***
4	0.019216	6.5	80.6	**
5	0.012586	4.3	84.9	*
6	0.010624	3.6	88.5	*
7	0.009058	3.1	91.5	*
8	0.007508	2.5	94.1	*
9	0.005668	1.9	96.0	
10	0.004870	1.7	97.7	
11	0.003744	1.3	98.9	
12	0.003186	1.1	100.0	
Total inertia	0.295120	100.0		

On the left-hand of the dimension, we can see the use of personal pronouns (PNP) in conjunction with punctuation markers (PUN) or modals (VM0). On the right, the occurrence of prepositions (PRP) and singular nouns (NN1) seems to corroborate our earlier findings that the use of nominals is characteristic of texts on this side of the scale. As an interpretation of dimension 1 was provided above, and its general structure is maintained in this plot, it might be of greater interest to look in greater detail at some other features of the current solution.

Dimension 2 POS trigrams that have considerable contributions on dimension 2 can now be inspected with slightly more accuracy in the graphical display of figure 5.6.

The POS trigrams with the highest contributions on this axis in the current correspondence analysis are the tags PUN-CJC-AV0 (ctr= 38), PNP-VM0-VVI (ctr= 26), NN1-PUN-NN1 (ctr= 29), PUN-VVB-PNP (ctr= 294), NN1-PUN-VVB (ctr= 207) and CJC-VVB-PNP with a contribution of 85. All of these have qualities above 500 and therefore represent relatively reliable mappings. In addition, the POS trigrams PUN-CJC-AV0 and PNP-VM0-VVI are the only combinations in this list that have positive coordinates on dimension 2. They can be detected in the top left of the cloud of points in 5.6.

On the other hand, POS trigrams that contribute to dimension 2 with negative coordinates are the trigrams NN1-PUN-NN1 (observable in the bottom right), NN1-PUN-VVB, PUN-VVB-PNP and CJC-VVB-PNP. These latter three trigrams occur in the bottom left hand of the plot, and in fact, could already be detected in figure 5.4 above, near the coordinate of the text by Speed (1697). Given this association between POS trigrams with high contributions on dimension 2 and coordinates close to one particular text, there is reason to suspect that dimension 2 is an axis that represents,

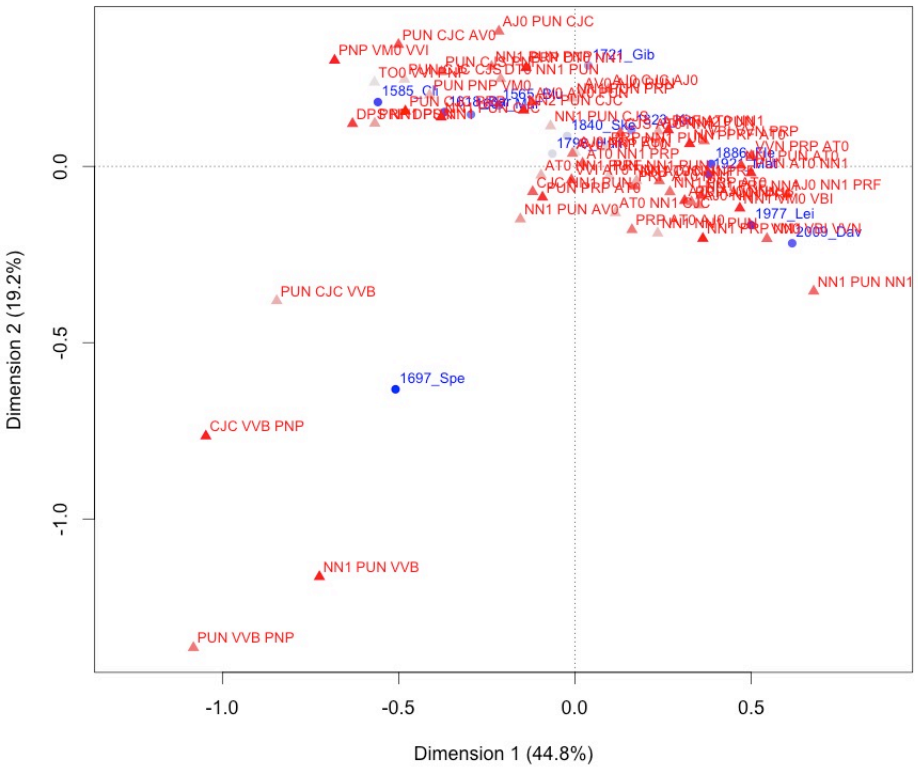


Figure 5.6: Symmetric plot of trigrams and sources (100^+ observations)

at least in part, the distinction 'Speed versus not-Speed'.

The tags that are not located in the lower left quadrant of the plot appear as the extremes in a rough diagonal running through the not-Speed group. An overlap may be observed between the early texts in the upper left and tags containing either a coordinating conjunction and an adverb after a punctuation marker,

(14) "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)

or a personal pronoun with a modal auxiliary and a lexical verb infinitive (i.e., the tag already mentioned above, PNP-VM0-VVI), e.g.,

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

Conversely, the singular noun before and after a punctuation marker (i.e., NN1-PUN-

NN1) seems to appear in the quadrant that is occupied by the late 20^{th}- and early 21^{st}-century text samples in the lower right of the plot:

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

So, although dimension 2 may primarily express the difference between Speed and not-Speed, there is also a trace of a diachronic difference in the use of modals and personal pronouns and combinations of (singular) noun tags. In fact, traces of a temporal distinction are not surprising if the cloud of data points is indeed spherical and follows a slanted, rather than a straight, line through dimension 1. Looking at the data points in figure 5.6, the plot indeed suggests that the cloud is best captured by a somewhat diagonal line from the top left to the bottom right. Nevertheless, such conclusions on patterns in the data can be misleading if we only take into account the first two dimensions. Before we turn to the third dimension of this plot, however, we will briefly look at the outlying position of the text by Speed, and associated POS trigrams.

The outlier in the bottom left-hand quadrant was already noticeable in the previous correspondence analysis. Its position and some of its features can now be interpreted more easily, as the four trigrams in the vicinity of Speed's late 17^{th}-century *The Gentleman's Compleat Jockey* are clearly visible.[29]

The tag relatively close to the location of Speed's sample in the plot (slightly higher and to the left) corresponds to the POS trigram sequence PUN-CJC-VVB: a punctuation marker, a coordinating conjunction, and a base form of the verb (e.g., ", and give", ", and feed", ", and bleed").[30]

Another trigram close to Speed's coordinates, even further towards the left and close to the border of the current window of the plot, reflects the sequence of a punctuation marker, a base form of a lexical verb and a personal pronoun (PUN-VVB-PNP). This sequence reflects sentence openings and phrases after the use of a comma or semi-colon, such as "take him", "saddle him" and "litter him", to name just a few corresponding patterns in the text by Speed. As may be inferred, 'him' in such instances refers to

[29] Again, it is noted that we are dealing with a symmetric plot here. Therefore, although these trigrams occur in the neighbourhood of the text by Speed (1697), we must be wary that their distance to this sample text should be interpreted with caution. Nevertheless, since no other textual sources appear in this quadrant, establishing an association between POS trigrams and this sample may be somewhat less controversial.

[30] This sequence of a CJC followed by a VVB mainly occurs after commas. Of the 301 CJC tags in this text, about two-thirds appear as 'and' ($n = 212$), with other categories including 'or' ($n = 65$), 'but' and 'nor' (respectively, 21 and 3 occurrences). 'And' is only occasionally found with semi-colons (n= 16), once after a colon and never after full stops. In fact, there is only one coordinating conjunction found after a full stop in this text, which is a ". But" sequence. This particular use of coordinating conjunctions in combination with punctuation thus also indirectly hints at the difference in status for the punctuation markers employed by Speed.

the discourse entity of the generic horse, and the tag for the base form of a lexical verb (VVB) in this combination as well as the previous POS trigram corresponds to the use of an imperative. Both tags seem particularly indicative of Speed's recipe-like manual.

The four combinations found near Speed's text in the plot may actually turn out to be overlapping, with the combination of PUN-CJC-VVB (39 observations in Speed, out of a total of 134 in the corpus) and CJC-VVB-PNP (41 out of 100) forming a continuation after a punctuation marker of the form PUN-CJC-VVB-PNP. E.g.,

(17) "[. . .] , and give him three of them a day" (Speed, 1697)

In turn, NN1-PUN-VVB (52 observations out of 106) may coalesce with PUN-VVB-PNP (64 out of 109), to form the 4-gram NN1-PUN-VVB-PNP. The utterance in (18) below is a particularly rich example, offering a combination of this specific POS 4-gram (the second underlined phrase), as well as one of its underlying trigrams (i.e., PUN-VVB-PNP at the opening of the sentence), in addition to the previous 4-gram (PUN-CJC-VVB-PNP):

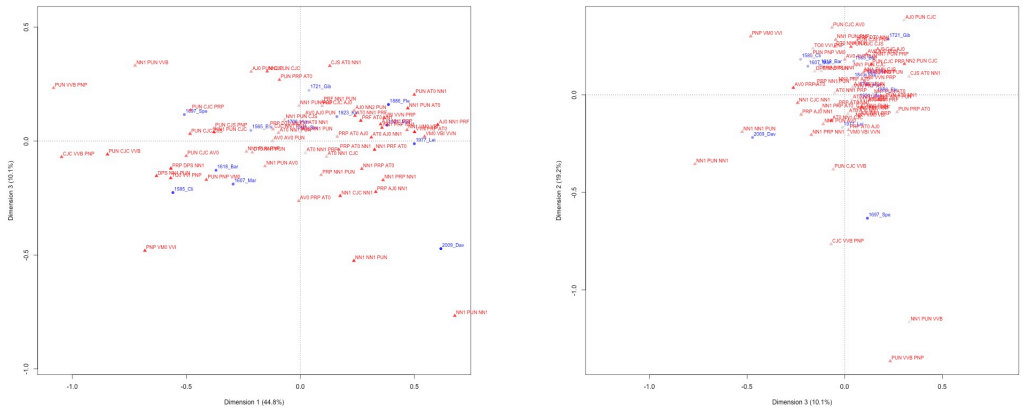
(18) ". Give him a due proportion of provender, litter him very well , and let him be clean rubbed down" (Speed, 1697)

Such 4-grams combinations in the sample mark the manual by Speed (1697) as rather procedural, and more so than the other texts in the corpus. An inspection of the original source text will lead most readers to suspect that the manual by Speed (1697) has a remarkable 'recipe-like' character, even in comparison to other sample texts in the current data set. Although it is by no means remarkable to find such phrases in an instructional texts, what is rather striking is that the use of these sequences turns out to be so frequent in Speed, and with that, indicative of his idiosyncratic prose style. One feature that might account for this is that Speed seems to refer to the horse with a personal pronoun in such procedural sections, which might be distinctive for his style of writing (or at least, the earlier portion of the current corpus).

Although the four POS trigrams in the lower left quadrant of the plot can be found across the corpus, high attestations in the manual by Speed seems worthy of note, as it greatly determines the position of this text in the plot. Conversely, the relative infrequency of these trigrams in the other instructional texts in the sample may be pointed out. Together, such differences in frequency counts on the trigrams mentioned here, between this text and the other texts in the corpus, results in this particular manual being located so far off the general cluster observed in the plot. In additional analyses in which Speed's sample is considered a supplementary row (thus not affecting the computation of the CA), the text appears to be assigned a position closer to the 17^{th}-century sample fragments (albeit a bit further to the right on dimension 1; see

section 5.3.5 and appendix C).

Dimension 3 As before, the inspection of principal inertias indicated that the contribution of a third principal axis was significant (this time with an inertia of 10.1%). The two plots in figure 5.7 provide perspectives on this dimension from above (fig. 5.7a) and from the right-hand side (fig. 5.7b) of the cloud of data points.



(a) Symmetric plot, 100^+ obs., Dims 1 & 3 (b) Symmetric plot, 100^+ obs., Dims 2 & 3

Figure 5.7: Symmetric plot of the third dimension (100^+ observations)

When we look at the perspective on the cloud in 5.7a, the text by Speed does not appear to have an extreme value on the third dimension. Rather, it is the text by Davies that takes up a unique position in the plot, in the lower right-hand corner. Note also that the texts with which it is usually associated, the manuals by Matheson (1921) and Leighton-Hardman (1977), appear approximately on the horizontal axis in this plot.

Not surprisingly, taking into account dimensions 2 in conjunction with the third dimension (cf. figure 5.7b) shows the pattern seen in the discussion of dimension 2 above: Speed (bottom right) versus not-Speed (texts in the top half of the plot). In addition, a rough diagonal may be observed from the mid-left of the plot, through the origin, to the top right. Here, too, the position of Davies (2009) almost seems to lie outside the main cloud of points. What the third dimension thus adds to the picture seen in figure 5.6 is that the text by Davies may also be somewhat of an outlier, although slightly less so than the manual by Speed. Although it seems excessive to suggest that dimension 3 represents a 'Davies versus not-Davies' distinction, dimension 3 helps to suggest a difference between the text by this author and other PDE texts in our corpus.

In terms of contributions, then, Davies indeed has both the highest contribution on

dimension 3 (ctr= 0.463) as well as the largest (negative) coordinate (i.e., -0.472). At some distance, this is followed by the samples of Clifford and Gibson, with contributions of respectively 0.101 and 0.113. Clifford, similar to Davies, has a negative coordinate here (-0.226), whereas Gibson has a positive coordinate of 0.223.

When we look at trigrams that contribute to this third dimension, it is in particular the earlier mentioned tags NN1-PUN-NN1 (with coordinate -0.767 and contribution 0.257) and PNP-VM0-VVI (-0.481, 0.126), as well as the nominal trigram NN1-NN1-PUN (-0.525, 0.099) which stand out.

Based on these data and an inspection of the plot, it may be tempting to infer that the third dimension might tie in with a frequent use of the use of singular nouns and punctuation markers. however, this seems a too quick a conclusion, especially given the fact that is exactly the NN1 and PUN tags are the most frequent POS classes in our corpus overall. They will therefore feature in a large number of trigrams.

What is more remarkable is the contribution offered by the combination of a personal pronoun, a modal verb and the infinitive form of a lexical verb, as seen above. The trigram PNP-VM0-VVI occurs 198 times in the corpus, and particularly in the early texts by Clifford (n= 35), Markham (n= 42) and Baret (n= 39), but also in the fairly late manual by Skeavington (n= 22). Although we cannot make any firm claims based on the occurrence of merely one tag, it could suggest that use of modals in combination with personal pronouns might be antagonistic to the use of singular nouns in combination with punctuation. That is, if text samples are characterised as making use of either one of these tags, the frequency for the other pattern will be marginal.

Other tags appearing in the vicinity of the modal tag PNP-VM0-VVI, with Early Modern connotations:

PRP-DPS-NN1 & DPS-NN1-PUN Potentially form a 4-gram cluster of a preposition, possessive determiner, singular noun and punctuation marker, e.g., *for/to your horse.*

TO0-VVI-PNP Infinitive marker 'to', infinitive of a lexical verb and a personal pronoun, e.g., *to give him.*

PUN-PNP-VM0 Another possible 4-gram in combination with the above modal tag PNP-VM0-VVI, in particular when it occurs just after a punctuation marker.

In conclusion, the correspondence analysis on the 58 most frequent trigrams shows that the same basic pattern may be found as in the previous CA, indicating that the structure of the data is relatively robust. A look at the 10 most frequent POS trigrams and their occurrences across samples may reveal addition patterns of stylistic association.

5.3.4 Distribution of the 10 most frequent POS trigrams across text samples

Reducing the set of POS trigram types even further, we may investigate the relationship between the most frequent POS trigrams in our corpus in terms of their occurrence across text samples. Using an association plot, cf. figure 5.8, we can determine the degree to which certain trigrams are associated with certain texts. The 10 most frequent POS trigrams selected here together account for approximately 8.77% of all trigram tokens in the data.

When interpreting an association plot, it should be kept in mind that these 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 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 (a negative value, in dark) conversely indicates that there are fewer observations for a certain trigram in a text than expected, based on the average.

Because the graphical output in figure 5.8 does not allow the plotting of labels for all text samples (horizontal) and POS trigrams (vertical), these labels can be derived from table 5.5 (with the first row in the table corresponding to the top line in the plot, and so on). Next to providing the corresponding labels for the axes in the association plot, the table provides absolute frequencies for these 10 POS trigrams.[31]

Inspecting the association plot reveals exciting results. For example, the signal observed for the second trigram from the top, NN1-PUN-CJC, shows a remarkably positive association with the early half of the corpus and a negative association with the second half. That is, Early Modern texts appear with lightly shaded 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 5.5 indicate that this POS trigram is frequently used throughout the corpus.

The opposite distributional pattern seems best visible for the fourth row, representing the trigram AT0-AJ0-NN1 (e.g., *a young horse*). In the middle period in our

[31] 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. In addition, these 10 trigram cell frequencies and their row and column totals are a subset of a larger contingency table, which might be problematic (Gries, 2014, p. 376).

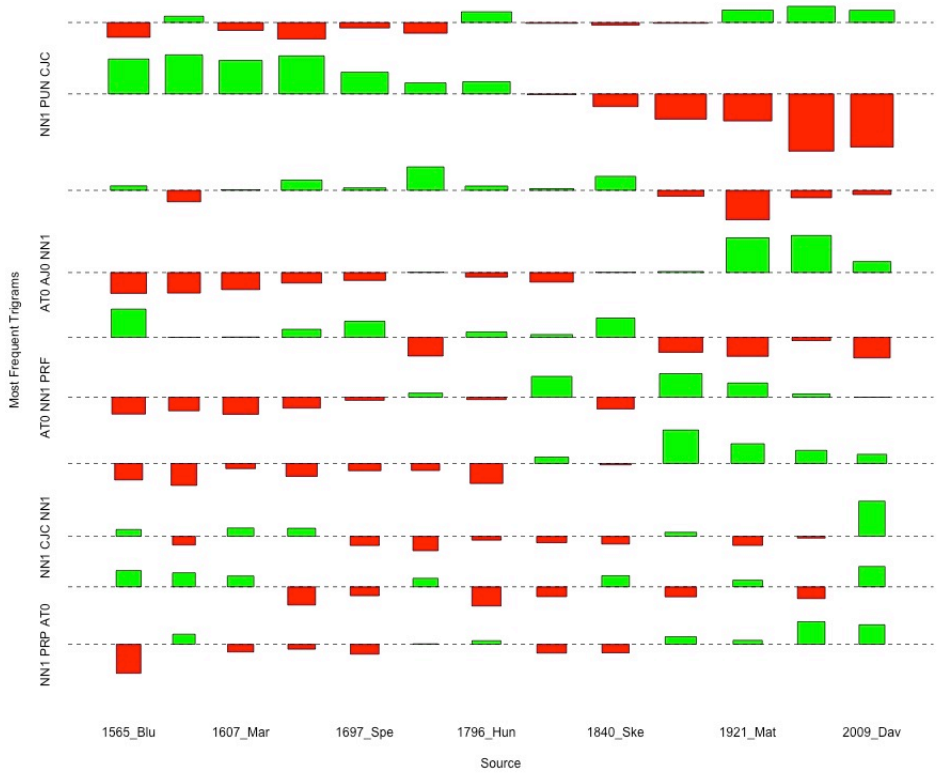


Figure 5.8: Association plot of 10 most frequent POS trigrams and sources

Table 5.5: 10 most frequent POS trigrams in the corpus (n)

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

corpus this tag occurs more or less with expected frequency (i.e., almost no deviation from the centre line), but this trigram 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 seems to be noticeable with respect to the sixth and seventh POS trigram, respectively AT0-NN1-PUN in (19-a) and AT0-NN1-PRF in (19-b):

(19) a. "available minerals in the soil," (Leighton-Hardman, 1977)
b. "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'. As a consequence, 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/PRF, as seen in the example by Leighton-Hardman. The combination illustrated in the text by Matheson, in turn, in occurs inside a prepositional phrase, i.e., "matters [...] which appertain to the welfare of all classes of horses" (Matheson, 1921), and thus also forms a 4-gram of this type. Such combinations, in this case nominal postmodifications strategies of varying complexity, may therefore be particularly indicative of the prose in the later period of our corpus.

Despite this seemingly obvious link in combination of the POS labels of rows 1, 6 and 7, however, the distributional pattern for the trigrams in rows 6 and 7 as seen in the association plot does not tie in with the distribution of the trigram in the first row. In terms of a less-than-expected use in the first period and a more frequent use in the second period, this pattern is not particularly discernible for the potential 'preceding' tag as captured by the trigram PRP-AT0-NN1 in row 1.[32] The distribution of the tag in row 1 rather seems to dovetail the pattern seen for the tag in row 10 instead, which is another tag with which PRP-AT0-NN1 may overlap. Now, however, the trigram combines to its left, e.g., NN1-PRP-AT0, rather than to its right, to form another 4-gram. An example of the combination NN1-PRP-AT0-NN1 may be found in the text by Davies, e.g., "food for the majority of horses" (Davies, 2009).[33] In sum, these findings make it seem likely that a thorough inspection of the tokens underlying the PRP-AT0-

[32]Similarly, although rows 2 (NN1-PUN-CJC) and 9 (PRP-NN1-PUN) show partially overlapping tags, their distributional pattern does not suggest that a large share of these tags are found together in the text samples in our corpus.

[33]Note also that neither the example from Leighton-Hardman nor Matheson comply with this pattern, since these trigrams are preceded by a NN2 and VVB tag, respectively.

NN1 trigram 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 PP headed by the preposition 'of' (PRF) in the current selection of texts.

Lastly, it may be observed that the pattern for the POS trigram in row 10 seems less affected by a periodic disposition, as seen by the somewhat erratic pattern of the bars in figure 5.8. This suggests that the use of this trigram is not particularly conditioned by considerations of style for a certain period. On the other hand, it may be observed that this trigram is in almost complementary distribution to the POS trigram seen in row 3 (AJ0-NN1-PUN), which may indicate that both are part of different (idiosyncratic) strategies or styles of writing. That is, where expected frequencies for PRP-AT0-NN1 in row 1 are negative, expected frequencies for AJ0-NN1-PUN (row 3) are positive, and vice versa. This indicates, for example, that 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. In addition, given the distributional similarity between trigrams in rows 1 and 10, a comparable relationship seems to exist between the use of PRP-AT0-NN1 (row 1) and AJ0-NN1-PUN (row 3) as that between rows 1 and 10. 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 ($\approx 0.14\%$) of all the trigram types found across the corpus.

5.3.5 Additional correspondence analyses

Given previous concerns, two additional correspondence analyses were carried out. One involves a correspondence analysis on the set of text samples and POS trigrams after trigram generation but selects only those trigrams that do not contain tags indicating a punctuation marker. The second additional correspondence analysis concerns a CA on the data set as used in section 5.3.2 on 50% of all tokens, but considers the manuals that are suspected to be outliers (i.e. Speed, and after some consideration, also Davies) as supplementary variables. The results of both analyses can be found in appendix C, but we will briefly summarise their main findings here. For visual inspection, the plots for dimensions 1 and 2 which belong to these correspondence analysis are provided in the appendix as well as at the end of this section (figures 5.9 and 5.10 here correspond to figures C.1 and C.3 in appendix C).

The first thing to be noticed in the analysis on the data set without trigrams containing punctuation mark tags is that a considerable share of the types and tokens in our corpus actually contain PUN tags (either in first, second and/or third position).

Of the original 7,305 POS trigram types found in our corpus, 5,841 were established to be devoid of a PUN tag. Conversely, this means that 1,464 of the types in the data set involve the use of at least one PUN tag (roughly 20.04% of types), representing 16,444 POS trigram tokens. If punctuation markers are found to be a problematic category this is indeed cause for some slight concern, as these numbers indicate that roughly one-third of the attested tokens (32.02%) in our corpus involve the occurrence of such a contested tag.

The correspondence analysis displays the same basic pattern as detected in section 5.3.2: both in terms of dimensional reduction (retention in the first three dimensions is 57.6%), as well as in the general structure that is observed in the cloud of data points in the symmetric plots of dimensions 1, 2 and 3 (see figure 5.9). Although some slight differences may be detected, for example in the total inertia in the solution and a denser cloud of POS trigram points near the origin, the relative distance between text samples is maintained. Although the plots and correspondence analysis in some respect show a clearer visualisation of the underlying data, we give preference to a model that incorporates both POS trigrams with as well as without punctuation tags. This is based on the fact that for the interpretation of results, a richer data set is preferable. For example, deleting POS tags with punctuation markers would ignore frequent patterns that are found in sentence openings (e.g., formulaic phrases that are used to switch or introduce a topic), which seems an undesirable consequence.

The second correspondence analysis on the trigrams with 100 or more observations but without texts that are suspected to be outliers shows that the visual inspection of the plot improves considerably. The relative position of POS trigrams is easier to interpret because the axis are not skewed so much towards the text by Speed, for example. In general, however, these visual displays corroborate the patterns in the original correspondence analyses above. For example, plots of dimensions 1 and 2 show that the POS tags in trigrams in the early part of the corpus often contain coordinating or subordinating conjunctions (CJC or CJS), personal pronouns (PRP), punctuation markers and modal verbs (VM0). On the other hand, trigram labels positioned around the later cluster of texts in the plot often contain labels connected with the use of nominals (cf. Payne & Huddleston, 2002), for example singular nouns (NN1), unmarked adjectives (AJ0), prepositions (PRP and PRF), and general articles (AT0). In addition, VVN tags for past participle forms of lexical verbs are still found in this region of the plot.

Wit respect to dimension 2, there is some suggestion that the positive end of of the vertical axis is now associated with a frequent use of unmarked adjectives (AJ0) and plural nouns (NN2), whereas tags in the negative domain are particularly seen to include singular nouns (sometimes even a combination of two successive NN1 tags).

An relevant question is where the combination of these tags, that is, AJ0-NN1 with a preceding or following POS tag, show up in the plot. Only three of these are seen to occur with a frequency of over 100 observations in this correspondence analysis. These are the combinations of an adjective plus singular noun, followed by either a punctuation marker, an 'of'-preposition or any other preposition (i.e., AJ0-NN1-PUN/PRF/PRP). Interestingly, their coordinates on dimension 2 are fairly low, either with positive or negative signs, suggesting that they are positioned near the origin in the plot. Although we cannot base any firm conclusions on the loadings of merely three POS trigrams, these results suggests a reinterpretation of dimension 2 as a nexus *within* the Late Modern English nominal style, between texts that go in for clusters of singular nouns versus texts that employ a stacking of adjectives (cf. the position of Gibson, and for example the position of the tag AJ0-CJC-AJ0 in the positive domain of this axis in figure 5.10).

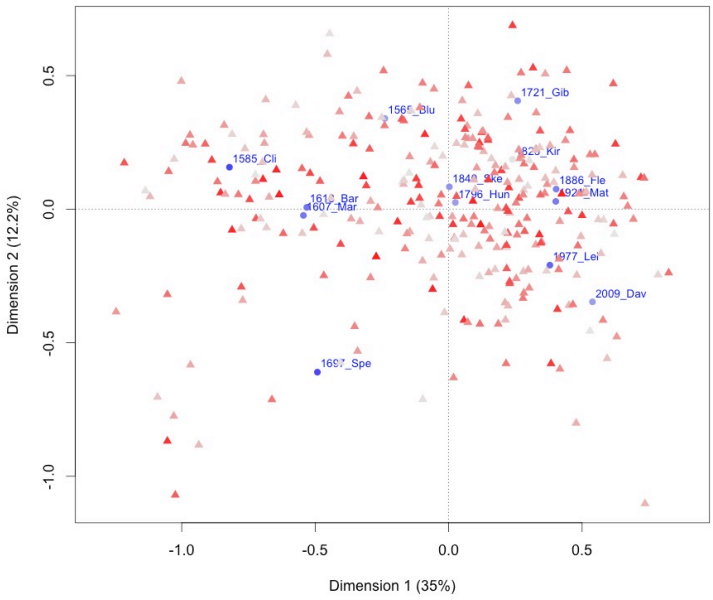


Figure 5.9: Symmetric plot of trigrams and sources (no PUN, 50% of tokens)

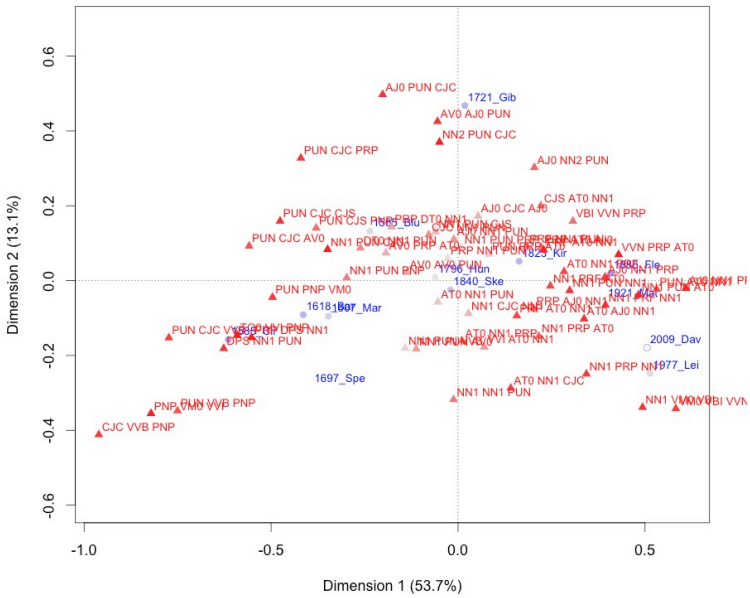


Figure 5.10: Symmetric plot of trigrams and sources (100^+ observations, supplementary variables: Speed, Davies)

5.3.6 Hierarchical clustering

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, p. 50), for example, has suggested the routine use of such hierarchical clustering methods when a correspondence analysis is carried out on a contingency table is. This procedure seems particularly relevant when there is a suspicion that the data is not being represented well due to the dimensional reduction of CA (e.g., either obscuring clusters in a two-dimensional plot which exist in high-dimensional space, or conversely, creating clusters in low-dimensional space that do not reflect the full complexity of the data set). Although there is no particular reason to assume that this is the case in the current data set, we may benefit from the use of HCA to corroborate the visual output obtained in the biplots of the two correspondence analyses in sections 5.3.2 and 5.3.3 above. In addition, we can compare the results of both CAs to see if there are significant differences in the clusters that are created when we apply HCA on

either the rows or the columns of both contingency tables (that is, the text samples or POS trigrams).

As hierarchical clustering of POS trigrams only tells us which trigrams go together, and not which trigrams go together in certain periods (i.e., with certain texts or groups of texts – the CA output is in fact much more suitable for this purpose), we will focus our analyses here on the clustering of text samples in particular. In a second step, we compare the dendrograms obtained on the basis of POS trigram generation to dendrograms based on a hierarchical clustering of the averages of POS tokens (in n per 1,000 words) for the 60 tag classes in these texts, using full text length (and thus without cut-off at 4,000 tokens). First, however, we will consider the two HCAs that go with the data as seen in the correspondence analyses above.

Method

As a note on methodology, 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.

In addition, 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 Tyrrk, 2013). Greenacre 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 (Greenacre, 1988, p. 44). As Baayen (2008, p. 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 assumption best. However, based on the suggestion by Greenacre (1988) to use Ward's method on the data in a contingency table, for comparability we will use this linking method for all hierarchical clustering analyses in this section. In addition, 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, p. 152) and squared Euclidian distance (cf. Greenacre, 2007; Murtagh, 2005).

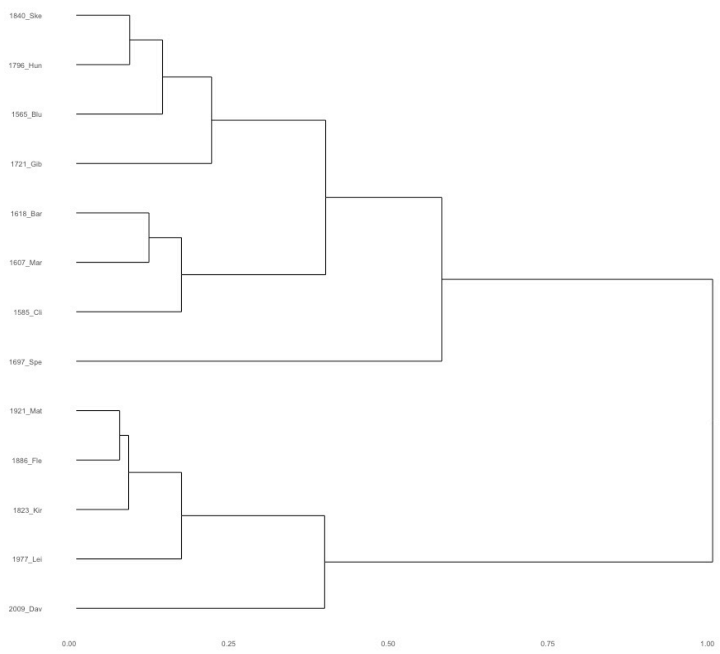


Figure 5.11: Dendrogram on text samples (POS trigrams, 100^{+} observations, method: Ward)

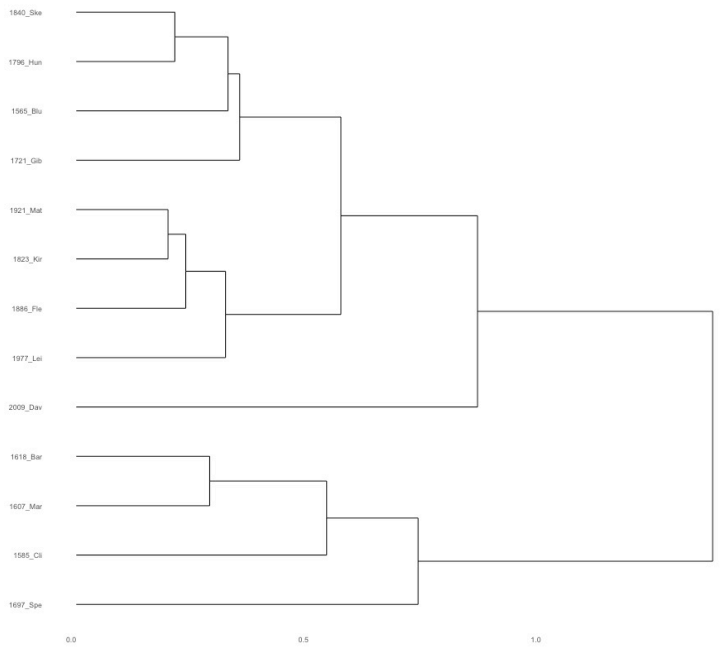


Figure 5.12: Dendrogram on text samples (POS trigrams, 50% of tokens, method: Ward)

Results – POS trigram frequencies

The dendrogram on the contingency table containing POS trigrams with 100 observations or more in the corpus (figure 5.11) 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 19^{th}-century, except for Skeavington (c1840)) and an early group. That the text by Skeavington is grouped with texts published before the 19^{th}-century may be surprising, but this has a correlate in the biplot of the axis with the highest inertias in the corresponding CA (cf. figure 5.6): Skeavington is positioned towards the left on dimension 1 in the plot, and even to left of the origin, where all other manuals published after the 18^{th}-century (in addition to Gibson (1721)) are positioned in in the positive domain of this dimension.

If we look more closely at the early cluster, there is a notable position of the text by Speed (1697), which is merged with the Early Modern group last (and notably, even after Skeavington). The remaining two larger clusters in the early section of the corpus consist of a late 16^{th}-century and early 17^{th}-century cluster of Clifford, Baret and Markham on the one hand, and a more varied group (in terms of date of publication) of the 18^{th}-century texts (Gibson and Hunter) and Skeavington in combination with, somewhat inexplicably, the late 16^{th}-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 can be made light of if we look at a plot of dimensions 2 and 3 (cf. figure 5.7b). In this plot, the manuals by Clifford, Baret and Markham (i.e., the earliest cluster of this HCA) are visible in the top-left quadrant, whereas the text samples of Gibson (1721), Hunter (1796), Skeavington as well as Blundeville (1565) all have positive coordinates on both dimensions 2 and 3, which positions them in the top-right quadrant in figure 5.7b. Although an inspection of the POS trigrams in this area (and on these dimensions) will be required to guide an interpretation of the clustering of these particular manuals, this finding illustrates how an interpretation of the data in a low-dimensional solution of the correspondence analysis may benefit from information conveyed by an additional hierarchical clustering analysis, as suggested by Greenacre (1988).

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 observed. For example, figure 5.12 shows that the two major groups that are distinguished in this cluster solution are now an early cluster in the bottom branch (the texts by Baret, Markham, Clifford and Speed, visible in the previous dendrogram somewhat in the centre of the tree). Note that, as with a crib mobile, the branches of the dendrogram may be turned around without changing the underlying structure of the clustering or the relative distance between text samples (cf.

Oksanen, 2014, p. 6). The fact that the manuals of Davies (2009) and Baret (1618) appear adjacent in this dendrogram, for example, should not be taken as a sign that these texts may be seen as neighbours in some sense. The order of text samples on the left-hand side of the dendrogram does not indicate their relative distance, with the main cue for proximity of the leaves captured in the joining of the branches.

The other large cluster in figure 5.12, the top-branch in the dendrogram, consists mostly of the Late Modern English text samples. However, a separate position for Davies may be distinguished, which is consistent with the picture seen in figure 5.4 of texts positioned on the right of the vertical axis. It is more difficult to distinguish other clusters of texts in this CA biplot, however, for which the dendrogram in 5.12 is again helpful: the larger Late Modern group is seen to consist of two remaining sub-clusters, one for the texts from the 19^{th}- (barring Skeavington) and 20^{th}-centuries, and one for the the 18^{th}-century texts by Gibson and Hunter, as well as the texts by Blundeville and Skeavington. It is additionally observed that the similarity in the use of combinations of parts-of-speech must be quite robust for the texts in this last cluster. In both dendrograms this cluster is built up in the same way, with first Hunter and Skeavington appearing as most similar, then Blundeville being added, and then Gibson being folded in before this entire group is merged with other clusters (either the clearly earlier section in the dendrogram in figure 5.11, or the 19^{th} & 20^{th}-century section in the 50%-of-tokens model in figure 5.12). These findings seem to warrant further study of the typical trigrams used in these four texts, as well as the textual patterns with which these POS combinations are associated. As a starting point for further study, the POS trigrams with positive coordinates on dimensions 1 and 2 in the corresponding CA may be investigated, as these particular texts appear close together in the top-right quadrant in figure 5.4.

Before we continue onto additional clustering, attention is drawn to the fact that the difference between these two plots 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. For example, it may be noted that while an important contribution in the structure of the cluster of Hunter, Skeavington, Blundeville and Gibson may be found in the POS trigrams with higher frequencies, expanding the number of trigrams to include a wider range of possible types does not change the inherent structure of this cluster in the dendrograms. As the hierarchical clustering analysis on 50% of the POS trigram tokens incorporates more of the complexity in the data set, we therefore slightly prefer the clusters found in this HCA over those in the previous agglomerative clustering solution.

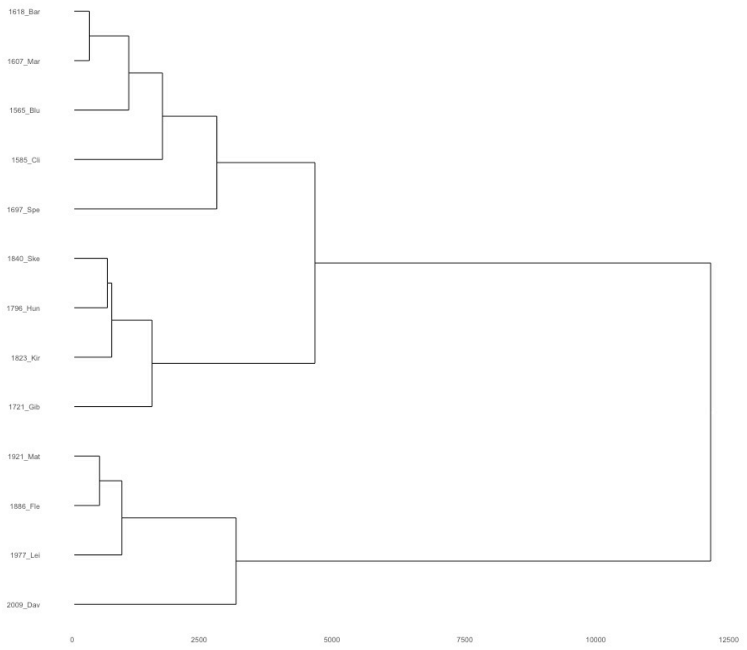


Figure 5.13: Dendrogram on text samples (POS averages, dist.: squared Euclidian, method: Ward)

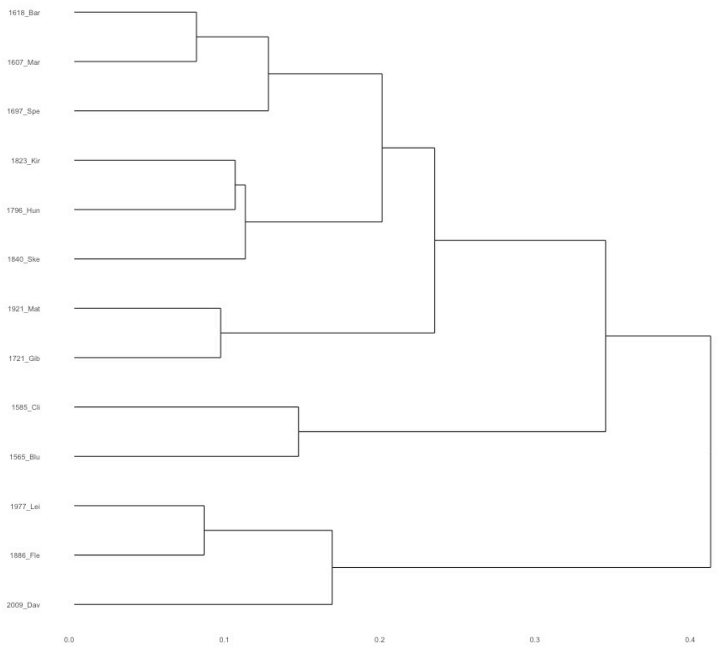


Figure 5.14: Dendrogram on text samples (POS averages, dist.: Spearman corr., method: Ward)

Results – standardised POS frequencies

Using the same linking method (Ward), we can also compare the results obtained by hierarchical clustering analyses on the POS trigram frequencies with another perspective on the same data set. Basing our dendrograms not on frequencies of POS trigrams but rather on standardised average frequencies of POS tags found in these sample texts (n per 1,000 words; see also Tyrrk, 2013, for a similar experiment), we are thus taking into account all of the information in the current data set (and somewhat in contrast to the data on which the previous two plots are based).

An inspection of figures 5.13 and 5.14 shows that the more easily distinguishable clusters can be identified in the former. Three major clusters are visible in the dendrogram in figure 5.13, which is based on squared Euclidian distances.[34] The bottom cluster contains texts published as of the middle of the 19^{th}-century up to the present, with a bit of an exceptional position for Davies as this manual is clustered very late in the HCA. The top half of the dendrogram contains the earlier part of the corpus, with the top branch containing all texts published in the 16^{th}- and 17^{th}-centuries. A bit of a special position here is reserved for the text by Speed, which is the last single manual to be joined with this early group. Given its date of publication this late merger is not entirely surprising (cf. the eight decades since Baret), but what is remarkable is that this HCA picks up on the position of both Speed and Davies as the latest manuals to be joined to other groups. This is noteworthy in light of the fact that these two texts were suspected to be outliers in the correspondence analyses in sections 5.3.2 and 5.3.3, but importantly, our current clustering is based on standardised POS frequencies and not the POS trigram frequencies used above (cf. figures 5.11 and 5.12). There is reason to suspect that the use of parts-of-speech in these texts is therefore markedly different from that of their near-contemporaries.

What remains to be observed in the dendrogram is a Late Modern English cluster that contains the texts of the 18^{th}-century and the first two texts of the 19^{th}-century. Note that the earliest manual in this cluster is somewhat different from the later texts, as the sample by Gibson is clustered last (we have seen the special position of this text in the CA biplots of dimensions 1 and 2 above as well). When comparing this group of four texts with the cluster solution in figures 5.11 and 5.12, however, an interesting difference is observed: whereas the clustering of Skeavington, Hunter and Gibson shows the same pattern as in the earlier two dendrograms, the position of Kirby in the current dendrogram is filled by the 16^{th}-century text by Blundeville in both figures 5.11 and 5.12. Although this is caused no doubt by the particular combinations

[34] Note that an effect of using squared Euclidian distance is that we increase the importance of large distances and conversely reduce the importance of small distances between the standardised counts for the part-of-speech categories in these text samples.

of similar POS trigrams found in these four texts in the previous two dendrograms, the current grouping on the basis of standardised POS frequencies seems to be far more in line with intuitions regarding periodic styles of writing.

If we now turn to figure 5.14, The dendrogram based on Spearman correlations in this case seems to prefer to merge separate leaves which are then folded into other, similar types of smaller clusters rather than iteratively add single leaves to existing clusters. As can be seen, it is far more difficult to discern neat clusters of text samples in this dendrogram.

Nevertheless, a late cluster is visible in the bottom section of the tree in the combining of Leighton-Hardman, Fleming and Davies (the most formal texts?), and another cluster which is clearly discernible consists of the two 16^{th}-century texts by Clifford and Blundeville. Remarkably, a cluster of the early 20^{th}-century text by Matheson and the early 18^{th}-century text by Gibson are added last to the remaining branch of a temporally quite central group of 17^{th}-, 18^{th}- and 19^{th}-century manuals. In the latter group, two minor branches may be distinguished: one for the 17^{th}-century (Baret, Markham and Speed) and one for the (very late) 18^{th}-century text by Hunter and the two texts from the first half of the 19^{th}-century by Kirby and Skeavington. In general, then, the agglomerative hierarchical clustering analysis using Spearman correlations seems to offer quite a number of clusters which are difficult to account for (cf. the clustering of the manuals by Gibson and Matheson, which are exactly two centuries apart).

Many features may be highlighted in a comparison of the four dendrograms just discussed. Here, we highlight only two particular differences, one for the late cluster of text samples and one for a text in the early half of our corpus. For example, one difference in the average POS frequencies compared to the POS trigram frequency distances is the position of the text by Kirby. In both HCAs on the POS trigram frequencies, the cluster of 18^{th}- and 19^{th}-century texts by Matheson, Kirby, Fleming and Leighton-Hardman seems quite stable (cf. bottom branch in 5.11, and a cluster in the center of the dendrogram in 5.12). Although in different orders, Kirby merges with Matheson and Fleming early in the POS trigram dendrogram, after which Leighton-Hardman is added. Conversely, in the dendrograms based on standardised POS frequencies, the manuals by Fleming, Leighton-Hardman and Davies cluster relatively early, but Matheson only joins this cluster in the dendrogram using squared Euclidian distances, and both Matheson and Kirby appear in completely different branches of the tree in dendrogram in figure 5.14.

Several explanations can be forwarded that could partially account for this finding, although these are all tentative and require further investigation. For example, using the full text samples instead of a cut-off at 4,000 POS tokens before trigram generation could point to the possibility that the parts-of-speech found in the remainder of these

text samples not represented after the cut-off cause a significant skew in the data, significantly affecting the calculation of standardised frequencies. Another explanation could be the influence of POS trigram generation, which favours frequently occurring POS and POS trigrams and is less favourable for texts which employ a great number of POS categories to somewhat similar degrees. This is possibly compounded by the fact that we are dropping 50% of trigram tokens (i.e., parts-of-speech in hapaxes, etc.), and is thus unkind to texts with long tails of POS trigram types in particular (that is, texts containing a high number of low frequency trigrams). A third explanation for this difference could lie in the use of difference distance metrics in the clustering algorithm. As was already observed, squared Euclidian distance (on which figure 5.13 is based) has a different effect on the agglomerative hierarchical clustering than the use of Spearman correlations (cf. figure 5.14) or weighted chi-square distance (figures 5.11 and 5.12).

The other difference between both sets of clustering solutions for the text samples that we would like to highlight here is the position of Speed (1697). This manual was identified as an outlier in the correspondence analyses and is shown to appear in very varied positions in the clustering solutions. As an outlier in both POS trigram dendrograms, clustering very late with texts in the early section in figure 5.11 and the penultimate text to join another cluster in figure 5.12, the manual by Speed is relatively unproblematically clustered with the other 17^{th}-century texts in figure 5.14. In figure 5.14 Speed is merged as the penultimate manual, shortly before Davies but in a different cluster, similar to figure 5.12. However, we may note here that this HCA is based on squared Euclidian distance here, which overemphasises large distances in general (see our note above). According to Tyrrk (2013, p. 193fn20), Ward clustering is sensitive to outliers in the data. Since we use this method in all our hierarchical clustering solutions, we cannot estimate the degree to which the clustering behaviour of the manual by Speed is affected by this fact. Nevertheless, the position of Speed remains intriguing in itself: sometimes merged with the very early texts (cf. figure 5.13), sometimes with its contemporaries in the 17^{th}-century (figure 5.14, and marginally in figure 5.12) or as a general late addition to the earliest part of the corpus (in 5.11). Although clustering somewhat earlier, the text by Gibson seems to show somewhat of a similar erratic pattern across these four dendrograms, which could be investigated further in addition.

Lastly, it would have been interesting to study these patterns of merger using Variability-based Neighbourhood Clustering, a clustering algorithm developed by Hilpert and Gries (2009) specifically with diachronic data in mind. Rather than clustering unrelated leaves of the dendrogram, this algorithm only clusters leaves or branches that are temporal neighbours. Although it would have been beneficial to add this procedure

to our current data set for the sake of comparison, this visualisation may have obscured some of the interesting stylistic similarities, as measured by part-of-speech features, in texts by for example the non-neighbours Hunter and Skeavington (early mergers in all HCAs).

Only briefly, we turn towards the clusters of parts-of-speech as found in the dendrograms in figures 5.15 and 5.16 positioned on the last two pages of this chapter. Given the large number of leaves with POS labels, these figures are printed vertically rather than horizontally for ease of inspection.

The clustering of parts-of-speech using squared Euclidian distances shows a considerable difference between singular nouns (NN1) and all other POS categories. However, some interesting clusters can be observed in the remaining branches of the dendrogram. For example, the top branch contains coordinating conjunctions (CJC), unmarked adverbs (AV0), personal pronouns (PNP) and plural nouns (NN2) clustered together, joined by unmarked adjectives (AJ0), articles (AT0) and general prepositions (PRP, barring 'of'). At some distance, punctuation markers (PUN) are also clustered with this group. Although an interpretation of this cluster is highly tentative, the tags in this cluster may indicate a close relationship between regular listings (coordinating conjunctions and punctuation), as well as tags that may be involved in nominal pre- and postmodification.[35]

In the remainder of the dendrogram, we see a branch which includes a wide variety of POS labels (all 50 remaining tags in the tagset). However, there is one larger cluster discernible at the left (top) of this varied branch which includes (not in exact order of clustering) the preposition 'of' (PRF), past participle, base and infinitive forms of lexical verbs (VVN, VVI and VVB) in combination with modal auxiliaries (VM0), coordinating conjunctions (CJS) and general as well as possessive determiners (DT0 and DPS). This last cluster of tags shows some similarity to features seen before, e.g., combinations of modals and lexical verb forms, and may be associated particularly with texts and trigrams with positive coordinates on dimension 2 in the correspondence analysis in §5.3.2. In addition, another smaller cluster of three tags is discernible on the right-most part of the varied cluster, which may be a nexus concerned particularly with the verb 'to be': it contains the tags for the infinitive and -s forms of the verb 'be' (respectively VBI and VBZ) in combination with the to-infinitive (TO0).

Figure 5.16, on the other hand, shows a much more layered clustering pattern. In fact, it is in fact much more difficult to detect relevant clusters in the current dendrogram, which uses Spearman correlations as a distance measure. However, it can be seen that there are some meaningful clusters to be discerned.

[35] Going against a strong indication of the latter interpretation, however, is the fact that other tags that could be associated with this function, e.g., preposition 'of' (PRF) and conjunction 'that' (CJT), occur in another branch of the dendrogram.

Singular nouns (NN1) cluster with the past participle form of the verb 'to have' (VHN), unmarked adverbs (AV0) and articles (AT0) in a larger cluster that also contains wh-determiners (DTQ) and wh-adverbs (AVQ), as well as punctuation (PUN), subordinating conjunctions (CJS) and proper nouns plus non-count nouns (respectively NP0 and NN0). Not entirely surprising, this nominal branch of the tree (the cluster containing tags on the right-side of the dendrogram) also contains VVG tags, i.e., -ing forms of lexical verbs, suggesting the inclusion of (nominal) gerunds in this cluster as well. Remarkably, plural nouns (NN2) are not part of this cluster and instead appear somewhat towards the left of the dendrogram, in a cluster with possessive determiners (DPS), unmarked adjectives (AJ0) and personal pronouns (PNP). The nearest cluster to this second nominal group is a branch that contains some familiar tags as well (see the left-most branch of the dendrogram): coordinating conjunctions (CJC) and prepositions (both 'of' as well as other prepositions; PRF and PRP), in conjunction with lexical verb base and -ing forms (VVB and VVN) and -s forms of 'be'. That modal auxiliaries (VMO) do not not cluster with some of these tags might be surprising, given earlier patterns of association encountered in the current data set. Instead, modals are seen to cluster here with the possessive marker 's or ' (POS) and superlative adjectives (AJS), and no direct explanation for such a clustering comes to mind. After all, it needs to be borne in mind that the patterns of association presented in these hierarchical clustering analyses using the current standardised POS frequencies does not entail that such tags appear close together in the texts of the current corpus.

5.4 Chapter summary

As stated in the introduction to this chapter, our aim is not to make any claims as to a definitive diachronic trend based on the current data set. Rather, the goal here is to try to gain some further insight into diachronic stylistic profiling of instructional prose using data-driven statistical techniques based on frequencies of part-of-speech clusters.

Hierarchical clustering in §5.3.6 and as applied by Tyrrk (2013) serves as a useful method for the characterisation and categorisation of texts, but this method only indicates that a frequent use of one part-of-speech label is associated with another. What it cannot do, for example, is indicate whether singular nouns frequently occur in clusters or certain multi-word expressions, which is why an n-gram approach to patterns in the surface structure is particularly helpful. That we are using POS labels rather than lexical input for n-gram generation and correspondence analyses helps to identify such frequently occurring chunks 'one level below the surface', and avoids interference of spelling variation, among other things. What correspondence analysis offers in addition is a visual display of both angles on the same data: as POS-trigrams across

diachronically ordered texts, and text samples according to frequency of use for POS clusters – greatly enhancing the establishing of associations between both sets.

Although correspondence analysis aids in visualising the richness of the data, the selection of patterns has necessarily remained somewhat eclectic. Selecting patterns based on certain frequencies or statistical cut-off points is one possible heuristic for approaching the interpretation of the wealth of data, but unfortunately neglects many interesting patterns of association. Nevertheless, it seems remarkable that 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. Although minor differences do arise, these patterns are upheld in light of variation based on clustering technique used (i.e., correspondence analysis or hierarchical clustering analysis), different metrics (standardised POS frequencies or absolute counts of POS trigram clusters) and distance measures in hierarchical clustering. It is thus assumed that the patterns found in the current data are robust and represent genuine stylistic differences between the texts in the corpus.

Irrespective of the influence of texts that may be regarded as outliers, the data clearly shows a cluster of texts which corresponds to the Early Modern English manuals in the corpus, as well as a cluster which points to texts published after the second half of the 19^{th}-century. What has remained somewhat underexposed is the position of texts in the centre of these correspondence plots, which largely represents texts composed and published in the 18^{th}- and early 19^{th}-centuries. On the most important axis in the correspondence analysis biplots, the dimension related to diachrony, these texts appear as intermediaries. This reflects the fact that in terms of frequencies for POS trigrams, these manuals also take up a middle position between both the early and late cluster of text samples in our corpus – i.e., not nominal enough to be categorised as texts of the 20^{th}- and 21^{st}-centuries, but not modal, verbal or pronominal enough to rank as texts from the early cluster.

One issue that needs addressing, however, is whether in using the current methodology and experimental set-up, we are exploiting idiosyncratic authorial differences rather than (dis)similarities in periodic-specific prose styles. Although it is highly unlikely that the texts in our corpus are arranged in this way purely by chance and based on factors unrelated to diachrony, the danger of sampling bias is quite apparent, with often only two samples per century for the time span covered by our corpus. Several possible extensions of the current methods present themselves for further study into establishing how the diachronic trace in the data should be interpreted. For example, the use of supplementary data points in correspondence analysis, as applied in section 5.3.5, may test the current solution by adding horse manuals currently not part of the corpus to inspect their position in relation to the text samples currently selected.

An even greater challenge, although methodologically comparable, would be to include texts from different registers but with comparable dates of publication. Although this would introduce the confounding factor of register difference, a variable avoided in the current dissertation as much as possible, it could shed more light on the status of instructive-informative prose as a sub-register itself.

Another avenue for further study would be to divide the current text samples into smaller fragments and carry out correspondence analyses with these texts. The internal consistency of our current text samples may be tested in this fashion, although such an analysis will have to face the problem of potentially very small frequencies.Another option would be to simply select a greater proportion of text from the same source material, largely disregarding the global discourse topics which form the basis of the current text sample selection criteria.

What needs to be borne in mind, however, is that both methods represent *exploratory* statistical techniques only. In general, statistical methods cannot replace theory. Using the techniques described above can only aid in detecting or uncovering latent patterns that we fail to notice with the naked eye. It is maybe particularly for this reason that Jockers refers to the use computational techniques for the detection of trends in text as "distant reading" (Jockers, 2013). Subconsciously, we might detect patterns in the data that we fail to identify when we consciously start looking for linguistic features on which such stylistic intuitions are based. Style is such a phenomenon that is much more amenable to an approach that treats such patterns as a gradient notion, and not a feature checklist used to distinguish between different members of a set. In addition, this is related to the change of perspective which is captured in the oft-cited statement by Sinclair that "language looks rather different when you look at a lot of it at once" (Sinclair, 1991, p. 100).

In addition, it might be argued that for the detection of prevalent procedural sequences which mark the manual by Speed as an outlier (cf. section 5.3.5), for example, we do not need to resort to elaborate statistical or computational methods. However, as Stubbs has pointed out,

> "even if quantification only confirms what we already know, this 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, p. 6)

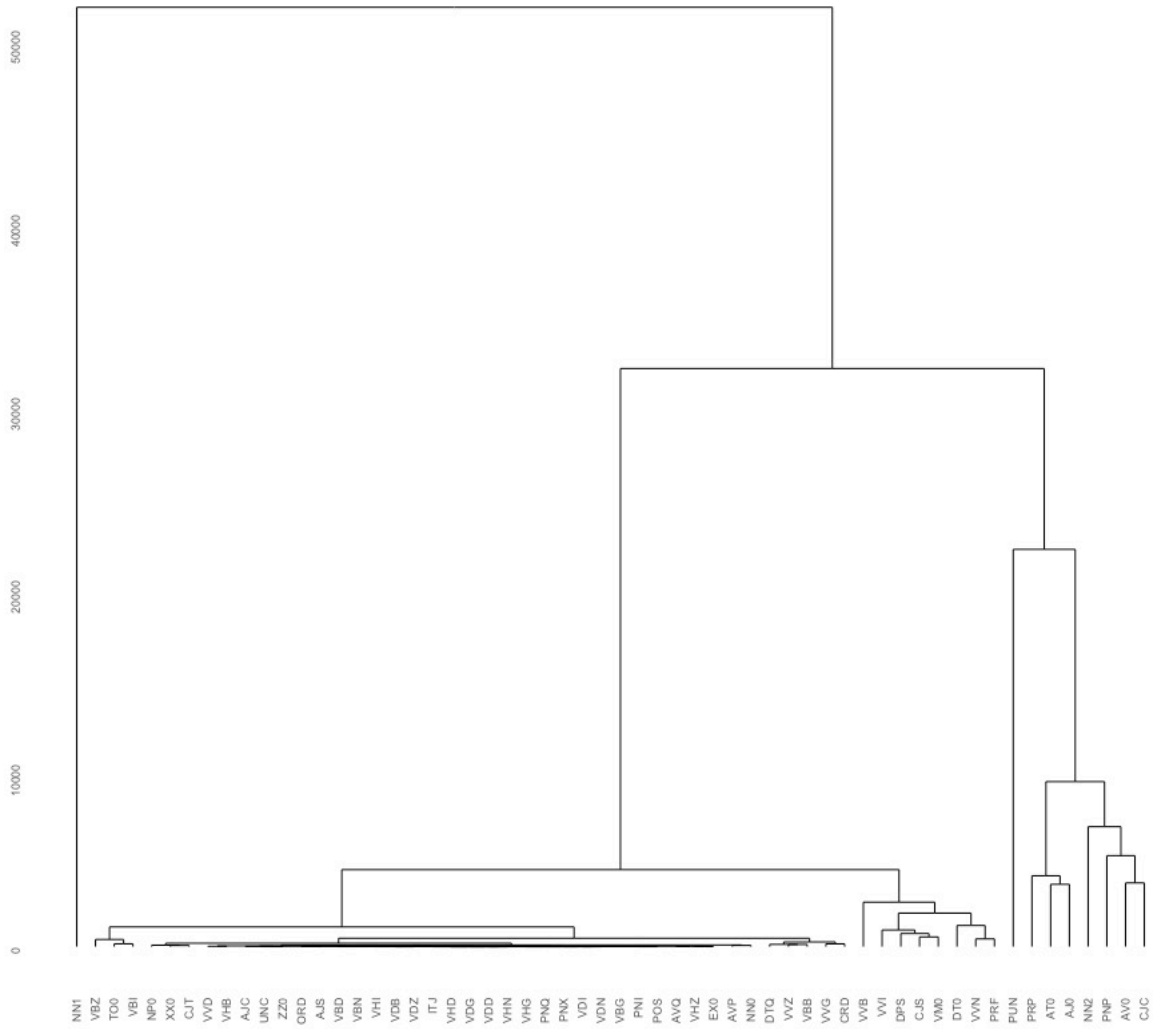


Figure 5.15: Dendrogram on standardised POS tag frequencies (dist.: squared Euclidian, method: Ward)

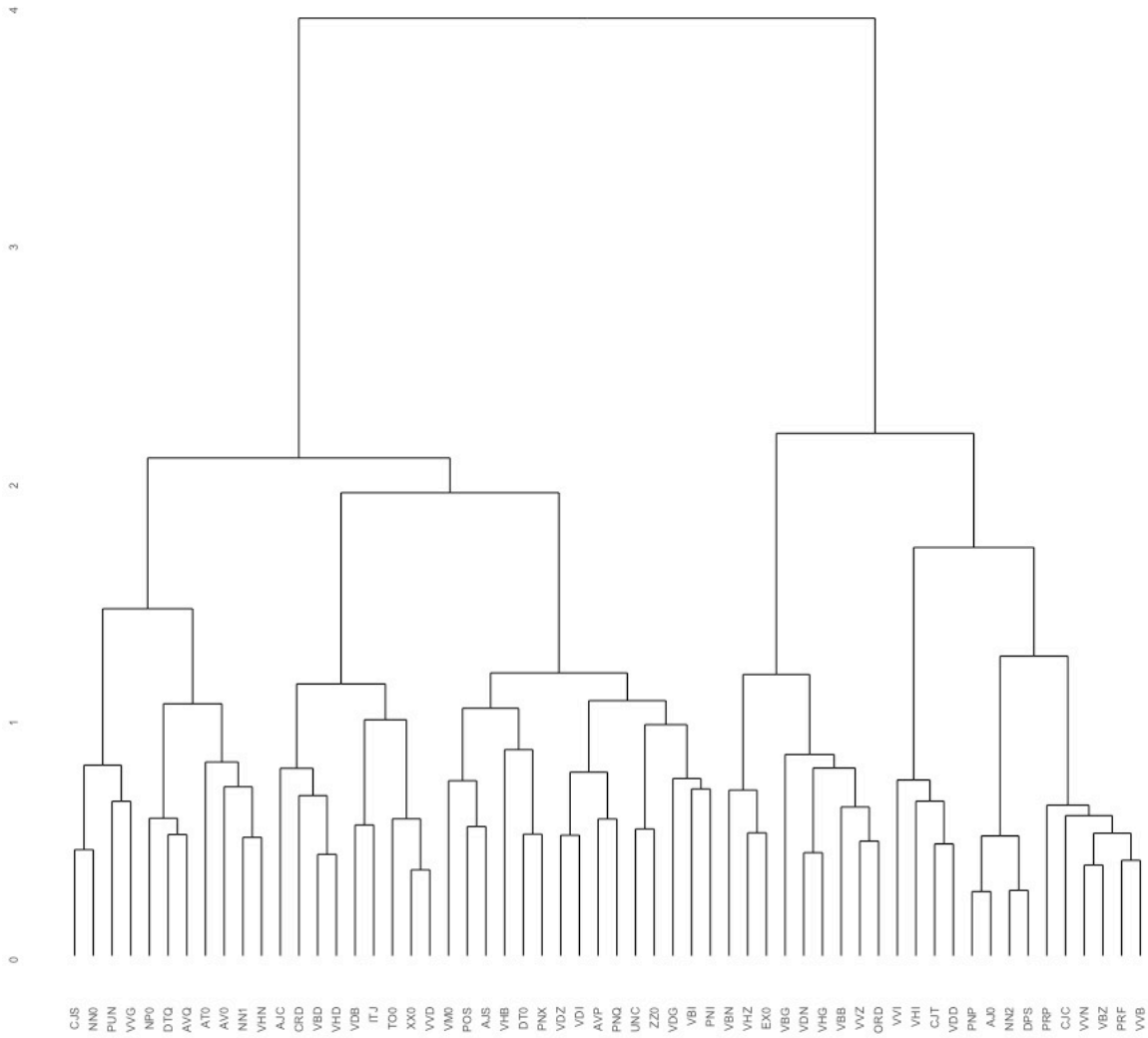


Figure 5.16: Dendrogram on standardised POS tag frequencies (dist.: Spearman corr., method: Ward)

Part III

Discourse Organisation

Chapter 6

Referential Coherence in Instructional Writing

6.1 Introduction: Discourse structure and textual organisation

In his deductive typology of discourse types, Longacre (1996) distinguishes two elementary dimensions which underly a classification of discourses: Agent Orientation and Contingent (Temporal or Chronological) Succession (cf. Diller, 2001; Martin, 1992). The intersection of these binary parameters results in a four-way classification of 'Notional Text Types' (NTT; see Table 6.1). According to Longacre (1996, p. 9), procedural discourse is oriented at "what is done or made, not on who does it". In turn, expository discourse is claimed to be oriented at Themes rather than Agents (Longacre, 1996, p. 9). Note that these two notional text types which are central to the texts in our corpus, i.e., procedural and expository discourse, are primarily distinguished by way of the second dimension in the model, Contingent Succession. Whereas procedural discourse is positive on this feature (i.e. there is a sequence in which the elements of discourse are ordered), expository discourse is organised not temporally/chronologically, but logically (Longacre, 1996, p. 9).[1]

In reality, the definitive categorisation of a text or its constituent parts (let alone a complete subgenre) to one particular Notional Text Type according to textual function or purpose is not without problems, even Longacre (1996, p. 9) admits. Nevertheless, the two parameters in the model serve as a useful stepping stone with regard to referential coherence and (global) discourse progression in this type of prose.

[1] Furthermore, Longacre (1996) recognises other, subsequent binary dimensions (e.g., Projection, Tension) which allow a more fine-grained subdivision of discourse types beyond the most essential classification in this four-way matrix. Although such a subdivision generally provides a more specific characterisation of discourse types, it does not seem relevant with respect to the current discussion.

Table 6.1: Four-way matrix of notional text types and parameters (Diller, 2001, p. 12; see also Longacre 1996, p. 10)

	+ Agent Orientation	– Agent Orientation
+ Cont. Temp. Succ.	Narrative	Procedural
– Cont. Temp. Succ.	Behavioural	Expository

Peikola, Skaffari, and Tanskanen (2009, p. 5) positions instructional writing as belonging to the Behavioural and Procedural discourse text types in the model by Longacre (1996) above. At first sight, such a categorisation seems compelling, particularly given the fact that exhortation is explicitly mentioned as an example of the Behavioural text type by Longacre (1996). However, the suggested categorisation Peikola et al. (2009) also suggests that instructional discourses can appear as –Agent Orientation and +Contingent Succession (i.e., Procedural; see 6.1), as well as the opposite, +Agent Orientation and –Contingent Succession (i.e., Behavioural).

Thus, if instructional discourse can appear to be both containing as well as lacking Agent Orientation *and* Contingent Succession, any of the four notional text types (Narrative, Procedural, Expository and Behavioural) suddenly turn out to be possible surface forms for instructional texts (or more accurately, labels for the characterisation of an instructional text).[2] The suggestions by Peikola et al. (2009) may thus be very useful for a precise description of the various discourse forms in which texts with a general instructive purpose may have appeared in the history of English, but it also severely complicates an accurate description of the "deep structure" (Longacre, 1996, p. 8) of the most appropriate discourse type.

In the current contribution, we take as an important starting point the fact that instructional writing of the form found in contemporary popular (scientific) publications in our sub-genre is characterised by an overall *absence* of Agent Orientation (as opposed to the salient representation or expression of 'Agent' in the Narrative and Behavioural text types). As mentioned before, both Procedural as well as the Expository notional text type are characterised by a general lack of an orientation towards Agents. Given that one of the main foci of the current section is to identify how such an alternative orientation (i.e., –Agent) may be expressed in the grammar, we disregard for the moment the disambiguation along the parameter of Contingent Succession. Instructional

[2]The only way in which the suggestion by Peikola et al. (2009) would stand is if these categories are mutually exclusive: a piece of instructional writing would then either occur as containing Agent Orientation if there is no Contingent Succession, or with Contingent Succession but this blocking the use of Agent Orientation. In this way, instructional texts would occur as an (exclusively) Behavioural text type, or as a Procedural text type only. Such an interpretation cannot be straightforwardly inferred from Peikola et al. (2009), however.

texts are therefore generally characterised as *procedural-expository* prose. This may be appropriate particularly in present-day instructional writing, where the 'how' often seems to go hand in hand with the 'why', possibly to lend credence to the suggested procedure or methodology (barring perhaps the most bare recipe instructions).

In narrative, it is the protagonist that is arguably the logical anchor for the discourse. If there is no a Longacrian-Agent that is the central object of orientation in Procedural and Expository writing, however, one may wonder what it is exactly that constitutes the 'anchor' for the discourse.

6.2 Discourse structure, referential coherence and textual progression

6.2.1 Centering Theory: Referential coherence and focus of attention

Introduction

Centering Theory is a basic model for discourse entity tracking and referential coherence. From its outset, the theory was intended as a formal framework for modelling the attentional state of discourse participants within a global theory of discourse structure by Grosz and Sidner (1986). Centering, initially termed 'focusing', is concerned with (local) focus of attention (cf. Grosz, Weinstein, & Joshi, 1995; Walker, Joshi, & Prince, 1998).[3] The two other components in the model of discourse structure model, intentional structure and linguistic structure, will only receive a brief mention here. The first refers to the level of intentions or discourse purposes of the interlocutors and is related to the rhetorical relations within, and relational coherence of, a discourse. The second, linguistic structure, refers to the sequence of utterances that make up the discourse's actual linguistic realisation. It encompasses the linguistic constructions and devices used, e.g., words and phrases, intonation, changes in tense and aspect (Grosz & Sidner, 1986).

Although attentional state is related to cognitive processes involved in discourse processing, including the inference load placed on the hearer or reader, these matters go far beyond the scope of the current dissertation (the reader is referred to Komen, 2013, p. 25ff, for a discussion and some useful references). However, a brief example might clarify some assumptions regarding attentional state, local focus, perceived coherence and the suitability of centering theory for anaphor resolution.

Centering tries to provide a model of the discourse entities and their linguistic form. One of its key aims is to explain differences in the perceived coherence and inference

[3] Grosz and Sidner (1986) make sure to note, however, that attentional state is a property of the discourse, and is not directly attributable to the discourse participants.

load in discourses such as (1) and (2) (examples are taken from Hudson (1988), as found in Walker et al. (1998, pp. 6-7)):

(1) a. Jeff helped Dick wash the car.
b. He washed the windows as Dick waxed the car.
c. He soaped a pane.

(2) a. Jeff helped Dick wash the car.
b. He washed the windows as Dick waxed the car.
c. He buffed the hood.

Semantic or inferential theories for discourse interpretation as proposed by Hobbs (1985), for example, do not address the degree of incoherence found in such example narratives (Walker et al., 1998). Instead, such theories often allude to the fact that the pronoun 'he' in (1-c) can only co-specify 'Jeff' due to the semantic content of the verb 'soaping', and similarly the pronoun in (2-c) can only refer to the discourse entity represented by 'Dick' due to the nature of 'waxing' and 'buffing'. Centering Theory, on the other hand, bases itself more on the supposed inferential load placed on the hearer or reader. Since the discourse entity referred to as 'Jeff' is at the centre of the discourse in the first two sentences of (2), a sudden-shift to a different discourse entity ('Dick') in (2-c) is unexpected and will cause processing difficulties on the part of the receiver of the message. It is for this reason that example (2-c) is argued to be less coherent than (1-c), and centering offers a formal model that captures such perceived incoherence. Since the theory almost exclusively focuses on the coherence established by the entities populating the discourse model, centering theory should primarily be regarded as a framework for referential coherence (more specifically, entity coherence).

As the examples in (1) and (2) illustrate, it may be clear why CT has received attention for issues of anaphor resolution: pronominalisation is an excellent indicator of local focus of attention in English (cf. Brennan, Friedman, & Pollard, 1987). However, it should be noted that this is a fortunate side-effect, and not the only or primary purpose for which the formal framework of centering theory may be used. In addition, because centering theory tries to make claims that help explain local focus phenomena rather than design specific algorithms to model them (cf. Poesio, Stevenson, Eugenio, & Hitzeman, 2004b, p. 310), from a computational linguistic perspective it is rather seen as a *linguistic* than a computational theory.

General framework

The semantic entities occurring in a text or conversation, that is, the objects, relations or properties linguistically realised as NPs, are termed *centers of attention* (also *centers*,

or simply C) in centering theory.[4] Centers are uniquely associated with an utterance U, which is itself part of a larger discourse segment DS in discourse or text D. As every discourse move (utterance) sets up a number of discourse entities (centers) that can be used in the following discourse, utterance i in discourse segment m of discourse n is said to contain (a set of) **forward-looking centers**: $C_f(U_i, DS_m, D_n)$. For sake of brevity and simplicity, we disregard the level of the discourse segment and overall discourse for the moment, and assume that all utterances consist of the same segment DS_m in discourse D_n.

For every utterance, two additional centering labels are of importance: the **backward-looking center** $C_b(U_i)$, and the **preferred center** $C_p(U_i)$. The backward-looking center, as the name suggest, is the member of the C_f-set that links the current utterance U_i with the directly preceding utterance U_{i-1}. It is often observed that the backward-looking center is the discourse entity with which utterance U_i is most centrally concerned, paralleling the aboutness-topic of Reinhart (1981). However, as will be addressed below, the link that is established between discourse entities in two successive utterances will prove to be more important than this aboutness-feature.

The preferred center, on the other hand, is the forward-looking center of the current utterance U_i that is most likely to be picked up as the backward-looking center of the *following* utterance (i.e., $C_b(U_{i+1})$). The identification of a preferred center thus involves a *prediction* regarding the C_b of the following utterance. In addition, it may be noted that the C_p of the current utterance might coincide with its C_b.

Forward-looking center ($C_f(U_i)$): linguistic realisation of a discourse entity in utterance U_i

Preferred center ($C_p(U_i)$): the highest ranked C_f in utterance U_i

Backward-looking center ($C_b(U_i)$): the highest ranked C_f of the *previous* utterance (U_{i-1}) that is realised in utterance U_i (cf. Constraint 3 below)

The simple three-sentence narratives of (1) and (2), repeated here as (3) and (4), might clarify these notions:

(3) a. Jeff helped Dick wash the car.
[C_f-list:Jeff, Dick, car] [C_b:?] [C_p:Jeff]

b. He washed the windows as Dick waxed the car.
[C_f-list:He, windows, Dick, car] [C_b:He (Jeff)] [C_p:He (Jeff)]

c. He soaped a pane.
[C_f-list:He, pane] [C_b:He (Jeff)] [C_p:He (Jeff)]

[4] In line with the general literature on the topic, we will retain American English spelling when referring to specific components of the theory, e.g.,*center*, instead of British English *centre*.

(4) a. Jeff helped Dick wash the car.
[C_f-list:Jeff, Dick, car] [C_b:?] [C_p:Jeff]

b. He washed the windows as Dick waxed the car.
C_f-list:He, windows, Dick, car] [C_b:He (Jeff)] [C_p:He (Jeff)]

c. He buffed the hood.
[C_f-list:He, hood] [C_b:He (Dick)] [C_p:He (Dick)]

The difference in perceived incoherence alluded to in the previous section can be illustrated by the fact that the link to the previous utterance in (3-c) is also the most likely candidate to be picked up as the backward-looking center based on (3-b). That is, $C_b(U_i) = C_p(U_{i-1})$.[5] In contrast, the labels in (4) show that the backward-looking center in (4-c) is not the center you would expect based on (4-b), leading to incoherence (i.e., the C_p of (4-b) is 'He (Jeff)', whereas the C_b of (4-c) is 'He (Dick)'). According to Walker et al. (1998), it is both this center shift as well as the use of a pronoun for the new backward-looking center that contributes to the perceived incoherence of a discourse. Such center shifting will be explained in detail below.

Identifying the preferred center, or predicting which entity is most likely to occur as the backward-looking center of the next utterance, is based on a (partial) ranking of the forward-looking centers of each utterance. This ranking is guided by language-specific features and/or preferences, reflecting the fact that languages differ in the way discourse entities are marked. Finding an accurate prediction of which discourse entity is most likely to be picked up as the C_b of the next discourse move requires a language-specific algorithm that ranks the forward-looking centers of an utterance. There is a rich literature on the issue of ranking alone, both cross-linguistically as well as to the precise factors affecting ranking in English (see for example Walker, Iida, & Cote, 1994; Di Eugenio, 1998; Turan, 1998; Strube & Hahn, 1999; Pearson, Stevenson, & Poesio, 2001).

As is often observed, the ranking of the C_f-list of an utterance reflects the *discourse salience* of the forward-looking centers of an utterance (Walker et al., 1998). For present-day English, the single most important factor is assumed to be grammatical function, although surface order and information status also seem to affect ranking (cf. Poesio et al., 2004b, for an overview of the literature).[6]

[5] This requirement has been termed the 'Cheapness'-principle by Kibble (2001), following a proposal by Strube and Hahn (1999), and refers to the inferential load placed on the receiver. Cases in which $C_b(U_i) = C_p(U_{i-1})$ obtains are inferentially cheap, as opposed to inferentially expensive cases in which $C_b(U_i) \neq C_p(U_{i-1})$.

[6] Miltsakaki and Kukich (2000) offer a more detailed ranking based on previous crosslinguistic studies (SUBJ >IND. OBJ >OBJ >OTHERS >QIS, PRO-ARB), in which QIS denotes quantified indefinite subjects (e.g., everyone, people) and PRO-ARB stands for arbitrary plural pronominals (e.g., we, you; Miltsakaki & Kukich, 2000).

C_f-ranking by grammatical function: SUBJECT > OBJECT > OTHER

Next to these notions that set up the basic CT framework, there are a number of rules and constraints that are perceived as central in the literature on centering theory. A brief overview of these rules and constraints, as well as their suggested improvements, might prove essential for the application of a CT framework to textual progression.

Centering Theory - rules & constraints

The exact formalisation and interpretation of the rules and constraints of centering theory have received considerable scrutiny in the literature. It is noted that the three constraints and two rules used here, largely based on Walker et al. (1998), apply for every utterance U_i in a discourse segment DS consisting of utterances $U_1, \ldots, U_m$.

Centering Constraints

Constraint 1: There is exactly one backward-looking center $C_b(U_i)$

At first glance, Constraint 1 simply seems to state that every utterance has one backward-looking center. In a footnote, however, Walker et al. (1998) remark that much applied and empirical work to date has interpreted this statement as meaning that there is *not more than one* C_b, allowing for the possibility that an utterance has no C_b at all. On the basis of such interpretations, Poesio et al. (2004b) have proposed a strong ('exactly one C_b') and a weak ('at most one C_b) form of Constraint 1. (The issue of C_b-less utterances will be addressed again in section 6.2.4. For now, it will suffice to remark that there appears to be more empirical evidence that supports the weak version of Constraint 1 than its strong counterpart (cf. Poesio et al., 2004b).) One specific case of the C_b-less utterance, however, is allowed across the board: this concerns segment-initial utterances. Such utterances are not required to contain a backward-looking center by virtue of their not having at their disposal a $C_f(U_{i-1})$-list in the same discourse segment m to which they can refer back.

Note in addition that constraint 1 does not necessarily imply that utterances might not have more than one entity linking back to the previous utterance . Rather, it merely stipulates that the linked entity that has the highest rank in the C_f-list of the current utterance will receive the label $C_b(U_i)$. It seems important to realise that the identification of the backward-looking center thus also crucially depends on the ranking of the C_f-list.

Constraint 2: Every element of the $C_f(U_i)$-list must be realised in U_i

Constraint 2 concerns the set of forward-looking centers that is associated with the current utterance U_i. Although the statement is relatively straightforward, the matter of how *realises* should be interpreted has received some attention (see section 6.2.2).

Constraint 3: The backward-looking center $C_b(U_i)$ is the highest-ranked element of $C_f(U_{i-1})$ realised in U_i

A crucial aspect of Constraint 3 involves the relationship between utterance links and the terms in the C_f-list, both for U_i as well as for U_{i-1}. It was already established that the preferred center of the previous utterance ($C_p(U_{i-1})$) is the most likely candidate to appear as the backward-looking center of the next utterance. Constraint 3 underscores that for the prediction $C_p(U_{i-1}) = C_b(U_i)$ to be true, this candidate has to be realised in the current utterance U_i. Although on the surface this claim is self evident, Constraint 3 can be used to illustrate an important mechanism of CT: if the highest-ranked element of $C_f(U_{i-1})$, and thus the $C_p(U_{i-1})$, is not realised in the current utterance U_i, the label of backward-looking center moves on to the next term in the ranked C_f-list of U_{i-1} which *is* realised in U_i. Thus, any $C_b(U_i)$ has to be part of both the $C_f(U_i)$-list as well as the $C_f(U_{i-1})$-list.[7]

In contrast to the constraints, the rules of centering theory as such do not specify the limits of the framework. Rather, these seem to relate to the predictions of centering in terms of (maximum) coherence. Violations of these rules do not invalidate the framework, therefore, but will instead lead to marked perceptions of incoherence.

Centering rules

Rule 1 ('Pronoun Rule'): If some element of $C_f(U_{i-1})$ is realised as a pronoun in U_i, then so is $C_b(U_i)$

Rule 1, also known as the pronoun rule, is somewhat connected to Constraint 3 in that it concerns itself with realisation. Whereas Constraint 3 is concerned with the fact *if* an entity is realised in two consecutive utterances, however, Rule 1 is more concerned with the *how*; that is, the linguistic form of an entity that features in two consecutive utterances.

At least three influential formulations of Rule 1 have been proposed which all try to capture the intuition that the entity which most significantly links the current utterance to the preceding utterance often has the form of a (personal) pronoun in English. Three formulations mentioned in Poesio et al. (2004b), in order of date of origin, are "If the C_b of the current utterance is the same as the C_b of the previous utterance, a pronoun

[7] We will return to this feature in the discussion of the Cache-model of Walker (2000) in section 6.2.2.

should be used" (Grosz, Joshi, & Weinstein, 1983), "The C_b should be pronominalized" (Gordon, Grosz, & Gillion, 1993), and "If any C_f is pronominalized, the C_b is" (Grosz et al., 1995). These postulations may suffice for short pieces of narrative. A text that centers on the same discourse entity at length, however, may appear unnatural if it only realises this entity as a pronoun and does not vary in the ways this discourse entity is denoted (e.g., resort back to a proper name or NP to describe the protagonist; see §(11) below).The formulation of Rule 1 as it is used here is based on Walker et al. (1998). It is not as strict in terms of pronominalisation, in addition to capturing the notion that more than one center of the previous utterance may be realised in the form of a pronoun in the current utterance. However, if this latter situation holds, then at least the $C_b(U_i)$ must (also) have the form of a pronoun to not lead to some degree of incoherence.

Thus, if an utterance contains more than one pronoun that realises an entity from the previous utterance, one of these pronouns should ideally be the C_b. Example (5-b) in (5) shows that a violation to Rule 1, might lead to a discourse that appears somewhat odd. This utterance contains several co-referential entities, but the only pronoun ('it') is used to refer to a discourse entity with a low C_f-ranking in the previous utterance.

(5) a. Jeff helped Dick wash the car.
`[`C_f`-list:Jeff, Dick, car] [`C_b`:?] [`C_p`:Jeff]`

b. Jeff washed the windows as Dick waxed it.
C_f`-list:Jeff, windows, Dick, it] [`C_b`:Jeff] [`C_p`:Jeff]`

c. He made sure to soap them well.
`[`C_f`-list:He (Jeff), them (the windows)] [`C_b`:He (Jeff)] [`C_p`:He (Jeff)]`

Use of a pronoun in (5-a) for the discourse entity denoted as 'the car' in (5-a) is arguably less problematic if the C_b of (5-b) were to have also been realised as a pronoun (i.e., 'He' instead of 'Jeff').[8] This can be deduced from the next sentence, (5-c), in which both another C_f, next to the one labelled as the C_b, has the form of a pronoun. Although (5-b) leads to some degree of incoherence and increased inferential load, also with respect to the discourse following it, it does not invalidate any of the central claims of CT.

Rule 2: There is a hierarchy of transitions between utterances. The Continue transition is preferred over the Retain transition, which is preferred over the Smooth-shift transition, which in turn is preferred over the Rough-shift transition. In

[8] In fact, the situation with co-referring pronouns for as many discourse entities in (5-b) as possible, that is, for 'Jeff' as well as for 'the car', seems to result in the most fluent narrative. 'Dick' cannot be unambiguously referred to with a pronoun here, however, because 'Jeff' is the preferred-center in sentence (5-a).

other words,

Hierarchy: Continue >Retain >Smooth-shift >Rough-shift

Rule 2 refers to the transition between utterances, particularly between its centers. It tries to capture the intuition that certain center transitions are more coherent than others, and should therefore be preferred in a stretch of consecutive utterances within the same coherent discourse segment.[9] The four canonical transitions are often represented in a matrix (see table 6.2).

Table 6.2: Centering Theory transitions

	$C_b(U_i) = C_p(U_i)$	$C_b(U_i) \neq C_p(U_i)$
$C_b(U_i) = C_b(U_{i-1})$	Continue	Retain
$C_b(U_i) \neq C_b(U_{i-1})$	Smooth-shift	Rough-shift

This matrix basically addresses two questions, (1) is the backward-looking centre of the current utterance also the backward-looking centre of the previous utterance ($C_b(U_i) \stackrel{?}{=} C_b(U_{i-1})$), and (2), is the current backward-looking center also the preferred center (i.e., anticipating the *next* backward-looking center; $C_b(U_i) \stackrel{?}{=} C_p(U_i)$)? The examples in (6) illustrate the four different transition possibilities after a brief introductory narrative.[10]

(6) a. Fred gave Sally a diamond ring.
b. He made her incredibly happy with such a gift.
c. (i) Fred was delighted to see Sally smile. [Continue]
(ii) Sally gave Fred a big kiss to express her gratitude. [Retain]
(iii) Sally rang her mother to tell her the news. [Smooth-shift]
(iv) The dog barked at Sally out of sheer excitement. [Rough-shift]

The utterance in (6-c-i) realises the same discourse entity as backward-looking centre as the C_b available in (6-b): 'Fred' = 'He (Fred)'. In addition, the discourse entity referred to as 'Fred' in (6-c-i) is the highest ranking member in this utterance's C_f-list, and is therefore its C_p. Given that this utterance satisfies both question (1)

[9] It is to be noted, however, that the CT framework as proposed in Grosz et al. (1995) was concerned with *sequences of transitions* rather than singular transitions, and the influential paper by Strube and Hahn, in which the authors propose a Functional Centering framework and outline the principle of transition 'cheapness', similarly makes use of transition *pairs* (cf. Strube & Hahn, 1999).

[10] Using pronouns instead of the proper names for the discourse entities referred to as 'Fred' and 'Sally' in (6) would have made these utterances probably more CT-coherent, following Rule 1. Not abiding by Rule 1 has the advantage of making the explanation of the transitions between utterances, as well as referents, slightly more lucid.

and (2) above, the transition from (6-b) to (6-c-i) can be identified as a Continue transition.[11]

Example (6-c-ii) contains the exact same discourse entities as (6-c-i). The discourse entity identified as the backward-looking center of the previous utterance, 'He (Fred)' in (6-b), also realises the C_b of the current utterance (via the proper name 'Fred') by virtue of being the highest ranked C_f of U_{i-1} following Constraint 3. Thus, $C_b(U_i) = C_b(U_{i-1})$ holds. Based on the grammatical function of these referents, however, 'Sally' is ranked higher than 'Fred' in the C_f-list of utterance (6-c-ii), and is therefore its preferred center (hence $C_b(U_i) \neq C_p(U_i)$). The label Retain for this utterance reflects the fact that although the backward-looking center of the current utterance is retained in U_i, it is no longer the most likely candidate to occur as the C_b of the ensuing discourse.[12]

The Shift transitions, on the other hand, have as their main identifying feature that the current utterance does not have the same backward-looking center as the previous utterance (i.e., $C_b(U_i) \neq C_b(U_{i-1})$). For example, although utterance (6-c-iii) realises an entity from the previous utterance (i.e., 'Sally'), and has this referent as its C_b following Constraint 3, 'Sally' is not the backward-looking center of the previous utterance (6-b). As the highest ranked member of the set of forward-looking centers of current utterance (6-c-iii), however, 'Sally' is both its C_b as well as its C_p. The label for this type of transition is a Smooth-shift, as this utterance shifts the center unmistakably to a different discourse entity (and in particular, one with a grammatical function that ensures a high-ranking in the C_f-list).

Finally, the utterance in (6-c-iv) illustrates the case where neither $C_b(U_i) = C_b(U_{i-1})$ nor $C_b(U_i) = C_p(U_i)$ is true. The discourse entity that realises the backward-looking center in (6-b) is not available in (6-c-iv), nor is the discourse entity from the previous utterance that *is* available in the utterance, 'Sally', its preferred center (instead, it is 'The dog'). Transitions such as these are termed Rough-shifts. They rank lowest in the transition hierarchy introduced in Rule 2, and are assumed to be least coherent of the four classical CT transitions. The fact that they do still establish a link to the previous utterance makes them at least *somewhat* coherent, however, in contrast to specifically the 'Zero transition' outlined below.

Kibble (2001) has insightfully decomposed the essence of Rule 2 and the transition matrix, showing that the first question is concerned with the issue of *coherence*, whereas

[11] Note that the transition between (6-a) and (6-b) cannot be identified as a Continue unequivocally. Although the discourse entity referred to by 'Fred' and 'He (Fred)' seems to denote the same entity, and in similar grammatical roles, the status of 'Fred' as C_b of utterance (6-a) cannot be established without the prior discourse. It is a Center Establishment transition, see below.

[12] Although it has been suggested that the Retain transition announces the introduction of a new discourse topic (Brennan et al., 1987), there has been little empirical support for this claim (cf. Karamanis, 2003, pp. 42-43). We will return to this issue briefly below, in section 6.4.3.

question (2) inquires about salience.[13] Kibble terms these questions the Coherence and Salience Test, respectively, and points out that in the way the transition hierarchy is set up, it can be observed that Coherence is assumed to be more important than Salience ("a stronger requirement" Kibble, 2001, p. 581). That is, the transitions that satisfy the Coherence Test (Continue and Retain) are preferred over the Shifts, after which the Salience Test determines the order *within* the two respective categories (Continue VS Retain and Smooth-shift versus Rough-shift, and in this order). In addition, Kibble has remarked that whereas Salience might be more a matter at the sentence level, Coherence is rather a matter at the level of the text or discourse:

> 'salience' is a matter for sentence planning, choosing a verb form or some other construction that makes the C_b prominent within a clause or sentence, while [coherence] – ordering propositions in a text to maintain referential continuity – is a matter for text planning. (Kibble, 2001, p. 581)

Although this remark was made with reference to the task of Natural Language Generation in particular, this matter will become relevant below where we largely ignore the domain covered by the Salience Test, and build on the distinction offered by the Coherence Test.[14]

Additional transitions Three additional transitions are occasionally mentioned in the centering theory literature. The first, the Center Establishment transition (Kameyama, 1986), was already briefly touched upon with reference to segment-initial utterances. Establishment transitions refer to cases where an utterance with a C_b is preceded by an utterance without a C_b. Such transitions are typically encountered at the start of a discourse, or after a discourse segment switch (DS_t to DS_u, for example).[15] If we allow for

[13]Note that Kibble (2001) speaks of 'cohesion' with regard to the question whether the C_b stays the same over two utterances, rather than of 'coherence'. Since there is no reason to assume that 'cohesion' in this context means anything other than referential coherence as outlined by Sanders and Maat (2006), we avoid using the term proposed by Kibble (2001) here. One reason for the use of 'cohesion' rather than coherence could be to distinguish between the notion of perceived textual coherence (global) and to the referential coherence between two entities in a discourse (local), but this is not spelled out in Kibble (2001).

[14]That the setting up of a link to the previous utterance might be a discourse rather than a sentence planning phenomenon might not be a new insight. Firbas (1964) already commented on the discourse (in)dependence of the notion of theme with reference to earlier Prague School work, and Halliday (1967) seems to have made claims to the same effect regarding his conceptualisation of Theme. However, what is interesting is that this differentiation within centering theory not only addresses the issue of text planning, but also tries to identify (or make claims as to) how the planning at the level of the *current* utterance might foretell which element will be picked up as the link in the *next* utterance.

[15]It is noted that in the literature, such segment-initial utterance transitions are usually considered a type of continuation (as suggested in Walker et al., 1994) and grouped with the coherent transitions Continue and Retain rather than with the shifts, by those who do not make use of a specific label for the Establishment transition. However, Poesio, Stevenson, Eugenio, and Hitzeman (2004a) advise

the possibility that utterances can indeed contain no C_b, two other combinations next to the Establishment transition are theoretically possible. These transitions are sometimes referred to as Zero and Null. The additional transitions outlined here, perhaps with an exception for the Establishment transition, are generally assumed contribute even less towards coherence than the canonical framework's Rough-shift transition.

Zero refers to a C_b-less utterance that follows an utterance that *does* contain a backward-looking center, whereas Null refers to the case where a C_b-less utterance follows another C_b-less utterance (Passonneau, 1998; Poesio et al., 2004b). Example (7) serves to illustrate such 'incoherent' transitions.[16]

(7) a. Fred gave Sally a diamond ring. [-]
`[`C_f`-list:Fred, Sally, diamond ring] [`C_b`:?] [`C_p`:Fred]`

b. He made her incredibly happy with such a gift. [Establishment]
`[`C_f`-list:He (Fred), her (Sally), gift] [`C_b`:Fred] [`C_p`:He (Fred)]`

c. Fred's mother called around after lunch. [Zero]
`[`C_f`-list:Fred's mother, lunch] [`C_b`: - ] [`C_p`:Fred's mother]`

d. The ring was discovered to be worthless. [Null]
`[`C_f`-list:ring] [`C_b`: - ] [`C_p`:ring]`

As was already noted above, the transition between (7-a) and (7-b) is identified as an Establishment transition due to the fact that the C_b of (7-a) cannot be established. If we now take a look at utterance (7-c), the inverse obtains: although the previous utterance (7-b) has a backward-looking center ('He (Fred)'), and the current utterance (7-c) has a preferred centre that looks forward to the next utterance (its C_p is 'Fred's mother'), it lacks a backward-looking center itself. Following Constraint 3, any single discourse entity that is part of the C_f-list of the previous utterance and that is realised in the current utterance, *has* to be its C_b. This situation does not obtain here, however, as none of the referential entities in the C_f-list of (7-c), i.e., 'Fred's mother' and 'lunch', are available in the list of forward-looking centers of (7-b). As a result, (7-c) has to be identified as a C_b-less utterance, while the transition is labelled as Zero. The last utterance, (7-d), typifies the Null transition. This transition denotes C_b-less utterances following another utterance for which no C_b can be identified.

Establishment Utterance with a C_b follows an utterance without a C_b

against such a modus operandi, stating that in their corpus, Establishments seem to behave more like shifts than continuations.

[16] Although the lack of entity coherence is apparent in the narrative in (7), the succession of utterances appears not to be entirely unintelligible. No doubt in part due to semantic and situational inferences that may be drawn, it also shows that entity coherence, and CT-coherence in particular, is only one aspect in the perception of discourse coherence and cohesion.

Zero Utterance without a C_b follows an utterance with a C_b

Null Utterance without a C_b follows an utterance without a C_b

In addition, the examples in (7) also show that it need not strictly be the case that a C_b-less utterance may not contain *any* forward-looking centers (i.e., no discourse entities whatsoever), although this is a pattern attested with some regularity. The only requirement for the Zero and Null transition is that no co-referentiality can be established between the elements of two consecutive C_f-lists, as is the case for both (7-c) and (7-d). These utterances are referentially 'unlinked', so to speak.

Seen from this angle, the Zero transition is merely a special case of an 'unlinked' utterance transition, as it denotes a transition for which there is no correspondence between consecutive C_f-lists (but one for which one of the centers of the previous utterance has a special, namely backward-looking, status). It might be for this reason that the Zero and Null transitions are sometimes grouped together under the label 'NoCB' in the literature (e.g., Friedrichs & Palmer, 2014). We reject this label on two grounds, however. First, the label 'NoCB' as it is used in the literature is not a label for a transition as such, but rather for the utterance under consideration. Once the transition between utterances is taken as the element of analysis, it seems only appropriate to distinguish between Zero and Null transitions. Second, it generally seems to make sense to consider as wide a range of transition classes as possible when applying the framework to actual corpus data for practical reasons. If deemed necessary, any additional collapsing of categories can after all still be applied post-hoc.

Lastly, we take discourse segment-initial utterances (such as (7-a)) to be a special case of the unlinked transition class. Such utterances may or may not contain a set of forward-looking centers themselves, but given that there is formally no previous utterance available, the C_f-list of the previous utterance is necessarily empty, resulting in another case of an unlinked, and therefore C_b-less, utterance. In contrast, Walker et al. (1994) take discourse-initial utterance to contain a C_b, but one which is underspecified and only established upon processing of the *second* utterance.

Our categorisation rather somewhat resembles proposals by Strube and Hahn (1996) and Kameyama (1998). Strube and Hahn distinguish between context-bound and context-unbound discourse elements in the ranking of the C_f-list in their *Functional Centering* perspective. Context-bound discourse elements are those elements in $C_f(U_i)$ that are also part of $C_f(U_{i-1})$, whereas unbound elements are part of $C_f(U_i)$ but not of $C_f(U_{i-1})$. The authors do not specifically Kameyama in turn distinguish between 'chaining' and 'establishing' in the context of CT transition categorisation, with Continue and Retain resorting to the former, and the Establish transition as well as the Shifts being identified as part of the latter category (Kameyama, 1998, p. 94).

As our current model takes Shifts to express a connection between two utterances in some sense (as there still is a correspondence between their C_f-lists), the categorisation 'linked' (Con, Ret, R-s & S-s) and 'unlinked' (Zero, Null) seems more appropriate, and roughly corresponds to the labels for 'coherent' and 'incoherent' transitions already used above. Note that the Establishment transition should also be considered a linked, and coherence enhancing, transition in such a perspective, but that discourse-initial segments are not (they are categorised here as Null). Table 6.3 outlines how the additional and canonical CT transitions can be categorised based on the availability of a backward-looking center in the previous utterance and the current utterance.[17]

Table 6.3: Additional CT transitions

	'Previous C_b' ($\exists C_b(U_{i-1})$)	'No Previous C_b' ($\nexists C_b(U_{i-1})$)
'linked' ($\exists C_b(U_i)$)	Con, Ret, R-s, S-s	Establish
'unlinked' ($\nexists C_b(U_i)$)	Zero	Null

Empirical evaluation of CT has suggested that the three transitions outlined here need to be taken into account when applying a CT framework to naturally occurring discourse – especially when the data consists of non-narrative text domains (cf. Poesio et al., 2004b). As corpus-based studies that make use of the CT framework have helped to underscore (see §6.2.4), the attestation of these 'incoherent' transitions depends partly on a number of parameter settings which are often left unspecified in theoretical papers on centering. Three underspecifications that may be of relevance for the current study will be discussed in section 6.2.2. Next to utterance domain demarcation and realisation of discourse referents, the issue of discourse segmentation might be said to be particularly interesting in light of the additional CT transitions outlined here.

6.2.2 Corpus-based studies on Centering Theory: Testing, evaluating and specifying implicit parameters of CT

Empirical studies that have sought to test theoretical claims as laid out in the general Centering Theory framework have shown that support for its rules and constraints crucially depends on a number of implicit settings usually left unspecified in the literature.[18] Speyer, for example, notes that only about a quarter of the utterances in his corpus of German prose were suitable for his research purposes when applying a strict realisation of centers as based on the standard examples in the literature (that is,

[17] For maximum clarity, the symbols $\exists$ and $\nexists$ express 'there exists' and 'there does not exist', respectively.

[18] Walker et al. mention some of these underspecifications under the header of 'open issues' in centering, especially in the section on utterance-level issues (cf. Walker et al., 1998).

simple, isolated NPs; Speyer, 2007, p. 91). The author concludes that prior to applying centering on natural occurring discourse, the researcher needs to make a number of preliminary decisions, e.g., whether or not allowing non-nominal centers of attention, whether centers may occur in embedded utterances, adhering to direct or indirect realisation, the distinction between linear versus hierarchical recency (cf. Walker, 2000) and the extent to which lexical and/or semantic similarity is required between co-referential centers (Speyer, 2007, p. 92).

The main aim pursued by Poesio et al. (2004b) is exactly to empirically evaluate the instantiations (i.e., settings) of a variety of these implicit CT parameters using a corpus-based methodology. For example, although we have so far attempted to remain neutral as to the interpretation of the term 'utterance', it is not entirely clear whether this underspecified term corresponds best to the level of the sentence, the finite clause, a tensed clause or a T-unit (cf. Fox, 1987) in the literature. Although in part driven by the aim to establish which configuration results in the most optimal CT-setting (that is, specify the parameter instantiations in such a way as to result in the most coherent discourse, viz. a discourse that poses least processing difficulty on the part of the reader or hearer), one of the key findings of Poesio et al. is that changing parameter settings necessarily results in a trade-off between different metrics of 'coherence' as based on the various centering rules and constraints. For example, changing instantiations so as to decrease violations of Rule 1 will lead to more violations of Constraint 1 and Rule 2 (Poesio et al., 2004b). In light of these findings, it seems appropriate to specify a number of CT parameters that are relevant to the current research purposes.

Discourse segmentation

As a result of the theory's explicit restriction to utterances *within* discourse segments, Walker et al. have remarked on the general shortcoming that "any claim of the theory cannot be tested on two utterances that span a discourse segment boundary" (Walker et al., 1998, p. 20). The demarcation of discourse segments is crucial for the validity and applicability of both rules and constraints, however. Given the formal character of the CT framework, it is somewhat remarkable that there has been relatively little consensus on the formalisation of discourse segments; or alternatively, for how a segment boundary is to be formally recognised in a discourse.

Segmentation is said to be dictated by the component of intentional structure in the discourse structure model of Grosz and Sidner (1986). As a consequence, discourse segmentation is taken to be a feature that lies *outside* the domain of attentional state underlying Centering Theory. However, Grosz and Sidner (1986) have also not explicitly specified *how* discourse intentions may be used to identify segments in a discourse (Poesio et al., 2004b, p. 325), only that it should be possible to identify for

each segment a single discourse segment purpose (DSP; Grosz & Sidner, 1986, p. 178). Although Walker et al. (1998, p. 24) mention that the cache-model by Walker (1998) allows a somewhat looser relationship between intentional structure and attentional state with respect to discourse segmentation and discourse structure in general, Passonneau and Litman (1993) has shown that it is already difficult to identify discourse segment boundaries reliably.

Preliminary attempts in the CT literature have sought to provide some heuristics for discourse segmentation. Two of the simplest methods are to either regard the whole text as one single segment, or to use layout features such as paragraph or section boundaries as an indicator of a discourse segment boundary. Another method, suggested by Walker (1989), is to regard each paragraph as a discourse segment unless its initial sentence contains either a pronoun as subject, or a pronoun "whose agreement features are not matched by any other C_f in the same sentence" (Poesio et al., 2004a, p. 26).

The method used by Passonneau (1998), on the other hand, is informed by the set of canonical centering transitions, and tries to asses if and how such transitions correspond to a discourse segment boundary. In other words, it is attempted to establish how accurate centering transitions are in serving as cues for discourse segmentation. Passonneau (1998) has naive subjects annotate segments boundaries in the spoken narratives of the *Pear Stories* (Chafe, 1980) using two separate three-way transition classification schemes (one with Continue, Retain and Shift, and one with Continue+Retain, Establish and Null; cf. Passonneau & Litman, 1993). Although it is established that the Shift and Null transitions occur most frequently at boundaries in the two respective classifications, it is also shown that these transitions are attested even more often *within* segments (i.e., at non-boundaries) in the annotated narratives.[19]

To study the correlation between segment boundaries and centering transitions, Poesio et al. (2004a) extend the method pursued by Passonneau (1998) and make use of all seven utterance transitions as outlined above.[20] The authors conclude that the correlation between discourse segmentation and the coherence status of transitions is not perfect, the raw data suggest that Establishment, Zero and Null transitions are indeed more frequently associated with boundaries than the four classic transitions (cf. Poesio et al., 2004a, p. 65). Nevertheless, it is decided to collapse the transitions in two categories in such a way that the shifts are categorised with the incoherent transitions. This decision seems somewhat remarkable in light of the empirical data, i.e., the figures

[19]Despite the use of two different classifications, the experiment seems to hinge on the assumption that there is a connection between segment boundaries and shift transitions. Given the fact that even the Rough-shift transition still retains some link to the previous discourse, namely in the form of a lowly ranked but nevertheless available C_b as indicated above, however, we argue that such shifts may not be the strongest segment boundary cue.

[20]Since the corpus used by (Poesio et al., 2004a) is not annotated for segments by naive subjects, it is assumed here that segmentation is based on text layout features.

in the table on which this decision is based, but the authors regard this as the best collapsing strategy available (cf. Poesio et al., 2004a, p. 68).[21] About a third to a quarter of the incoherent transitions occur at boundaries (32% for Establish, 32% for Zero and 26% for Null, respectively), whereas these numbers are decidedly lower for other transitions (19% of Continues and 12% of Retains occur at boundaries; the χ^2-test with 6 degrees of freedom is shown to be highly significant, $\chi^2 = 39.1, p \leq .001$). In fact, it can be seen that shifts occur least frequently at boundaries (7% of Smooth-shift and 12% of Rough-shifts, equivalent to 4 and 9 cases) and well below the overall average of 24% transitions occurring at boundaries. The resulting collapsed table reported a χ^2(df=1) = 11.1 at $p \leq .001$, whereas a χ^2-test with the shifts collapsed with the Continue and Retain transitions reports χ^2(df=1) = 31.84 at $p \leq .001$ (cf. Table 6.4; Pearson chi-squared with Yates' continuity correction).

Table 6.4: Collapsing transitions (based on Poesio, Stevenson, Eugenio, & Hitzeman, 2004a, p. 65-66)

	Non-boundary	Boundary	Total
Con+Ret+SSH+RSH	313	47	360
Zero+Est+Null	434	178	612
Total	747	225	972

Using a variety of CT settings the authors conclude that, like Passonneau (1998), their method has failed to establish that transitions are accurate indicators of segment boundaries. However, they also report that their tests indicate that it is also "very unlikely" that centering transitions and segment boundaries are completely *unrelated* (Poesio et al., 2004a, p. 68).

Using the paragraph boundary as a heuristic for discourse segmentation, Friedrichs and Palmer (2014) also compare the distribution of utterance transitions within and across paragraphs. The authors observe that more than half of the paragraph-initial sentences in their large-scale corpus contains a discourse entity that is realised in the final utterance of the previous paragraph (Friedrichs & Palmer, 2014, p. 139).

Despite the assertion by Walker et al. (1998) that Centering's claims cannot be tested across discourse segment boundaries, it thus seems ill-advised to apply a strict separation of centers across discourse segment boundaries. Walker (2000) completely

[21] This table is based on indirect realisation with U=finite clause and GFTHERELIN ranking. In the immediately preceding section §5.2, on the relationship between the type of transition and the linguistic realisation of the subject (e.g., pronoun or full NP), the authors do in fact categorise the Zero and Establishment transition with the Shifts. It is stated there that Establishments behave more like Shifts than Continues (contra Walker et al., 1994), at least with respect to this feature, and the successive collapsing strategy increases the χ^2 value for most statistical tests in the section (Poesio et al., 2004a).

abandons the within-discourse segment restriction for the retrieval of antecedents, for example, which is made possible by her proposal of a cache, instead of a stack, model of attentional state (see Walker, 1996, 2000).

Without a resolution of this theoretical issue and the resulting ramifications, we remain somewhat agnostic as to the precise status of segmentation in relation to transition type. The restriction on CT rules and constraints to apply only *within* discourse segments (DS) is thus somewhat relaxed (it may rather be in effect at the level of the discourse, D). However, we do suggest that it may be relevant to establish which transitions correspond to text layout cues that may indicate segmentation, such as paragraph and section boundaries. In addition to deciding the correct label for an utterance transition, a tag for either 'boundary' or 'non-boundary' is added therefore. The incoherent transitions of Establishment, Zero and Null are of particular interest in this light, as these are not normally taken into account in large scale corpus-based studies of CT barring Poesio et al. (2004b) and Friedrichs and Palmer (2014) on contemporary English.

Realisation

Direct and indirect realisation Realisation refers to the way in which a discourse entity is linguistically realised at the level of the utterance. Although Grosz et al. (1995) point out that an exact interpretation of realisation depends on the semantic theory one adopts, two broad forms of realisation may be recognised: direct realisation and indirect realisation (Poesio et al., 2004b). *Direct realisation* refers to cases in which a noun phrase in the current utterance encodes a discourse entity which is already in the mental discourse model. *Indirect realisation*, on the other hand, refers to cases in which a noun phrase in the current utterance does not realise such a discourse entity explicitly, but rather realises an entity which stands in an associative relationship with a discourse entity already in the mental discourse model (i.e., a form of bridging reference; cf. Clark, 1977; Clark & Haviland, 1977; Irmer, 2009).

Allowing indirect realisations in a CT context effectively means expanding the possibilities for establishing links between utterances to, for example, entities that are elements of a partially ordered set (POSET; cf. Ward, 1988; Birner & Ward, 1998; Ward et al., 2002) or functional dependency relation (cf. Walker et al., 1998). Conversely, restricting realisation to only direct mentions may result in large proportions of incoherent transitions, as C_bs may only co-refer discourse entities which enter into a direct *identity* relationship across an utterance boundary.

Walker et al. note that a restriction to direct realisation, that is, observing a strict identity relationship between referential entities, in all likelihood "misses important generalizations" in terms of coherence establishing coreference through indirect rela-

tions (Walker et al., 1998, p. 5fn4). Indeed, Poesio et al. (2004b, p. 325) state that indirect realisation "can play a crucial role in maintaining the C_b". Nevertheless, it is suggested here that the identification of two entities which take part in an associative relationship crucially depends on information not directly available in the co-text, but rather in the cultural context and/or store of real world-knowledge of the interlocutors (see for example Givn, 1992, p. 12ff.). Such relationships thus contribute to the (perceived) discourse coherence only if the hearer or reader is able to access or retrieve this information (see also Kehler, 2004).

Form of the referring expression Third person pronouns (*they, it*) may be considered to be the easiest form of encoding an identity relationship between referential entities in a discourse. Given the particular non-narrative nature of the discourse under study, however, it also needs to be established if and how such co-referential link between referential entities may be expressed by way of (complex) NPs. We take the *restriction of a mental entity* by way of modificational strategies as crucial in determining whether any identity or associative relationship exists between referential entities that are expressed by indirect realisation (and with particular attention for the utterance transition expressed). See the examples in (8) and (9) (note that the C_b for both initial utterances is established on the basis of the previous discourse, and pre- and post-modification of complex NPs is indicated in the C_f-list by square brackets for relevant referential entities).

(8) a. Horses need more protein when tissue is being laid down for growth [...].
`[`C_f`-list:Horses, protein, tissue, growth] [`C_b`:Horses] [`C_p`:Horses]`

b. Mature horses will most likely do fine on a lower protein percentage [...], depending on their workload. [Continue]
`[`C_f`-list:[Mature] horses, lower protein percentage, workload] [`C_b`:horses] [`C_p`:horses]`

c. Horses that are in intense training need more protein than the maintenance horse because they are developing muscle tissue; [Continue]
`[`C_f`-list:Horses [that are in intense training], protein, maintenance horse, muscle tissue] [`C_b`: Horses] [`C_p`:Horses]`
(excerpt from: Duberstein & Johnson, 2009/2012)

(9) a. Newborn foals may contain as much as 90% water!
`[`C_f`-list:[newborn] foals, water] [`C_b`:water] [`C_p`:foals]`

b. Water molecules are however very small. [Zero]
`[`C_f`-list:[water] molecules] [`C_b`:---] [`C_p`:molecules]`

c. Water is essential for: [Null]
[C_f-list:water] [C_b:---] [C_p:water]
(excerpt from: Davies, 2009)

Deictic markers (exhoporic reference) The difference in coherence established by way of direct realisation in the co-text, such as through the use of anaphora and other grammatical cues (e.g., inflectional marking) available in the linguistic string, versus such coherence that depends on world-knowledge and information in the discourse context, is also linked to the interpretation of first- and second-person pronouns. Although Poesio et al. (2004b) initially do not treat such deictic, or "exophoric" (Biber et al., 1998), entities as potential centers in the C_f-list on the basis of suggestions in Walker (1993), it is also acknowledged that second person pronouns "often seem to play an important role in maintaining the coherence of the discourse", especially in instructive texts such as in the pharmaceutical subdomain (Poesio et al., 2004b, p. 342; see also p. 318, 343 and 356). An example which serves to illustrate this can be found in (10). It seems textually coherent but contains C_b-less utterances if exophoric elements are not treated as potential centers of attention, and would therefore be considered CT-incoherent (example (7) from Poesio et al., 2004b, p. 318):

(10) a. You should not use PRODUCT-Z
b. if you are pregnant of [*sic*] breast-feeding.
c. Whilst you are receiving PRODUCT-Z ...

Note that in this context, Givn (2001, p. 459-464) distinguishes between "grounding referents" to the shared current speech situation, the shared generic-lexical knowledge and the shared current text. From a neuropsychological perspective, these three are said to correspond to "working memory, current attention focus", "permanent semantic memory" and "long term episodic memory" (Givn, 2001, p. 460-464).

We take the coherence that is established through such deictic markers available in the discourse context, in conjunction with bridging associations and potential recourse to a global focus entity (i.e., a global discourse topic; see also Poesio et al., 2004b, p. 333), to be outside the scope of Centering Theory proper. Centering is to be purely based on the co-text (i.e., may not resort to deictic markers for the first- and second-person) and can only rely on direct realisation. It is here that the difference between perceived coherence and the coherence that can be captured by the formal framework of CT ('CT-coherence') is particularly marked.

Category of the referring expression In terms of the linguistic category of centers of attention, Centering usually focuses explicitly on noun phrases and personal pronouns as sources of referring expressions. However, (Speyer, 2007) points out that neither Grosz et al. (1995) nor Walker et al. (1998) commit to a specific category for a constituent to serve as a center of attention (as long as it has the ability to refer). For German, this would squarely put referential and pronominal adverbs (e.g., *dann, hier, dafür, davon*) into the category of potential items that may occur in the C_f-list, especially as these contain nominal phrases etymologically (cf. Speyer, 2007, p. 93).

With respect to English, pronominal adverbs or pro-forms (cf. Stirling & Huddleston, 2002) such as *there* in adverbial connectors *therefore* or *therein* do not normally receive attention in the CT literature. In light of the history of English, however, the use of such adverbial connectors seems at least in part motived especially *because* of their special discourse function of establishing coreferential links between utterances (see also Lenker, 2010, 2011). Although the referential nature of connectors such as *therefore* has become somewhat obfuscated in Present-day English, such items deserve scrutiny in the analysis of historical texts – particularly when allowing for non-nominal or propositional antecedents.

Somewhat similarly, demonstrative pronouns (*this, that, these, those*) are usually disregarded in corpus-based applications of CT, even though their referential function can often be established readily. The fact that demonstratives may refer to either nominal or non-nominal (e.g., propositional) antecedents, however, makes their implementation somewhat problematic in a framework that focuses primarily on nominal constituents.A case in point is that although Poesio et al. (2004b) note the relevance of a discussion of the status of such pronouns, they do not support the use of demonstratives as potential centers for the C_f-list on account of a distortion of CT-coherence metrics (specifically, violations to Rule 1). Such considerations need not concern us here: as long as there is an antecedent in the previous discourse, demonstrative pronouns should be allowed to act as centers of attention even when referring to a discourse entity that is not realised by a NP.[22]

Utterance domain

Perhaps even more pertinent than discourse segmentation, at least for practical application of Centering on corpus data, is the demarcation of the utterance domain (for example, in generating a C_f-list). In the early literature on Centering Theory, the utterance domain was implicitly identified as that of the sentence (cf. Poesio et al., 2004b). Kameyama (1998) has questioned the optimality of the sentence as unit for

[22] In practice, such propositional or non-nominal centers will not be indexed as members of the C_f-list utterance for utterance, unless they serve as antecedents.

an update of local focus, however, pointing to problems of ranking the C_fs in complex sentences in written text consisting of multiple levels of embedding. Rather, it is suggested to consider not the sentence but the finite clause as primary unit for local focus update. Another option that is put forward is to consider all verbed clauses (verbal complexes), regardless of finite or non-finite (cf. Poesio et al., 2004b, p. 317).

Examples of each of these suggestions may be found in corpus-based applications. For example, Taboada and Zabala (2008) suggest that the best utterance level domain is the finite clause, based on a corpus study of both contemporary English and Spanish spoken dialogue. In turn, Miltsakaki and Kukich (2004, 01, pp. 30-3) suggest to take subordinate clauses as part of the same utterance as the matrix clause of the sentence, but argue that centers in subordinate clauses should be ranked below those of the matrix clause in the C_f-list. Although there is no explicit mention of identifying the sentence as the utterance domain, the examples in Speyer (2007) use the sentence as realising the utterance domain in the German texts under consideration.

Similarly, in a CT-based corpus study of texts in the history of English (from late Middle to Present-day English), Martnez Insua (2011) does not comment explicitly on the identification of the utterance with either sentence or clause. Although the study is only concerned with *there*-existentials and related constructions, and is thus arguably concerned mainly with the left-periphery, the author unfortunately does not offer any specific details as to how utterances and the utterance domain are identified.

The issue of specifying the utterance domain is highly relevant when taking into account the diachronic focus of the current study. As writing conventions regarding sentence length, embedding and punctuation are known to be subject to change in English (cf. Parkes, 1992; Horobin & Smith, 1999; Smith, 2012; Crystal, 2015), identifying the utterance domain with the sentence may be inadequate. Nevertheless, identifiying centers at the clause level may not do justice to what may be thought of as deliberate efforts in terms of text planning (in the sense of Kibble, 2001) conveyed by the organisation of content into sentences and larger units of text (paragraphs, sections, etc.). As Los et al. (2012), Meurman-Solin (2012) have noted, identifying the correct unit of analysis in studies of discourse structure is still under debate.[23]

6.2.3 Identifying previous utterances, the Cache-model and size of the "backward-looking window"

One of the crucial points for the Centering Theory framework, especially with respect to corpus-based application, is determining what counts as 'previous utterance'. Mitkov (2010, p. 611), for example, argues that limiting the "search space for antecedents"

[23]For the present case studies in sections 6.5.1 and 6.5.1, we will use the sentence as the discourse structure unit, noting that this might be an inadequate metric for the retrieval of centers of attention.

to the immediately preceding utterance, i.e., U_{i-1}, might be one of the major limitations for Centering Theory as a model for discourse processing. Most corpus-based CT studies indeed adopt some form of linear adjacency whilst specifying whether certain intrasentential, embedded utterances are not to be regarded as actual updates of the local focus (e.g. Kameyama, 1998; Poesio et al., 2004b; Karamanis, 2003). Suggestions for such embedded utterances include adjuncts, coordinated clauses and second elements of coordinated VPs, relative clauses and clause complements, for example. The status of these embedded utterances not only affects which entities populate the C_f-list, however, but even more importantly, which entities in the previous discourse are no longer available for subsequent mention, and may thus contribute to (CT-)coherence, due to the nature of the intervening utterance. By extension, this choice directly affects the summary of utterance transitions necessary for an assessment of a text's coherence.

Walker (2000) highlights the fact that the discourse structure model by Grosz and Sidner (1986) already addresses the problem of intervening segments for local focus update. What seems to be of import for finding an antecedent is not whether the previous utterance directly precedes the utterance in question (i.e., is linearly recent), but whether it precedes it on the same, or a dominating, discourse level (i.e., it is hierarchically recent; see also Speyer, 2007). It is thus theoretically possible to refer back to a center across another utterance if that intervening utterance is on a lower or embedded level in the discourse structure it can thus be readily 'popped off the stack' when returning to a dominating discourse level in the stack model by, Grosz and Sidner (1986). Speyer (2007) has called to the fact that the frequency with which such sub-discourses, or focus pops, occur in corpus data warrants a mention (and resolution) in the general CT framework.[24]

Although most corpus-based CT studies have indeed observed some sort of hierarchical recency in terms of intervening utterances since Kameyama (1998), such studies have usually focused on the intrasentential level and the particular syntactic properties of utterances (i.e., clauses) which are to be regarded as embedded (on the basis of whether they do or do not provide viable antecedents/C_fs), rather than on the level of the discourse structure and complete (sequences of) sentences that may constitute an intervening discourse level.[25] Limiting the 'backward-looking window' to such a restricted anterior domain, across a number of embedded clauses in the case of complex sentences but never across a full sentence, significantly affects centering metrics such as utterance transition frequencies. For example, if the clause introduced by the

[24] "Solche Einschübe sind hinreichend häufig, dass sie eine Erwähnung im Regelwerk der Centering Theory verdienten" (Speyer, 2007, p. 95).

[25] Grosz and Sidner (1986) speak of complete (series of) discourse *segments* that may be popped, but we refrain from this term here because an interpretation of their discourse segment is more elaborate than can be provided here, or in what was outlined in 6.2.2.

subordinating conjunction in the second utterance of (10) were to be taken as an intervening subordinate discourse level, hierarchical recency allows the first and third clause to be regarded as adjacent (and therefore appearing considerably more coherent than three consecutive C_b-less utterances). Such a method can be extended to the level of sentences, but would then require not only an assessment of the syntactic properties for local focus updates, but also an assessment and interpretation at the global discourse structure level.

Hahn and Strube (1997) experiment with a centering mechanism that uses such a segmented discourse model and which allows treating sentences or multi-sentence units as different levels of discourse ("Centered Text Segmentation Analysis" or simply "Centered Segmentation"). The authors incorporate discourse structure into a centering algorithm and test it on texts from three different text domains (technical reviews, a news article and two literary narrative samples). As a functional theory that incorporates thematisation and topicalisation, the authors connect their framework to the text linguistic framework for the thematic progression patterns in complete texts:

> Daneš (1974) also allows for the combination and recursion of these basic patterns [i.e., Thematic Progression type 1 and type 2]; this way the global thematic coherence of a text can be described by recurrence to these structural patterns. These principles allow for a major extension of the original centering algorithm."
>
> (Hahn & Strube, 1997)

In what they term a typical pattern (thematisation of rhemes, TP type 2) for detailed description of objects, Hahn and Strube note that such texts are characterised by a sequence of Center *Shifts* (Hahn & Strube, 1997, p. 109). Expository text, on the other hand, are characterised by a single-theme property (and probably by extension, continuous thematic progression, TP type 1; cf. Hahn & Strube, 1997, p. 111).

The authors note that with respect to the size of a backward-looking window, applying a strict 1-utterance anterior search space causes 28% of anaphors and 48.3% of textual ellipses in their German test set to fall outside the scope of the window (Hahn & Strube, 1997, p. 110). Such results illustrate that specifying the backward-looking window as U_{i-1} without a discourse structure that takes into account different levels of discourse (and their hierarchical distances) thus significantly affects the coherence evaluation of the texts under study (cf. Hahn & Strube, 1997, p. 110; specifically their Table 10).

Although the model is in competition with the Cache model of attentional state proposed by Walker (1996, 2000), and is even acceded to possibly be part of the latter, the authors note as two advantages that Centered Segmentation is not limited to a certain Cache size, in addition to the fact that their model does not suppose unlimited

retrieval from main or long-term memory but rather restricts the search space for antecedents to only those mental entities already mentioned in the prior discourse.

To underscore the relevance of the issue of the size of the backward-looking window for Centering Theory in general, Hahn and Strube have remarked that "[t]he extension of the search space for antecedents is by no means a trivial enterprise. A simple linear backward search of all preceding centering structures, e.g., may not only turn out to establish illegal references but also contradicts the cognitive principles underlying the limited attention constraint (Walker, 1996b)." (Hahn & Strube, 1997).

6.2.4 Corpus-based evaluations of coherence using Centering Theory

Most corpus-based Centering Theory research seems primarily concerned with empirical evaluation of a specific feature of the theoretical framework. What has perhaps received little attention is whether basic assumptions regarding concepts such as textual coherence, which underlie most of the hypotheses regarding the specifics of the CT framework, are supported by corpus data.[26] However, Poesio et al. explicitly mention the descriptive benefit that may result from applying CT to corpus data, since "knowing the extent to which real texts conform to centering preferences is an important goal in its own right" (Poesio et al., 2004b, p. 310).

The feature most central to our aims, and one which has attracted a great deal of scrutiny in the Centering literature, is the Rule 2 hierarchy of transitions. Some early studies have taken up the issue of whether to use single transitions (two consecutive utterances), as suggested by Brennan et al. (1987) and Walker et al. (1998), instead of sequences of transitions (utterance triplets), as suggested in Grosz et al. (1995) and Strube and Hahn (1999). Recent studies have adopted the single transition-setting and focus rather on how the Rule 2 hierarchy should be interpreted, as well as the degree to which there is empirical support for this hierarchy.

For example, Walker et al. (1998) assert that support for the occurrence of Rough-shifts is weak or nonexistent in naturally occurring discourse, based on data provided in Di Eugenio (1998) and Hurewitz (1998). Such a finding would fit the intuition that perceivably coherent discourse predominantly consists of entity-coherent transitions. However, although Rule 2 states that certain transitions are preferred over others, it does not explicitly state *how* this should be operationalised. Should the rates of occurrence for each transition category reflect their rank in the hierarchy, should there simply be more 'coherent' transitions (Continue+Retain) than Shifts (and other, even less coherent transitions) for a discourse to be perceived as coherent, or is it enough for Continue to be the most prevalent transition overall? In light of later corpus-based

[26] This view is perhaps not surprising given the assumptions regarding *maximum* coherence in the literature. We will address this issue in section 6.2.4.

work that takes into account greater corpus sizes, more varied text domains and a variety of languages, it appears that the view that coherent discourse predominantly consists of entity-coherent transitions cannot be easily maintained. Table 6.5 provides an overview of utterance transitions in a range of domains of naturally occurring text. Caution must be taken when comparing these figures, however, as the studies listed may slightly deviate in their operationalisation of CT parameters.[27]

Table 6.5: Coherence transition statistics in corpus-based CT studies

Source	Text Domain(s)	Continue	Retain	S-shift	R-shift	Trans. (n)
Di Eugenio (1998)	Formal & informal written prose	61.1%	7.1%	'Shift': 9.7%		113
Passonneau (1998)	Spoken narrative prose (Pear stories)	44%	6%	'Shift': 50%		308
Hurewitz (1998)	Written prose & telephone conversations	41.6%	26.9%	7.1%	-	406
Strube and Hahn (1999)	Literature, newspaper & product reviews	31.6%	50.1%	8.5%	9.9%	543
Poesio, Stevenson, Eugenio, and Hitzeman (2004b)†	Pharmaceutical & museum descriptions	7.0 %	3.8%	3.7%	2.3 %	1,007
Friedrichs and Palmer (2014)‡	Newspaper articles, essays & letters	11.4%	8.3%	9.9%	10.6%	8,491

†: using the 'Vanilla' instantiation, ‡within paragraph transitions only

Transition statistics as a coherence metric

The data referred to by Walker et al. (1998) in Di Eugenio (1998) on formal and informal written Italian prose indicate that although Continue appears to be the most prevalent transition overall (61.1%), the transition frequencies do not reflect the canonical Rule 2 ordering. Shifts are found to be more numerous than Retains (9.7 versus 7.1%, respectively), and the frequency with which Establishment transitions are attested (19.5%) indicates that there is a fair amount of utterances for which no backward-looking center can be established at all (implying either a discourse segment change or two unlinked utterances, which includes linking to a discourse entity not available in the immediately preceding utterance, for example to an entity in global focus; cf. Di Eugenio, 1998, p. 128).[28] Results obtained by Passonneau lend even less support to common intuitions regarding transitions in coherent text and interpretations of the Rule 2 hierarchy. Her corpus data results in, on average, 44 % Continue, 6% Retain and 50% Shift transitions (Shifts were combined due to the low number of Rough-shifts, cf. Passonneau, 1998, p. 126).

In contrast, the data presented by Hurewitz (1998, p. 280; table 14.1) offers more support for the canonical ordering of the transition hierarchy. The 'coherent' Continue and Retain transitions occur both as the most frequent utterance transitions overall, as

[27] In addition, it may be observed that the percentages do not add up to 100% for some of these studies. This is because these studies included additional transition categories for which the proportions are as follows: Di Eugenio (1998): Est: 19.5%, Other: 2,7%; Hurewitz (1998): 'No C_b': 24.4% (R-s, Est, Null, Zero); Poesio et al. (2004b): Est: 18.8%, Null: 47.9%, Zero: 16.7%; Friedrichs and Palmer (2014): Est: 25%, 'No C_b': 34.7% (Zero, Null).

[28] It should be noted that since the author is mainly interested in pronouns in Italian, the utterances in her data do not contain any pronouns after a Rough-shift condition (Di Eugenio, 1998, pp. 127-8).

well as in order of coherence (41.6% and 26.9%, respectively). However, the data also indicate that nearly one in four utterance transitions is part of the more incoherent category, as 24.4% of transitions are categorised as 'No C_b' (in this case, Rough-shift as well as Null and Zero transitions, and quite probably Establish; cf. Hurewitz, 1998, pp. 278-9).[29]

Using the 'Vanilla instantiation', that is, setting parameters as suggested in mainstream CT literature, Poesio et al. (2004b) do not find a great deal of support for the Rule 2 hierarchy. Although it is true that the frequencies of the four canonical transitions are attested in order of 'coherence' (Con>Ret>S-s>R-s), the unlinked Null transition is the most frequent transition overall (47.9%), and the two most 'coherent' transitions Continue and Retain together constitute only slightly more than a tenth, respectively 7.0 and 3.8%, of the transitions found in the pharmaceutical pamphlets and museum object descriptions under study. Even when regarding the four canonical transitions only, Poesio et al. (2004b) note that Shifts are attested more than Retains, a finding consistent with every study listed in table 6.5, except for Strube and Hahn (1999) on German. In addition, Poesio et al. (2004b, p. 355) remark that support for the transition hierarchy is found to be weak at best in findings from psychological experiments, as it does not seem to be the case that the more coherent transitions put less inferential load on the receiver than more incoherent transitions cf., Gordon et al. (1993).

The figures for Friedrichs and Palmer (2014) are consistent with this general pattern in that the canonical transitions only account for a fraction of the total number of transitions in a recent large-scale experiment using newspaper writings as obtained from the Wall Street Journal corpus (Marcus, Santorini, & Marcinkiewicz, 1993). Continue is not the most frequent transition (11.4%), nor do the coherent transitions dominate overall (11.4% + 8.3% for Retains) or is the Rule 2 hierarchy reflected in the rates of occurrence of the transitions found in the within-paragraph transitions in their corpus.[30]

Perceived coherence, referential (entity) coherence and 'CT-coherence'

Such results seems to be problematic specifically for what has been termed the 'coherence-assumption' (cf. Miltsakaki & Kukich, 2000, 2004, 01); most studies on Centering Theory seem to assume, implicitly or explicitly, that published texts are (maximally) coherent. It follows, then, that with respect to the transition hierarchy outlined in Rule

[29] As there did not seem to be a proportional difference between transitions in her random control sample of spoken and written utterances, the figures for these two modes have been combined here.

[30] For between-segment transitions, the authors report lower percentages for Establish, Continue and Smooth-shift transitions, and higher percentages for Rough-shift, Retain and 'NoCB' (i.e., Zero and Null; but note that even between paragraphs, there is a remarkably substantial amount of Continue transitions, 7.8%; Friedrichs & Palmer, 2014).

2, it is similarly assumed that coherent texts will exhibit transitions that are deemed coherent, either proportionally or in terms of raw frequencies.[31]

Based on their large scale corpus-based studies of under-researched registers in Centering Theory, Poesio, Cheng, Hitzeman, Stevenson, and Di Eugenio (2002) and Poesio et al. (2004b) conclude that the coherence as measured by CT metrics is not sufficient to explain the perceived coherence of texts, especially given the large amounts of NoCBs in their GNOME corpus (see Poesio et al., 2004b, specifically §5.2.2 & 5.3.3). If texts in the domains under study are indeed perceived as coherent, as seems to be the case, other sources are to be considered as contributing to the (local) coherence of the discourse in addition to the entity-coherence as captured by CT (see for example Poesio et al., 2004b, p. 353). They note that,

> One clear conclusion suggested by our results is that entity-based accounts of coherence need to be supplemented by accounts of other factors that induce coherence at the local level. [...] A more sensible approach [than the current one], especially as we don't yet know all the factors affecting coherence, would be to be more explicit about the scope of centering theory, viewing it not as a comprehensive account of "local coherence," but only as an account of the contribution of entity coherence to local coherence. In other words, we could view (Entity) Continuity as only one among the preferences holding at the discourse level. A natural way to formalize this would be to include Entity Continuity among a set of constraints like those proposed by Beaver, which would also have to include further constraints specifying preferences for rhetorical and temporal coherence. (Poesio et al., 2004b, p. 357)

Since CT is restricted to the domain of entity coherence, the obvious *other* main candidate for establishing coherence in discourse is relational coherence (cf. Poesio et al., 2004b, pp. 353-354). Despite the frequent mention of such other sources of textual coherence in the literature, however, it is the more remarkable that the coherence assumption is strictly upheld in most applications of CT. Karamanis, for example, has noted that most corpus-based studies that provide Rule 2 transition statistics specifically rely on two absolute preferences for coherence, namely that (1) Continue should be the most frequent transition (termed the 'abs-Con' preference), and that (2) the number of NoCB transitions should be minimised (the 'abs-NoCB' preference; Karamanis, 2003, p. 45). The underlying assumption seems to be that when the transition frequencies of a text meet these criteria, this text is deemed coherent in terms of CT.

We take a slight, but important, deviation from the position taken by Karamanis (2003). The author asserts that if these two criteria are met, "CT is taken to be a reliable estimator of the entity coherence of a text" (Karamanis, 2003, p. 45). However, we do not assume that texts are (maximally) entity coherent. Rather than assessing

[31] This idea is explicitly put forward by Brennan et al. (1987), and reiterated by Walker et al. (1998).

whether CT can be taken to be a reliable estimator of a text's entity coherence (based on assumed maximum coherence), we assume CT *is* a reliable estimator of (a certain type of) coherence, termed 'CT-coherence'. Further, we ask *how* coherent a text is in terms of this metric. This stance reflects the intuition that CT-coherence is but one component contributing towards the perceived coherence of a text. Other contributing components might include relational coherence, coherence as established by contextual factors such as bridging references or situational discourse participants (both arguably sources of referential coherence), layout features on the page, etc.

The abs-Con and abs-NoCB condition are suggested to be 'robust' estimators of coherence because a discourse segment in which neither is violated would usually be judged to be coherent by human evaluation. This reliance on true positive evidence fails to account for the fact that discourses that do not meet these criteria are also often evaluated as coherent. In fact, as outlined summarily above, while most corpus-based studies fail to find any support for the abs-Con and abs-NoCB criteria in the texts under study, let alone conform to a frequency distribution of transitions as outlined in the Rule 2 hierarchy, textual coherence is still assumed (see also Kibble, 2001).[32]

Friedrichs and Palmer (2014) point to text domain-specific features in relation to perceptions of coherence. Although the authors note a certain prevalence of shifts in their corpus (10.6% R-shift and 9.9% S-shift within paragraphs), this does not lead them to conclude that the texts under study should be regarded as incoherent. Rather, because the corpus consists of published texts whose coherence is not called into question, they contribute this finding to another factor; what is suggested is that *text domains might vary in their permissiveness of (degrees of) CT-incoherence* as indicated by transitions such as center shifts. Rather than signalling incoherence, center shifting in texts from the Wall Street Journal is seen as a textual device used to make a text more appealing for the reader (see for similar observations Kibble, 2001; Poesio et al., 2004b).

The perfect acceptability of shifts in otherwise coherent texts is illustrated by the example in (11) (example (2) from Friedrichs & Palmer, 2014, p. 142; underscoring indicates C_bs as identified by the authors).

(11) a. Two dozen scientist reported results with variations of the *experiments* [...] by Fleischmann and Pons. [-]

b. The *experiments*$_{C_b}$ involve plunging the two electrodes into *"heavy" water.* [Establishment]

c. When an electric current is applied to the electrodes$_{C_b}$, the *heavy water*

[32]Even when the Rule 2 hierarchy is taken more leniently, with the only requirement being that Continue is the most frequent transition overall, the evidence is not unequivocally in support of Rule 2 as a robust estimator of discourse entity coherence (cf. Karamanis, 2003, pp. 33-4; and footnote 39).

did begin to break up, or dissociate. [Rough-shift]

d. Ordinarily, the breakup of the $\underline{water}_{C_b}$ would consume almost all of the electrical energy. [Rough-shift]

(wsj1550, shortened)

Although we broadly agree with the transition labels provided by Friedrichs and Palmer (2014) in terms of the general CT framework, it should be observed that we deviate from the analysis of utterances (11-c) and (11-d) below (cf. section 6.4.3). What seems most important to notice here is that the use of Rough-shifts seems fairly unobtrusive in terms of discourse flow, however, and does not seem to lead to strong intuitions of incoherence despite the CT utterance labels.

As it happens, (11-d) exhibits a phenomenon that can be argued to be an important coherence establishing device, identified as transcategorisation (Halliday & Matthiessen, 1999; Halliday, 2004) or syntactic recategorisation (Kastovsky, 1982, 2006). This issue has been highlighted briefly by Miltsakaki and Kukich (2004, 01, pp. 45-6), who mention that co-referential links established by nominalisations of a verb or verb phrase in the previous utterance are an unresolved issue in Centering. The example in (11) also neatly illustrates that the identification of transitions is far from straightforward, even when CT-parameter settings are thoroughly scrutinised.

In contrast to Friedrichs and Palmer, Miltsakaki and Kukich (2000) argue that center shifts can be taken as a proxy for the degree of incoherence in a text. However, as Friedrichs and Palmer (2014) also point out, Miltsakaki and Kukich use a corpus of student essays whose degree of perceived coherence exactly *is* subject to question. These two different interpretations of center shifts, as either contributing to incoherence or helping to make the reading more enjoyable, is reflective of the divergent views and the difference between perceived coherence and CT-coherence. (In addition, the latter point of view foreshadows that incoherence should be seen as a relative degree metric rather than an absolute criterion.)

CT-coherence should be seen not as an absolute requirement, but rather as a relative measure. If center shifts are used, they may indeed add 'spice' to the reading experience. However, there may be a point of saturation, after which a text will be perceived as incoherent (e.g., an overuse will create a feeling of continuous flip-flopping between referents). Such thresholds may be genre-specific, as Friedrichs and Palmer (2014) observe, which tallies with the claim by Hurewitz (1998, p. 277) that texts with similar discourse styles should have similar utterance transition metrics.

For the distinction between perceived coherence and CT-coherence, it should be kept in mind that in corpus-based applications of CT, centering is restricted to only a subset of possible referential antecedents (mainly noun phrases and third-person pronouns).

Although the referential entities covered by CT arguably provide a large proportion of usual co-referential entities found in discourse, other forms of referential coherence, for example demonstrative pronouns and pronominal adverbs, are much more problematic and are usually disregarded in applications of CT (see section 6.2.2 and for example the discussion of R1-pronouns in Poesio et al. (2004b)). The fact that Centering Theory in practice is not normally concerned with propositional antecedents, and only with antecedents realised as nominal constituents, makes that it may best be described as a theory of entity-coherence rather than a theory of referential coherence. However, there is no theoretical objection for not incorporating other forms of co-referential antecedents (nominal or non-nominal constituents; cf. Grosz et al., 1995; Walker et al., 1998).

Creating a descriptive summary of transitions for a text seems a fairly straightforward method to gain an insight into the overall CT-coherence of a discourse. However, Beaver reiterates a remark by Grosz and Sidner (1998) that "the sentence by sentence classification of transitions in [Brennan et al. (1987)] and indeed the bulk of later Centering literature do not provide any way to evaluate the coherence of complete texts" (Beaver, 2004, p. 34).[33] Studies in natural language generation, however, have tried to address this issue more fully for purposes of automated text generation.

Insights gained from Natural Language Generation: Choices available to an author

Kibble (2001, p. 585) argues that conforming to centering rules does not necessarily predict a fluent, interesting and smooth discourse. Adhering strictly to coherence demands, for example by continued use of a pronoun, or 'forced' use of a passive instead of a Shift + active verb, appears to hamper the readability (and supposedly, processing load) of a discourse segment. Based on their extensive corpus study of CT parameters, Poesio et al. have also concluded that "ensuring VARIETY seems to be as important a principle in discourse production as maintaining coherence" (Poesio et al., 2004a, p. 82) Kibble (2001) provides an example discourse from a pharmaceutical text to illustrate how a short sequence of Continue – Smooth-shift – Smooth-shift need not necessarily imply an incoherent text. Attempts to restructure the sentences of the text in such a way as to turn the Smooth-shifts into 'coherent' transitions (e.g., Continues or Retains) is unsuccessful. This leaves Kibble (2001, p. 582) to conclude that as far as the discourse content allows, the original author did indeed adhere to centering coherence as much as possible.

This prompts the conclusion that, with respect to empirical evaluation of the Rule

[33]Despite this remark, however, Beaver (2004) does not provide an in-depth account of such an agenda and merely outlines what such an evaluation may look like using his proposal of a hybrid between Optimality Theory (cf. Smolensky & Prince, 1983 [2004]) and Centering Theory.

2 transition hierarchy in natural occurring data,

> not only does corpus evidence fail to confirm the canonical ordering, but in fact corpus analysis itself is not sufficient to evaluate the claims of CT without taking into account the underlying semantic content of a text. That is, statistics about the relative frequency of occurrences of different transition types do not in themselves tell us much about which transitions are preferred in particular situations since they do not take account of the choices available to an author
>
> (Kibble, 2001, p. 582)

That is, coherence can only be established in light of the *options available to an author* (at every turn, discourse move or utterance, etc.). Clumsy, non(-CT)-coherent utterances cannot always be explained in terms of centering:

> The bottom line is that from a generation point of view, *centering is not enough*. Maximizing coherent transitions will not in itself produce optimally fluent and readable text; instead, a number of other factors have to be taken into consideration in order to minimize the inferential load on the reader, hold the reader's interest, and reflect communicative intentions.
>
> (Kibble, 2001, p. 585)

Karamanis (2003) goes a step further in addressing the issue of evaluating the coherence of complete texts using Centering Theory. Where previous corpus-based CT studies have informally provided an entity coherence measure by summarising transitions counts within a text, the author proposes to recombine the utterances of a discourse segment and compute a score for each alternative ordering, specifically concerning the utterance transitions it exhibits (for example, scores for transitions frequencies, whether these violates abs-Con and/or abs-NoCB, etc.). The original discourse segment is then compared to these alternative orderings to see if, given the available propositional content, it maximises coherence. The methodology thus allows an evaluation of the coherence of the initial discourse segment in light of other possible ways an author could have structured a text. Karamanis notes that "in order to account for the existence of dispreferred transitions in a text of attested coherence, one has to consider the choices available to an author when structuring a certain set of utterances. One way to do this is by employing a search-oriented strategy which views the preferences underlying R2 and C1 as a relative rather than an absolute measure of entity coherence" (Karamanis, 2003, p. 41). The examples by Karamanis (2003, p. 48) serve to illustrate that given a set of re-ordered alternatives, the coherence of a corpus example can be evaluated in an alternative way to the transition frequency summary and absolute preferences of abs-Con and abs-NoCB indeed.

Claims based on the methodology of Karamanis are purported to be able to "estimate the entity coherence of text spans longer than a pair of utterances" (Karamanis,

2003, p. 41). However, although it is true that the method offers a valuable metric for relative coherence and takes into account more than a descriptive summary of (local) between-utterance transitions, it also needs to be observed that the re-ordering of utterances only applies *within* discourse segments. The method is thus restricted to making claims about discourse segment coherence, and not that of the entire text. [34]

Thus, it seems that using Rule 2 via a scoring function using re-ordered transitions might still lead to sub-optimal assessments of coherence over longer stretches of text (i.e., incoherent transitions *between* discourse segments). The method proposed by Karamanis (2003, p. 50) seems to go for a seemingly 'locally bad but globally best', text structuring order for coherence transitions.[35] A summary of transition frequencies or proportions might thus still offer some benefit in terms of description of the text and in comparison to other texts in the same text domain.

Obviously, centering does not claim necessarily that it primarily tries to capture all aspects of acceptable discourse, but it is a case in point that CT-coherence is merely one of the building blocks of successful natural discourse: "Maximizing coherent transitions will not in itself produce optimally fluent and readable text; instead, a number of other factors have to be taken into consideration in order to minimize the inferential load on the reader, hold the reader's interest, and reflect communicative intentions." (Kibble, 2001, p. 585)

Although the intuition is probably true that once a topic has changed, the discourse flow does not continuously "flip-flop" back and forth between two discourse entities (Kibble, 2001, p. 585), such a claim probably mainly pertains to certain text domains. In types of discourse prone to linear progression, there might be less topic continuity (although admittedly, there would be little 'flopping back'), but this does not make the text domain 'incoherent' outright. As has been pointed out by Los (2012), modern English narrative does show quite a bit of topic switching, since inanimate entities might sometimes occupy subject position.

Although evaluating texts based on a re-ordering of sentences within a discourse segment might in some way approximate the 'choices available to an author', it should also be kept in mind that the operationalisation of these 'choices' is quite simplistic given that the method does not allow for a recasting of the propositional content. What

[34] Not only would applying this method to the whole text be computationally heavy, but given the assumption that discourse segments share some discourse purpose or or global discourse topic, re-ordering an entire text would probably cut across this level of discourse structure.

[35] But note that Karamanis (2003, p. 50) makes use of the 'locally worst, yet globally best' phrase to refer to using Rule 2 deterministically (rather than with a search-based approach). However, 'local' in his sense refers to the level of the utterance transition, and his 'global' refers to global within a discourse segment (!). Restructuring utterances within a discourse segment so as to minimise violations might make sense at the level of the paragraph, but may result in paragraphs that are themselves unconnected. His 'global' falls short in this respect, it is assumed. Note in addition that Karamanis (2003) makes use of a 5-category CT transition model: the four standard transitions plus NoCB.

we are interested in is indeed more of the latter kind: namely to investigate how, given a summary of transitions, a naturally occurring text may establish coherence, and be perceived as coherent, despite being CT-incoherent (for example, by not obeying abs-Con).

Having mentioned a reordering of the propositional content brings to mind Hallidaian agnates, although these are not the alternatives as envisioned by Karamanis (2003). Rather, it is established how, within a discourse segment, the utterances can be reordered to result in the maximum amount of Continue transitions (the Abs-Cont condition) and no violations to NoCB transitions. That said, it is also asserted that the current NLG perspective is only one way to study the issue of authorial choice, and that it is outside the scope of the current study to investigate the option of changing the semantic or propositional content of the utterance to make the text more CT-coherent (Karamanis, 2003, p. 47).

In addition, Kibble remarks that both hearer and speaker are involved in negotiating cognitive load: "while a sequence that conforms to the cheapness principle may reduce the cognitive load on the hearer, it can actually *increase* the load on the speaker owing to the need to plan ahead beyond the current utterance" (Kibble, 2001, p. 586).[36] Even despite these improvements regarding the principles underlying a classification of CT transitions, it is concluded that, "referential continuity as specified by CT may play an essential part in computing the overall coherence of utterance transitions but it is only one of the determinants of discourse structure" (Kibble, 2001, p. 586).

Next to prompting investigations such as Karamanis (2003), which adopts this sentence restructuring methodology for maximising utterance coherence within discourse segments, the remarks by Kibble (2001) are a case in point that different domains may have different preferences regarding, and allowances as to, the transition hierarchy. As a result, the choice for transitions is constrained to some extent by discourse conventions or preferences, and it is these that partly guide the choices available to an author – or at least, helps to restrict the range of available options.

Text domain or genre effects in non-narrative prose

The coherence of a discourse is readily affected by (text structuring) features particular to a certain text domain (e.g. Kibble, 2001; Karamanis, 2003; Poesio et al., 2004b; Friedrichs & Palmer, 2014). This text domain influence usually receives only passing mention, however, as empirical usually studies seem interested mainly in scrutinising the internal workings of the CT framework and not in an (exhaustive) descriptive characterisation of a particular text or text domain. The current section will provide

[36] See also Levelt (1989).

some of the text domain-specific descriptive findings that can be gathered from corpus-based studies on CT. These may guide intuitions as to how much 'coherence' we may expect to find in the text domain under consideration.

A number of authors have commented on the influence of specific genres or text domains on CT-coherence, specifically in those cases where it concerns deviations from (entity) coherence assumptions as guided by intuitions based on narrative discourse (Strube & Hahn, 1999; Kibble, 2001; Karamanis, 2003; Poesio et al., 2004b; Speyer, 2007; Friedrichs & Palmer, 2014). Against general assumptions underlying Rule 2, Strube and Hahn remark that the use of transitions which in isolation seem to require considerable inferencing load are readily acceptable in certain text domains (Strube & Hahn, 1999, p. 332). Similarly, Kibble (2001) notes that given the lack of support for the Rule 2 transition hierarchy in corpus-based studies,

> A preponderance of Shifts over Continues may reflect the domain and content of a text rather than the author's organizational goals. In fact, it can be seen that sequences of Smooth Shifts are rather natural in certain kinds of narrative or descriptive text
>
> (Kibble, 2001, p. 581)

This comment suggests that (1) the transition hierarchy is not reflected in the rates of occurrence of transitions as exhibited in the corpus data (and therefore, at least abs-Con is misguided), that (2) transitions are thought to be largely dependent on an author's discourse organisational aims, but also that (3) such aims may be overridden by the semantic content of a text or the specifics of a certain text domain (and possibly, that (4) at least descriptive text and some forms of narrative show unexpected patterns based on general CT assumptions).

The third point, the influence of the text domain on the textual organisation of a discourse, is also reflected in the remark by Karamanis (2003) that text structuring in natural language generation is usually regarded a genre-specific problem (citing work on natural language generation by Reiter & Dale, 2000). Thus, the particulars of the structure of a text are perceived as being guided by the *discourse purpose* of a genre (i.e., description of objects in a museum catalogue, in the case of Karamanis, 2003).

Strube and Hahn (1999) have also noted that in the type of discourses to which centering has been mainly applied, the predominant text phenomenon on which referential coherence relies is pronominal reference. However, there is evidence that this is not the case for all text domains, however, and the authors note that some texts might rely more heavily on bridging inferences and the type of coherence established through world knowledge as the main source of referential coherence (Strube & Hahn, 1999, pp. 320-1):

> We claim that [an] extended version of ranking constraints is necessary to analyze texts from certain genres, e.g., texts from technical or medical domains. In these areas, pronouns are used rather infrequently, while functional anaphors [i.e. bridging inferences] are the major text phenomena to achieve local coherence.
>
> (Strube & Hahn, 1999, p. 321)

In addition,

> While the basic Cf ranking criteria are sufficient for texts with a high proportion of pronouns and nominal anaphora (e.g., literary texts, newspaper articles about persons), it is necessary to refine the ranking criteria in order to deal with expository texts, e.g., test reports, discharge summaries. These texts usually contain few pronouns and are characterized by a large number of inferrables, *which are often the major glue in achieving local coherence.*
>
> (Strube & Hahn, 1999, p. 323; emphasis mine)

Such remarks are supported by evidence from register studies, which finds that spoken registers (conversation and speeches) mainly make use of anaphoric pronouns and full nouns, whereas written registers largely rely on full nouns to realise referring expressions (Biber et al., 1998, p. 118). Interestingly, Friedrichs and Palmer note that the category of NoCB transitions is more prevalent in *essays+letters* than in newspaper *articles* (43.4% versus 32.6% of transitions within each text domain). This result is accounted for by the fact that newspaper articles usually make use of a limited set of protagonists, whereas letters and essays "change their focus on different entities as they progress" (Friedrichs & Palmer, 2014, p. 141).[37]

These findings by Friedrichs and Palmer link up with the discourse model of Longacre (1996), especially the influence of the dimension of Agent Orientation. The literature on referential topic coherence (cf. Sanders & Maat, 2006) predominantly focuses on how CT-coherence is established in narrative prose. An obvious target for investigations of grammatical realisation of referential entities in this type of discourse puts into focus pronominals. Non-agentive discourses largely make use of other devices to construe coherence, however, and based on findings from genre-sensitive studies of referential coherence, it would put noun phrases into focus in the current text domain in particular, especially for texts in the later half of our corpus.

Specific studies on CT, grammar and IS

Speyer (2007) has devoted specific attention to the relationship between referential status, syntax and information structure by studying C_b-status and the left periphery

[37] On the other hand, this also somewhat contradicts the reported finding by Biber et al. (1998, p. 116) that out of four registers (conversation, speeches, news and academic writing), newspaper writing has the highest frequency of referring expression (see also Biber, 1992).

of the clause in German. In terms of transition frequencies, the author does not find any marked differences in the proportional distribution between genres. Although Speyer (2007) does not provide any specific figures, this finding might be the result of corpus sampling rather than a contrast between English and German, as Strube and Hahn do find differences between text domains in written German (that is, depending on the C_f-ranking chosen; cf. Strube & Hahn, 1999, p. 331).

A finding more central to Speyer's aims, however, is that there is a correlation between utterance transition type and grammatical realisation of discourse entities in the left periphery: the more CT-coherent the transition, the more likely it is that the C_b occurs in the *Vorfeld*/prefield and establishes a strong link with the previous context there. In contrast, transitions that are less coherent in terms of the Rule 2 hierarchy are more likely to exhibit prefield constituents or phrases that signal a contrast (or perhaps more accurately, a 'delimitation' in the sense of Krifka and Musan (2012)) or a poset relationship (Speyer, 2007, p. 104). 76.61% of Continue transitions have the C_b in the prefield, and of the remaining 40 cases, 87.5% have the C_b as the first element of the middle field (Speyer, 2007).[38] The suggestion that there is a correlation between grammatical status and the position of constituents in the German clause might not be a new insight, nor that this relation is much stricter in English than it is in German. Nevertheless, these finding by Speyer (2007) highlight the relationship between grammatical status, position and referential status of discourse entities. In addition, it tallies with observations regarding the Theme-Rheme literature, according to Speyer (2007, p. 104), which we will turn to presently.

With reference to centering theory and diachrony, Martnez Insua (2011) uses an expanded version of CT (termed Meta-informative Centering Theory (MIC); cf. Wodarczyk & Wodarczyk, 2013) to investigate the role of *there*-existentials as discourse coherence establishing devices in the history of English. Given the current research agenda, this study deserves particular mention for its choice to use CT to investigate a single construction in diachronic perspective, and specifically given its focus on the connection between information structure and discourse coherence after the late Middle English period. Unfortunately, Rule 2 and the transition hierarchy receive only passing mention, and Martnez Insua (2011) does not provide any descriptive utterance transition metrics for the investigated corpus material. Nevertheless, it is suggested that the *there*-constructions under analysis might contribute to CT-coherence by 'softening' Rough-shift transitions in historical English texts (Martnez Insua, 2011, p. 110). The author concludes that "[there-existentials] have the capacity to serve discourse coher-

[38] Somewhat remarkably, the Retain transitions scores lowest in terms of C_bs in the left periphery: 35.44% of Retain transitions have a C_b in the prefield. As for the remaining C_bs, Retain also scores worst in comparison with the other canonical transitions, with only 45.10% of the non-*Vorfeld* C_bs occurring at the head of the middle field (Speyer, 2007, p. 103).

ence by continuing with the same local topics as previous utterances, as 'smooth-shift' transitions from one [center of attention] to another" (Martnez Insua, 2011, pp. 112-3). As essentially a diachronic form-to-function mapping of *there*-existentials in terms of CT, this study provides a helpful cue in the analysis of utterance transitions and discourse organisation in the current corpus. Before we move on to such an analysis, however, we will review how the literature on Thematic Progression (cf. Dane, 1974) offers a perspective on discourse organisation from a more text linguistic angle.

6.3 Thematic Progression and textual organisation

6.3.1 General outline

Following from the Prague School insight that utterance rhemes 'push the communication forward' (cf. Firbas, 1964), one of the functions of sentence *themes*, on the other hand, is that these may serve as discourse anchors which play an important constructional role in the organisation of text. It is this property that is exploited in the Thematic Progression framework by Dane (1974). Thematic Progression , henceforth also TP, encompasses "the choice and ordering of utterance themes, their mutual concatenation and hierarchy, as well as their relationship to the hyperthemes of the superior text units (such as the paragraph, chapter, ...) to the whole text, and to the situation" (Dane, 1974, p. 114). Dane argues that such an overview of utterance themes, as well as their (hierarchical) relationships, can offer a "skeleton of the plot" Dane (1974, p. 114). According to the author, this outline provides an important aid in explaining a text's coherence albeit perhaps not the only one, Dane (1974, p. 114).[39] Although the purpose of the TP framework seems to be primarily descriptive, in offering a way to chart the thematic organisation of a (coherent) text (cf. Dane, 1974, p. 127), it serves to illustrate the relationship between text structure on the one hand, and the study of sentence level phenomena as known from the traditional Prague School's Functional Sentence Perspective (FSP) on the other.

An artefact of the latter may be the attention Dane devotes to the difference between simple, complex and condensed utterances. Although these may be described in more general grammatical terms (e.g., the provided examples of 'composed utterances' involve paratactic/coordinated clauses or phrasal constituents, and 'condensed utterances' seem to illustrate cases of hypotaxis or subordination), they serve to illustrate a focus on sentence level phenomena for which FSP is generally known.[40]

[39] Dane explicitly uses the term text*connexity* rather than text *coherence* to refer to the (results of) textual organisation of a piece of writing. The linguistic concept 'coherence' is preferred as a theoretically neutral term, rather than as a label to describe a characteristic of a particular text (Dane, 1974, p. 114).

[40] Next to apposition and coordination, Dane (1974, p. 117) mentions nominalisation and relativisa-

With respect to text and discourse structure, and in contrast to general assumptions in the Centering Theory literature,Dane stresses that "coherence (connexity, continuity) is not a necessary property of text" (Dane, 1974, p. 114). Rather, discontinuity and uneven degrees of (referential) coherence, within and across texts and text fragments, is readily evidenced in studies on text coherence.[41]

It is clear from Dane (1974) that the author ascribes to the basic distinction between information status (i.e., familiarity, Givenness) and thematisation (i.e., Aboutness, Salience). Specifically, although thematic status and known (given) information may often fall together, those cases in which themes convey new information "undoubtedly calls for, and justifies" a distinction between Givenness and (the status of) theme (Dane, 1974, p. 108).

Crucially, the conceptualisation of the utterance theme that Dane (1974) employs for his Thematic Progression framework is closest to the Beneš-basis (Dane, 1974, p. 112).

> Our conception of the utterance theme stands near to E. Beneš's characterization of the "point of departure" (Cz. "východisko", G. "Basis") as "the opening element of the sentence") that "links up the utterance with the context and the situation, selecting from several possible connexions one that becomes the starting point, from which the entire further utterance unfolds and in regard to which it is oriented" (1959, 216).
>
> (Dane, 1974, p. 112)

From the above it should be clear that Dane does not have in mind Givenness as the central feature of the theme, but rather something that is either captured by primary position in the sentence (point of departure in a very linear sense) or, more to the point, something that is expressed by way of Salience/Aboutness (the referent which the remainder of the utterance says something about Dane, 1974, pp. 106-7). [42]

Particularly relevant for an application of TP is how Dane (1974) approaches the issue of identifying utterance themes. Although it is mentioned that objective criteria for finding utterance themes is required (reference is made to Firbas' CD methodology), Dane also notes that for the current purposes, a "rough determination" suffices

tion as grammatical devices that may be used to form complex or condensed utterances.

[41] See for example the references made to work by Hausenblas (1964) and Trost (1962). The latter deals with an old distinction between a connected style (*harmonia glaphyra*), which tends towards "a close linking up of the sentence with the text", and a disconnected style, known as *harmonia austera*, which can be characterised by "a clean-cut independence of each sentence" (Dane, 1974, p. 114).

[42] In addition, there is some evidence to suggest that "[that which] links up the utterance with the context and the situation" should be understood not only as directly co-referentially linked to an entity that came before, but also in more indirect ways of realisation (i.e., bridging inferences). A discourse entity that is inferentially linked to some entity in the context or situation (e.g., part of set, part of whole, antonym, etc.), while also being 'discourse-new', may thus still be denoted as having thematic status (cf. Dane, 1974, p. 110).

(Dane, 1974, p. 114). The methodology employed involves a number of *wh*-questions to determine the FSP-structures of particularly the rheme, and indirectly deriving the utterance theme from these. This procedure satisfies both the criterion for being an objective procedure in addition to being a purely linguistic method for identifying themes, according to the author (Dane, 1974, p. 115).

6.3.2 A framework for Thematic Progression

Based on a number of scientific and other professional texts in Czech, complemented by some cursory reading of German and English materials, Dane (1974, p. 118) arrives at three elementary patterns for the connection between the themes of consecutive utterances. It is these three basic patterns which provide the foundation of the Thematic Progression of texts. (Figures 6.1 - 6.4 are adapted from Dane (1974); text samples are taken from Duberstein and Johnson (2009/2012).)

Thematic Progression types

Type 1 - Simple Linear Progression (thematisation of rhemes) Rhematic material is consistently promoted to the theme of the consecutive utterance.

Example: "Horses can easily digest *nonstructural carbohydrates*$_{R_1}$, mostly in the small intestine. *These sugars and starches*$_{T_2}$ are primarily found in grains (e.g., corn, oats, barley) and provide a more concentrated form of energy than structural carbohydrates" [$\mathrm{R}_1 \Rightarrow \mathrm{T}_2$]

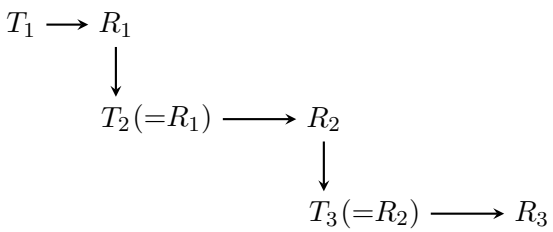


Figure 6.1: Simple linear Thematic Progression (type I; Dane, 1974)

Type 2 - Thematic Progression with a continuous (constant) theme The theme of a sentence remains the same over two or more discourse utterances (i.e., these themes realise the same discourse entity, but not necessarily in identical wording).

Example: "*Carbohydrates*$_{T_1}$ will most likely be the largest part of the horse's diet. *They*$_{T_2}$ can be divided into two groups: structural (fiber) and non-structural (sugars and starches)." [$\mathrm{T}_1 \Rightarrow \mathrm{T}_2(= T_1)$]

Type 3 - Thematic Progression with derived themes ("hypertheme") The utterance themes of a number of consecutive clauses are not connected to the material in the theme

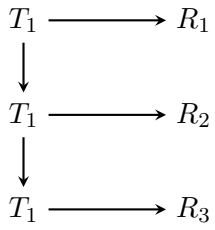


Figure 6.2: Continuous theme Thematic Progression (type II; Dane, 1974)

or rheme of the immediately preceding utterance, but refer to an overarching theme (of a paragraph, a larger text section, etc.).

Example: "*A growing horse*$_{T_1}$ generally needs between 12 and 18 percent crude protein in its diet for proper growth and development. [...][43] *Mature horses*$_{T_2}$ will most likely do fine on a lower protein percentage (8 to 12 percent), depending on their workload. *Horses that are in intense training*$_{T_3}$ need more protein than the maintenance horse because they are developing muscle tissue; however, most will still do well on a 12 percent protein feed." [T_1; T_2; T_3; Hypertheme (T) ≈ protein requirements for horses of varying ages and workloads]

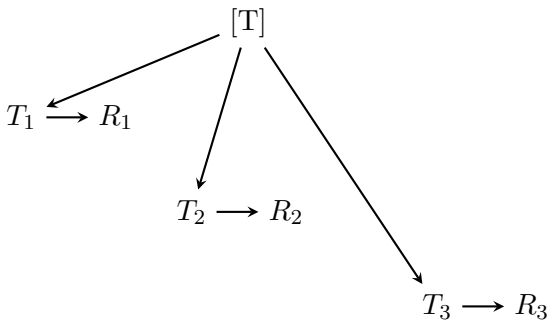


Figure 6.3: Hypertheme/derived theme Thematic Progression (type III; Dane, 1974)

Overarching patterns of Thematic Progression These three basic forms of thematic progression can be combined, according to Dane. Some of these more elaborate, 'higher order' TP-types occur frequently enough to be characterised as a regular pattern themselves. An example is the 'Split Rheme' (sometimes also referred to as TP type VI), for which a schema is provided in (12). This form of thematic progression takes several rhematic elements from one utterance and exhaustively explores each of these as new themes in succession cf., Dane (1974, p. 120-1).

[43]The intervening utterance here elaborates on the information in the first sentence, but it is on an embedded, sub-discourse level. It reads, "Horses need more protein when tissue is being laid down for growth (i.e., young horses in rapid growth phases, gestating mares in their last trimester, and lactating mares that need to produce large quantities of milk)."

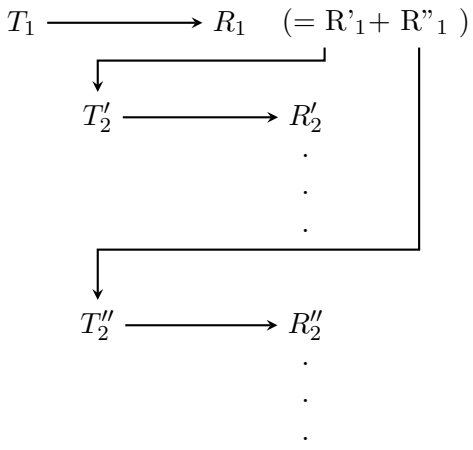


Figure 6.4: 'Split' rheme Thematic Progression ('type IV'; Dane, 1974)

From the figures and description in Dane (1974), it can also be inferred that the exploration of these new themes usually takes up more than one utterance (cf. TP type III), and that within each subordinate level, here indicated with ' and ", the utterance-to-utterance progression of themes itself makes use of the three basic TP types. For a general outline of the progression of themes, and especially with reference to their contribution to referential and/or textual coherence, it is thus sufficient to rely on these elementary TP types. It appears that, if one is interested in such 'higher order' patterns, overarching structures such as the 'split rheme' seem to follow quite naturally from a combination of the basic thematic progression types with the hierarchy of discourse levels of a text following proposals by Hahn and Strube (1997). Hahn and Strube (1997) seem to have been the first to note the natural overlap between Centering Theory and Thematic Progression, exploiting it for their Functional Centering perspective (cf. Strube & Hahn, 1999).

6.3.3 Corpus-based Thematic Progression studies

In the original paper by Dane (1974), there seems to be a suggestion that certain genres or text domains prefer certain TP types over others. According to Fries TP type 1, simple linear Thematic Progression, appears to be typical of scientific writing (cf. Fries, 1981). TP type 2, Continuous theme, seems typical of narrative, with a protagonist being introduced as discourse entity and then recurring as theme in consecutive sentences (via pronominal usage, etc.). Descriptive texts are often characterised by (sections of) thematic progression via derived themes, for example the description of socio-geographical objects in the examples provided by Dane (1974) for TP type 3. This type of thematic progression seems also typical of enumeration. Although no specific examples are provided, Dane (1974, p. 120) notes that extralinguistic signs often

help the present the subject matter for this form of Thematic Progression. One might think of it occurring quite naturally in conjunction with layout features such as bullet points. In addition, Hahn and Strube (1997) note that the derivation of TP type 3 usually requires some inferences and semantic interpretation, as this form of progression of utterance themes, and the coherence that depends on, or is caused by, such interconnectedness, relies on "conceptual generalisation hierarchies" (Hahn & Strube, 1997, p. 105).

Based on application of the model to texts in a variety of text domains (i.e., genres), a number of scholars have suggested improvements to the initial Thematic Progression model by Dane (1974), both in terms of its methodology as well as its mechanics. Francis (1989) carries out a corpus-based study on a random selection of newspaper articles in three different sub-genres, after pointing out that patterns of thematic progression are often illustrated using carefully chosen, prototypical sample fragments rather than randomised texts from a particular domain that provide a full summary of the complete text (reference is made to Fries, 1981; Martin, 1985). In this way, the often suggested relationship between genre and particular TP type preferences may be investigated more adequately.

Dubois (1987) suggests to discard of the third TP type, hypertheme, as she only finds one example of this form of thematic progression in her collection of written slides and transcribed recordings of an annual meeting of experimental biologists. In another empirical study of (bio)medical texts, Nwogu and Bloor (1991) found examples of derived themes in one text domain (medical research papers), but none in the popular medical texts that were examined. Such findings not only highlight that particular text domains might prefer certain forms of thematic progression, but also suggests that the formality of the register may affect whether certain TP types are deemed appropriate in a particular subject area or professional field (that is, their usage may be conventionalised).

The register studied by Downing (2001), in turn, is taken from the popular domain entirely; it consists of texts from newspaper and magazine articles in the leisure section. Of the three texts that are discussed, Downing (2001) does not find any instances of TP type 3 in the first text (the author rather notes the stylistic effect created by the alternation of TP types 1 and 2), whereas the second and third text, both descriptions of holiday destinations, are respectively noted as consisting of TP type 3 only (with the exclusion of one sentence, which shows simple linear progression), and a combination of types 1, 2 and 3. Thomas (1999) finds all three TP types represented in English essays by French students of telecommunications engineering, although to varying degrees: 31% of utterances display linear thematic progression, 45% constant theme and 24% 'hypertheme' thematic progression.

With a more historical focus, Downing (1995) uses a revised version of the Thematic Progression framework by Martin (1992) to study Chaucer's *General Prologue* to the *Canterbury Tales*. The author observes that the pilgrims, once introduced, are in most cased maintained using TP type 2, Continuous Theme (Downing, 1995, p. 158). Reference is made to the notion of topic continuity by Givn (1983), and it is observed that the patterns of thematic progression seen in the *General Prologue* tie in with intuitions regarding narrative discourse in general. In addition, Downing (1995) observes that the description of the attributes of the pilgrims often elicits non-canonical word orders (Downing, 1995, p. 159).

TP and genre

Studies such as Nwogu and Bloor (1991), Dubois (1987) and Downing (2001) seem to suggest that there is a certain relationship between text domain and preference for TP types. Nevertheless, Fries (1995, p. 8) comments that there is only weak support at best for the idea that Thematic Progression correlates with genre in a review of the literature to date (see Ghadessy, 1995; Hasan & Fries, 1995, for collections of empirical papers on Thematic Progression). More specifically, Fries (1995) suggests that rather than the genre in which it occurs, the use of specific forms of thematic progression may be influenced by the *discourse purpose* of a text or text segment, linking up with the intentional level of Discourse Structure as identified by Grosz and Sidner (1986).

This seems in line with the position held by Ariel (2007) on the relationship between register and the distribution of grammatical constructions with information structural properties. Ariel suggests that although corpus statistics may indicate that registers might be (uniquely) characterised by the use of certain grammatical or information-packaging constructions, this should be thought of as an artefact of the use of certain discourse purpose(s) associated with a particular register, rather than an intrinsic linguistic feature of the register itself (Ariel, 2007, p. 288). Both Ariel (2007) as well as Thomas (1999) point to the observation by Swales (1990) that genres are shaped by recurrent communicative events and associated communicative purposes, which drive the association with certain linguistic realisations. In addition, and with particular reference to thematic progression, Thomas notes that

> [A]s Swales (1990) and many others have shown, it is not the subject matter itself but the genre which determines many aspects of textual functioning. The thematic networks established in popular scientific accounts and research articles would seem to be more constrained by such factors as purpose and intended readership.
>
> (Thomas, 1999, p. 141)

Similar to observations on CT regarding allowances for incoherent utterance trans-

itions in a text, thematic progression patterns are thus subject to text domain-specific variation. As both frameworks provide perspectives on the structuring of discourse, this only seems fair: if communicative purposes shape Discourse Structure, determinants of such a structure will vary according to circumstances. In a similar way, the grammatical constructions of Ariel (2007) can be seen as linguistic expressions of discourse moves, and are thus also subject to such factors.

Since both TP as well as CT provide (partial) theories of discourse structure, and as their units of analysis (i.e., utterance transitions and thematic progression patterns) are subject to at least some of the same text domain-specific situational constraints, it is tempting to suppose that both frameworks tap into an underlying aspect of discourse structure. And indeed, several Centering studies mentioned above have noticed an overlap between utterance transitions and the patterns described in the Thematic Progression literature (e.g., Friedrichs & Palmer, 2014; Hahn & Strube, 1997; Speyer, 2007; Strube & Hahn, 1996). In the following section, we will explore this overlap between TP and CT more closely, and offer a combined account which draws in some of the information structure literature.

6.4 A hybrid account of Centering Theory and Thematic Progression

One of the primary reasons for taking CT as part of our theoretical model is that it offers a formal framework which allows an account of the degree of referential coherence in discourse, as based on assumptions of coherence in PDE prose. This is particularly beneficial if we want to investigate why an Early Modern text may appear so incoherent to an audience that is accustomed to – and trained to recognise – present-day conventions of textual coherence. Despite Centering Theory having received criticism for its inability to capture all forms of (referential) coherence adequately, it seems beneficial still as a model that offers a structured framework for large-scale annotation whilst also providing an existing body of work for comparison across various text domains and genres of PDE. In addition, it has been suggested that CT is compatible with a text linguistic model for the organisation of discourse (cf. Hahn & Strube, 1997), as well as being highlighted as a relevant coherence theory in current accounts of Discourse and Information Structure (see explicit references to the theory in Kruijff-Korbayov & Steedman, 2003; Mitkov, 2010; Vallduv, to appear).

Meanwhile, Smith has employed the TP framework by Dane (1974) fairly recently, claiming that the theoretical validity of these patterns is still very much applicable (Smith, 2003, p. 245).[44] In the context of a large-scale corpus-based CT study of

[44] Nevertheless, the author decides to rename the Linear theme TP as "Focus-Topic Chaining" (Type

natural text, Friedrichs and Palmer (2014) too, note how some relevant CT patterns overlap with those investigated in the Thematic Progression literature. The most extensive establishing of an overlap between Centering utterance transition and Thematic Progression patterns, however, is evident from the Functional Centering framework by Strube and Hahn (1996, 1999). In particular, and explicitly exploited in Hahn and Strube (1997), these authors have develop an hierarchical discourse segmentation algorithm that derives Thematic Progression patterns from centered data, thus bridging the implicit relationship between local focus of attention and global discourse structure.

Although far less formal, our aim is similar in extending this overlap between Centering Theory and the Thematic Progression of texts. In addition, we include a brief discussion of the status of the hypertheme pattern (TP3), which is largely left aside in such CT-oriented accounts (see section 6.4.3 below). Before a practical application of such a combined account can be attempted, however, the overlap between the frameworks of TP and CT needs to be scrutinised – and more comprehensively than by merely noting a similarity between TP patterns and utterance transitions as seen in CT data.

6.4.1 Aboutness, salience and themes/topics in CT

Salience and Aboutness

At first glance, CT seems to avoid the quagmire regarding sentence level notions of IS, e.g., theme/rheme, topic/comment, focus/background, etc.[45] Although the theory is concerned with the attentional state of discourse participants (e.g., local versus global focus, and center of attention), it does not explicitly address what the current utterance is 'about' (cf. the aboutness-topic and compatible notions of IS). What centering in essence claims to track in a given discourse utterance is 'what came before' (the C_b) and 'what is most likely to come next' (C_p). Nevertheless, on closer inspection the latter notion of 'what is most likely to come next' seems indeed based on sentence-level phenomena such as grammatical role and IS cues, which are used to guide the C_f-ranking. In most of the CT literature, however, such C_f-ranking phenomena are considered language-specific determinants which are regarded as practicalities rather than issues with which the centering mechanism itself should be concerned.

The notion that such sentence-level phenomena are not related in any way to what

1), while Continuous theme (Type 2) is termed "Topic Chaining" (following Givn, 1983) and the hypertheme pattern (TP 3) is rebranded as an "Unchained" connection of topic phrases (cf. Smith, 2003, pp. 245-6).

[45] In fact, Beaver notes that despite the close relationship between Centering concepts and existing terms in the domain of Information Structure, a lack of consensus in linguistic theories on the use of notions such as *topic* and *focus* may actually have been the main reason for the introduction of a novel terminology by early Centering theorists, so as to avoid pollution (cf. Beaver, 2004, p. 12fn11).

came before and only affect a *prediction* of what is to come next (or in other words, that C_b-status and C_f-ranking are essentially unrelated), is however an important assumption which establishes a connection between CT and Discourse Structure and Information Structure. Specifically, it ties in with the distinction between Coherence and Salience in the transition matrix as identified by Kibble (2001). That is, whether or not a discourse entity is realised in the current utterance (i.e., has C_b-status) is a matter of text planning (or textual organisation), and in a broader sense, of Discourse Structure. Information Structure, on the other hand, is involved in how the discourse entities that *are* realised in the current utterance are ranked, based on grammatical and word order transformations. It is therefore primarily a matter of *sentence planning.*[46]

Salience, however, and particularly the notion of 'local focus of attention' at the level of the utterance, could thus in fact be seen as a proxy for 'aboutness' based on structural or functional features of the utterance. Nevertheless, Speyer (2007) takes care to distinguish between the concepts of Aboutness and Salience, noting that, formally, there is no place for 'aboutness' in Centering Theory:

> [S]treng genommen ist im Formalismus der Centering Theory für Konzepte wie ‚Aboutness' gar kein Platz; die Theorie bezieht sich nur auf Givenness und Salienz. Aboutness ist hier eher ein Epiphänomen
>
> (Speyer, 2007, p. 90)

That there is such a categorical difference between Aboutness and Salience is however contradicted by CT's theoretical groundwork laid out by Grosz and Sidner, who regard the notion of aboutness according to Reinhart (1981) and 'local focus of attention' as largely overlapping concepts (Grosz & Sidner, 1986, p. 192). Similarly, Grosz et al. (1995, p. 206) resort to using of the term 'aboutness' in discussing a sample fragment in which rapid switches between C_ps take place, illustrating that Salience and Aboutness are closely related concepts in the early CT literature. Put differently, it is argued here that C_f-ranking may be seen as a formal device for identifying aboutness (see also Strube & Hahn, 1996, 1999).

With respect to Salience, one may in fact question to what extent finding 'the most salient entity in the current utterance' (following the interpretation of the $C_p$by Kibble, 2001), or along lines of 'the next utterance will most likely refer back to entity x in

[46] See a remark by Downing (2001), however, in which she notes that Halliday had noted early on that "thematization is independent of what has gone before" (Halliday, 1967, p. 17). This idea is challenged by Dane, who addresses this suggestion by stating that "such a conclusion appears very doubtful in the [*sic*] light of the fact that the choice of the themes of particular utterances can hardly be fortuitous, unmotivated, and without any structural connexion to the text" (Dane, 1974, p. 109). However, what Halliday probably intended here is that *in principle*, thematisation is not dependent on what has gone before. Even though they may be related, they must be regarded as distinct dimensions/systems/etc. See also Reinhart (1981), who argues that topic-hood is determined in part by the linguistic structure of the utterance (semantic, syntactic, intonational), and partly by the (discourse) context of the utterance.

the current utterance' (cf. Grosz et al., 1995, p. 212), is ultimately any more formal than deriving an aboutness-topic, usually paraphrased as 'an utterance being about X' (although more accurately, such topics are defined as "the expression whose referent the sentence is about"; cf. Reinhart, 1981, p. 57). Given this overlap, it seems justified to conclude, with Speyer (2007), that it is appropriate to assume that the CT framework offers both a shorthand for the aboutness-topic, as well as the notion of 'theme'.

Themes and aboutness-topics in CT

Several Centering Theory studies have drawn explicit links between sentence ('aboutness') topics and the Centering mechanism. Specifically, but not exclusively, Brennan et al. (1987), Ward (1988), Poesio et al. (2004b), Beaver (2004), Speyer (2007) identify the backward-looking center (C_b) in Centering Theory as the 'sentence topic' or aboutness-topic. Beaver (2004, p. 25), however, takes care to remark that when postulating the aboutness-topic, Reinhart (1981) strongly argued *against* the definition of topics in terms of givenness:

> topics and old information are clearly distinct phenomena. Representing old information is neither a sufficient nor a necessary condition for an expression to serve as the topic expression.
>
> (Reinhart, 1981, p. 78)

Reinhart continues her observation by noticing that while topics are to be identified independently of the notion of givenness, old information may help to determine the topic in a given situation (cf. Reinhart, 1981, p. 78).

In contrast to Centering studies that conflate the notions of topic with the C_b, others such as Strube and Hahn (1996, 1999), Hahn and Strube (1997, p. 212) equate the notion of topic, or rather, theme, with the C_p (and although it is not stated explicitly, all other forward-looking centers barring the highest ranked C_f should then be analysed as part of an utterance's *rheme*, it is supposed here). This seems consistent with the views of both Dane (1974) and Reinhart (1981) in the sense that given the choice between designating the C_p or the C_b as theme, the latter should be discounted on the grounds that Givenness is definitely not a criterion for thematicity/topic-hood. Hahn and Strube argue that although not every utterance has to have a *given* element, and thus a C_b, each utterance "must have a theme and C_f as well", which is then implicitly its C_p (cf. Hahn & Strube, 1997, p. 112). In keeping with the IS-oriented Functional Centering framework of Strube and Hahn, we thus take the theme or aboutness-topic as an utterance's most salient entity, and in effect the correlate of the C_p (following e.g., Dane, 1974; Hahn & Strube, 1997; Strube & Hahn, 1996, 1999).

$C_{b/p}$ and the Mathesius-theme

Incidentally, on the hoof of conflating the notions of *theme* and $C_{p/b}$, the Centering Theory terminology allows a distinction between two aspects of theme that have attracted debate ever since its influential definition by Mathesius (1939). Mathesius notion of theme, widely taken up in the English speaking linguistics community by way of Firbas (1964, p. 268), is defined as,

Mathesius-theme "That which is known or at least obvious in the given situation, and from which the speaker proceeds in his discourse" (Mathesius, 1939, p. 234)

As may be clear, the centering concept of C_b captures the first element within this definition concerned with Given, Known, etc. In turn, the element of 'Aboutness', or its proxy Salience, is captured by the C_p. The Centering framework thus allows a clear separation of these two aspects, and it shows that only when an utterance's C_b and C_p coincide, we can speak of a referential entity being a true theme in the Mathesian sense.[47] More recently, Smith (2003, p. 197fn12) has also observed the close connection between the formal notion of C_b in Centering Theory, and how it harks back to one of the two key aspects of the Mathesius-theme, i.e., of linking up to what has gone before. Note, however, that the phrase "from which the speaker proceeds in his discourse" may be taken up 'metaphorically', as in having it coincide with aboutness, as do Reinhart (1981) and most Formal Functionalists (see for example Kruijff-Korbayov and Steedman (2003) for this insight, and Newmeyer (2001) for a description of Formal Functionalism). Taken more literally, as is common in most studies with a Systemic Functional orientation, it is regarded as referring to the first element in the linear order of the sentence in English (cf. Fries, 1981; Huddleston, 1988; Matthiessen & Halliday, 2014).

6.4.2 The Cheapness principle and the Thematic Progression test

Although both Hahn and Strube (1997) as well as Friedrichs and Palmer (2014) have recognised that thematic progression types may be identified in utterance transitions patterns in centered data, neither seems to offer a comprehensive mapping of one framework onto the other. Such an overlap between the textual patterns of thematic progression and the utterance transitions in Centering Theory seems within reach, however, and may be best illustrated using the 'tests' that Kibble (2001) has identified for distinguishing between Rule 2 transitions. Next to the Salience-test and the Coherence-test,

[47] For an excellent, early analysis of the bi-componential nature of the Mathesius-theme, see for example Fries (1981). Givn (1990, p. 902fn15) has also commented on these two different elements of the general Prague School theme, that of 'being given information' and 'being talked about', as have Reinhart (1981) and Smith (2003).

which result in the canonical 2x2 utterance transition matrix (cf. figure 6.2), the insights of the Functional Centering proposals by Strube and Hahn are of relevance here.

As Kibble (2001) has recognised, the question whether $C_b(U_i) = C_p(U_{i-1})$ as proposed by Strube and Hahn is indeed an underlying core issue in Centering Theory. As an intermediary to arriving at this Cheapness test, however, Strube and Hahn (1996, 1999) propose an additional test, $C_p(U_i) \stackrel{?}{=} C_p(U_{i-1})$, for utterances that have passed the Salience test. Incidentally, this test seems exactly the central distinction between the Type 1 and 2 thematic progression patterns of Dane (1974). That is, given a co-referential entity that is maximally salient in the second utterance of a pair (theme/$C_p(U_i)$), the question is whether this entity is also the theme of the previous utterance, or whether it is derived from a lower ranking position (its rheme).

Given this overlap between the additional utterance transition requirement of Strube and Hahn and the Thematic Progression patterns, we will use the term 'Thematic Progression test' to refer to this particular distinction. Adding this test as a full-fledged condition to the canonical utterance transition matrix creates a model that is able to incorporate both CT transitions as well as TP patterns.

6.4.3 Transition tests as separate dimensions

In contrast to Strube and Hahn, who take the Thematic Progression test $C_p(U_i) \stackrel{?}{=} C_p(U_{i-1})$ as some form of intermediate condition for utterances that satisfy the Salience test (cf. Strube & Hahn, 1999, p. 333), we choose to regard it as a third dimension in the transition matrix. Figure 6.5 provides a model which incorporates all three independent 'tests' in Strube and Hahn (1999) as separate dimensions, aiding in the distinction between utterance transitions. (Note, however, that although the first two dimensions reflect the canonical set of transitions, the expressions for the Salience test have been reversed to reflect the relative importance of the forward-looking centre in our current account.)

Although we envision that the canonical CT utterance transition matrix offers a view of this cube from the front, it is immediately obvious from this figure that the four canonical CT transitions do not appear in one plane at the front of the cube. Both the fact that the Retain and Rough-shift transitions are found in the back, as well as the fact that the Expensive Continue and Expensive Smooth-shift are found *behind* their regular counterparts Continue and Smooth-shift, are direct consequences of the canonical CT perspective not acknowledging the depth in the matrix added by the Thematic Progression test. However, as a result of CT generally being blind to the dimension added by the TP-test, as well as the lower two front quadrants being empty (see below), nothing intervenes from seeing the Retain and Rough-shift transitions from the canonical CT point of view (i.e., facing the cube from the front). This is

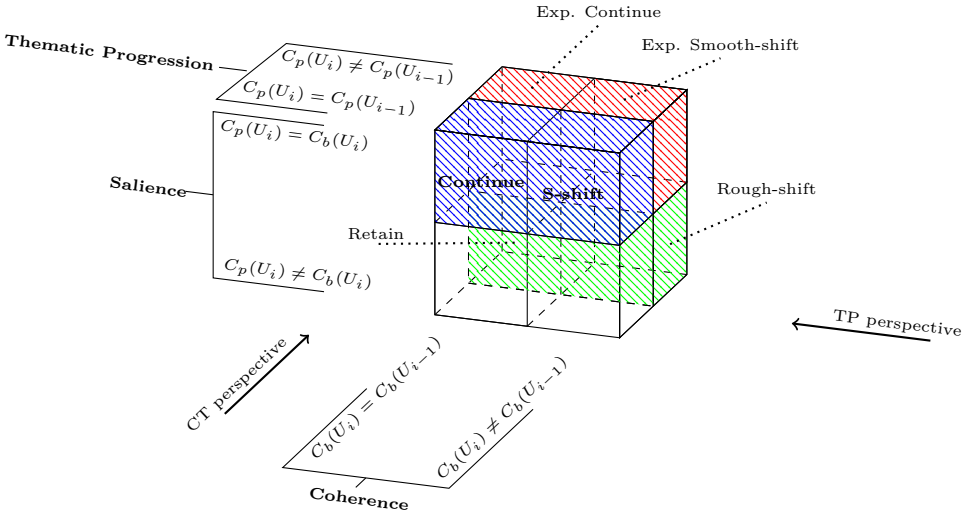


Figure 6.5: Hybrid CT/TP-model of utterance transitions

not the case for the Expensive transitions, however, which is why we suggest these transitions have remained 'under the radar' until Strube and Hahn (1999). Conversely, it has already been observed that there is no attention in the model by Dane (1974) for what is expressed by way of the Coherence test. This is illustrated best by the fact that the Continue and Smooth-shift, Expensive Continue and Expensive Smooth-shift, as well as Retain and Rough-shift can all be collapsed into respectively three blocks containing two transitions each (cf. the coloured blocks in figure 6.5). On the basis of the Thematic Progression literature, no distinction is thus possible between the transitions within these blocks (for the identification of Retain and Rough-shift as TP3 transitions, see below). Given the current model, we suggest the TP framework looks at this utterance matrix cube 'from the right'.

It was already mentioned that the lower two quadrants at the front of the cube are empty. This reflects the fact that the filling of these slots is theoretically impossible. As observed by Strube and Hahn, "Retain and Rough-shift fulfill $C_p(U_{i-1}) \neq C_p(U_i)$ without further extensions" (Strube & Hahn, 1999, p. 333). The observation that these transitions do not comply with the Thematic Progression test seems valid, but can be clarified somewhat. Using centering terminology, any $C_p(U_{i-1})$ that is realised in U_i would necessarily have to be $C_b(U_i)$ due to Constraint 3, since there is no situation in which $C_p(U_{i-1}) = C_p(U_i)$ while at the same time $C_p(U_i) \neq C_b(U_i)$. Stated differently, this statement becomes somewhat of a truism: if there is no coreferential link between the entities in two successive utterances, the current theme cannot be the previous utterance's theme. As a result, Retain and Rough-shift necessarily appear in the back instead of the front of the cube, where they satisfy $C_p(U_{i-1}) \neq C_p(U_i)$.

In terms of the Thematic Progression perspective, the TP test helps to distinguish

between TP types 1 and 2, as was noted above. Since these TP types have to satisfy the precondition that there is a coreferential link between two utterances, the transitions under scrutiny occupy the top plane of the cube, where they both satisfy the Saliency test ($C_p(U_i) = C_b(U_i)$). The Continue and Smooth-shift transitions comprise the blue block, Continuous Theme/TP type 2, which corresponds to the requirement that the discourse entity that realises the current theme is the same entity that realises the previous utterance's theme ($C_p(U_i) = C_p(U_{i-1})$). Expensive Continues and Expensive Smooth-shifts in red, on the other hand, show the Thematisation of Rhemes (TP type 1) pattern: the discourse entity that realises the theme of the current utterance is not co-referential with the previous utterance's theme ($C_p(U_i) \neq C_p(U_{i-1})$), but since $C_p(U_i) = C_b(U_i)$, it must necessarily realise a discourse entity from the *rheme* of the previous utterance (i.e., any $C_f U_{i-1}$ barring the preferred center).

The green block at the back of the bottom plane contains utterances that may be identified as hypertheme (TP type 3). However, our analysis that this category coincides with the Retain and Rough-shift transitions is not entirely uncontroversial, and warrants some further discussion.

Type 3, Retain/Rough-shift, new themes and 'unlinked' centering transitions

In the formal discourse structure accounts of Hahn and Strube (1997) and Smith (2003), the TP type 3 pattern is claimed to resist the intuitive establishing of connections between consecutive utterances. Instead, such utterances are thought to be related only by inferential linking: Hahn and Strube do not make use of the hypertheme-pattern (TP 3) because its inclusion would require conceptual (i.e., bridging) inferences which are beyond the scope and focus of their hierarchical discourse structure algorithm (Hahn & Strube, 1997, p. 105), and Smith points to the requirement of salient poset relationship which are inferrable from the discourse context for the licensing of such connections (Smith, 2003, p. 246; note the connection to indirect realisation in CT). Here, we argue that it is possible to (partly) fit these TP 3 patterns into a formal model of centering utterance transitions, although this requires a change of perspective with regard to mainstream CT (reversing the expressions in the Salience test, as is done in figure 6.5, may aid in this changed point of view).

Based on the notion that the theme/topic is identical to C_p, we argue that the Retain and Rough-shift transitions can be gathered under the label of TP 3. That these transitions have not as yet been recognised as Type 3 transitions is not entirely surprising, since their inclusion in this category depends not so much on the pattern that they show in terms of the traditional CT perspective (i.e., a non-salient C_b in the second member of a pair of utterances, or equivalently: a C_b being relegated to the

'rheme'), but rather what they, at first sight, lack, namely a C_p which is linked to the immediately preceding utterance. Exactly because $C_p(U_i) \neq C_b(U_i)$, this opens up a 'free' C_p slot which paves the way for filling by an entity that is unlinked to the previous utterance: either a completely brand-new entity, or an entity that is available in the long-range co(n)text.[48] In this way, Retain and Rough-shift transitions, but in essence any utterance for which $C_p(U_i) \neq C_f(U_{i-1})$, are natural members of the hypertheme Thematic Progression pattern (type 3; but see 6.4.3).

In addition, in both the TP as well as the canonical CT framework, the possibility of completely brand-new themes or C_ps seems somewhat underdeveloped. As has just been argued, although canonical CT can accommodate such 'new' C_ps in the Retain and Rough-shift transitions, the framework to date has focused almost exclusively on the entity that establishes the co-referential connection to the previous utterance, i.e., the C_b. For example, CT is usually applied with the assumption that utterances cannot be C_b-less, as is noted above. Similarly for TP, and based on the three main TP types, the promotion to theme of any referential entity seems to require at least *some* form of earlier mention: either developed out of the themes (TP2) or rhemes (TP1) of the immediately preceding utterance, or by way of bridging inferences to entities in the previous discourse (TP3). To the extent that it is theoretically possible, completely brand-new (unmediated, in the scheme of Strube & Hahn, 1999) discourse entities, which are inferentially unlinked or uninferrable with respect to anything that came before, are not well accounted for. This suggests that although Dane allows texts to vary in degree of (textual) coherence, at least a certain degree of referential coherence seems in some way implicitly assumed.[49]

In fact, the class of TP 3 utterance transitions is probably larger than just the Retain and Rough-shift transitions. A problem for the detection of particularly the Retain and Rough-shift transitions in practical applications of CT, and maybe particularly in text domains that do not deal with a finite set of referential entities, seems to be that although a shift to a new C_p may be observed, the link to the previous utterance (in the guise of the C_b) is not realised at all. The unlinked Zero and Null transitions, outlined in section 6.2.1, are examples of transitions which may not seem to violate textual coherence intuitively, but in which the sheer lack of a C_b leads to an incoherent transitions in CT terminology. Strictly speaking, such transitions satisfy both $C_p(U_i) \neq C_p(U_{i-1})$ as well as $C_b(U_i) \neq C_p(U_i)$, and could therefore be classified as TP type 3.

[48] Note that this analysis also crucially hinges on the identification of theme as the most salient entity, C_p, and not the given entity, C_b.

[49] It needs to be remarked that the TP framework presented in Dane (1974) is not intended as a full-fledged theory of discourse structure, but primarily as a descriptive model of frequently recurring discourse patterns found in the texts studied. The author indeed remarks that other, minority thematic progression patterns may be distinguished (but unfortunately without providing any specific examples; cf. Dane, 1974).

However, given that there is no C_b to begin with, no distinction is possible on the basis of the Coherence and Salience tests, and these utterances therefore cannot be incorporated in a model such as figure 6.5, which is based on the traditional Rule 2 matrix.[50] Although we will assign thematic progression patterns for such transitions irrespective of their centering status (C_b, C_f-ranking and tests for transitions), they are a special class of transitions slightly outside the model proposed here (cf. the status of additional CT transitions in table 6.3).

Lastly, the shift of focus from C_b to C_p as the theme of the utterance also has some bearing on the Retain-Shift hypothesis, which has received wide attention in the Centering Theory literature (cf. Brennan et al., 1987; Grosz et al., 1995; Karamanis, 2003; Kibble, 2001; Poesio et al., 2004a; Strube & Hahn, 1999). In general, this hypothesis states that over three utterances, the relegation of a $C_b(U_i)$ to a lower salient position in a consecutive utterance U_{i+1} (leading to Retain for the first transition), signals an intention to (Smooth-)shift the center in U_{i+2}. Poesio et al. (2004b, p. 330) have already shown that empirical support for this pattern in their non-narrative text domains is generally lacking, and although Karamanis cites a number of studies which have tentatively supported its existence, problems surrounding the patterns attested leads the author to rename it the *Retain:Smooth-shift inadequacy* of CT Rule 2 (Karamanis, 2003, p. 43). Given our identification of C_p as theme or aboutness-topic, we suggest that the *actual* shift (in aboutness) has already taken place in a Retain transition: because the backward-looking center is no longer the most salient entity, aboutness has shifted to this other entity that *is* the most salient entity of the utterance under scrutiny. In other words, it is not Givenness that should provide the most important cue for aboutness (as is still maintained in Brennan et al., 1987, and later CT studies; p. 155), but rather: Salience.[51] In addition, as already illustrated by the caution with which the original formulation of the Retain-Shift hypothesis in postulated i.e.,"Retaining may be a way to signal an intention to shift", Brennan et al. (1987, p. 156), there may be other ways in which a shift can come about, which seems more in line with empirical findings.

Problematic case: C_f-ranking in U_{i-1} when more than one 'rhematic center' is realised in U_i

Friedrichs and Palmer (2014) draw a link between a number of Rough-shift transitions in their corpus and the rheme-to-theme pattern of thematic progression (TP 1). In the scheme just presented, however, Rough-shifts are analysed as generally leading to

[50] Similarly, the Establish transition has a backward-looking center but cannot satisfy the Coherence test because there is no previous C_b.

[51] See also e.g., Vallduv (to appear) for the alternative C_b = aboutness-topic view. However, it is also important to note that irrespective of whether one adopts a view of aboutness as based on Salience or Givenness, the formal Centering mechanism already incorporates both notions.

a TP type 3 pattern, which would not fit the intuitive analysis of the example provided by Friedrichs and Palmer. We use example (11) above, reprinted here as (12), to illustrate this difference. Note that the example in (12) reflects centering annotation by Friedrichs and Palmer (2014), with backward-looking centers again underlined and the co-referential antecedents in the previous utterance in italics, while our analysis indicates backward-looking centers with square brackets where it deviates from Friedrichs and Palmer (2014):

(12) a. Two dozen scientists reported results with variations of the *experiments* [...] by Fleischmann and Pons. [-]

b. The *experiments*$_{C_b}$ involve plunging the two *electrodes* into "heavy" water. [Establishment]

c. When an [electric current] is applied to the *electrodes*$_{C_b}$, the heavy *water* did begin to break up, or dissociate. [Rough-shift]

d. Ordinarily the [breakup of the *water*$_{C_b}$] would consume almost all of the electrical energy. [Rough-shift]

(= example (2) in Friedrichs & Palmer, 2014)

Part of the reason for our diverging analyses rests on the fact that Friedrichs and Palmer take the assigning of theme to be based on givenness (i.e., theme as known information), and the concomitant identification of the C_b as the theme of the utterance. Hence, linear progression (TP1) ensues in their example (2c), here: (12-c), since "electrodes" is picked up from the rheme of the previous utterance. In contrast, since we do not take the theme to be identical with C_b but rather with C_p (i.e., aboutness as based on salience instead of givenness), and given that $C_p(U_i) \neq C_b(U_i)$ in the examples of Friedrichs and Palmer (2014, cf. the Rough-shift label), we analyse this example as showcasing TP 3. In our analysis, the theme of (12-c) is the 'new' entity "electric current" (= $C_p(U_{2c})$, indicated by brackets), and not the entities "electrodes" or "heavy water" which are in competition for C_b-status. However, we note that in formal CT approaches the exact identification of C_p, and accordingly the TP pattern, additionally depends on whether a distinction should be made between matrix clauses and subordinate clause in C_f-ranking, which as yet seems an unresolved issue (cf. Kameyama, 1998; Poesio et al., 2004b).

Beyond this particular case, the example provided by Friedrichs and Palmer does indeed flag up a problematic situation. It concerns cases where the current theme is derived from the previous rheme, but (1) there is more than one rhematic entity from the previous utterance in the current utterance, and (2) the entity that is realised as the current utterance's theme is in a lower rank in the previous utterances C_f-list based on traditional ranking criteria. Under such circumstances, we seem to be dealing

with linear theme progression (TP1), but because the current theme is not the highest ranked entity realised in the current utterance (cf. Constraint 3), the Salience test is not satisfied and we end up with a Retain or Rough-shift transition. In our current model, such transitions are necessarily analysed as TP3, which does not match the intuitive analysis of their Thematic Progression pattern. There does not seem to be a straightforward way out of this: the main problem is that CT's Constraint 3 states that $C_b(U_i)$ *must* be the highest-ranked entity realised from the previous utterance, whereas TP only distinguishes between the most salient entity in the previous utterance in absolute terms (i.e., its C_p) versus any other entity that may also be realised in that utterance.

The obvious culprit here is C_f-ranking, and although some utterance-level information might improve results of C_f-ranking (e.g., functional centering's reliance on familiarity status), it is also observed that success rates for any form of automated C_f-ranking come with a margin of error; be it based on naive textual order, grammatical ranking or familiarity status of nominal referents, or other stringent ways to generate a preferred order of referential entities (cf. Poesio et al., 2004b; Strube & Hahn, 1999).

Here, we do not adhere too strictly to such predictions based on C_f-ranking for two reasons. First of all, this particular problem crucially involves the connection between sentence planning and text planning, i.e., that the entity which will be the theme of the next utterance is (partly) determined by where it surfaces in the current utterance. We remain agnostic with regard to this point, and do not assume that there *is* a necessary connection between the two (cf. Kibble, 2001). Second, we are dealing with a body of diachronic texts for which it is doubtful that there is consistency and/or homogeneity with respect to conventions that govern the processes which we have denoted here as sentence planning and text planning. It is assumed here that late 16^{th}-century texts have been composed with different discourse conventions in mind for what constitutes a good sentence, as well as how to connect such sentences successfully in a piece of prose, than a late 19^{th}-century text, for example.

In general, we do not assume a strong correlation between the position of a referential entity in the current utterance and the likelihood with which it will be taken up as the most salient entity of the next utterance. In addition, since proposed methods of C_f-ranking are not yet able to sufficiently capture the ways in which sentence-level structural prominence marking (i.e., salience/aboutness) is achieved in a corpus of diachronic texts, we will not rely on automated C_f-ranking. Rather, we manually assign the status of both C_p as well as Cb, and will only make distinction between the forward-looking center that has C_p-status (theme) and any other (unranked) C_fs (rheme). Cases such as (12-c) will be noted, however, in which there is more than one non-thematic entity from the previous utterance that is realised in the current utterance, but where

the traditional assigning of *Cb*-status based on grammatical role ranking deviates from the identification of a C_b by intuitive assigning of a co-referential link between the entity that is the current theme and its position in the previous utterance. In a canonical CT scheme, Constraint 3 necessitates such utterances to be analysed as Retain/Rough-shift, but because we assign no particular C_f-list ranking (and by extension: ignore or flout Constraint 3), such cases are more easily accommodated in the current model.

6.4.4 Manual annotation of referential centers of attention

A full-fledged account of referential coherence should be able to deal with demonstrative and relative pronouns with a clear referential value (i.e., a coreferential antecedent), as well as referents in the discourse co-text (i.e., indicated by exophoric antecedents such as first and second person pronouns). Failing such a mechanism, manual theme assignment seems the best way to assign centering status. Thus, the ranking of the C_fs to find the C_p is done manually here (i.e., 'searching for C_p/C_b and theme' rather than 'computing C_p algorithmically').

For example, 1^{st}- and 2^{nd}-person pronouns, both singular and plural forms, are taken as implicitly available as part of the C_f-list. Although they may not be explicitly available in the co-text, the usual domain of centering theory, we take these pronouns to be implicitly available in the context of any form of discourse in which a speaker/author and a hearer/reader can be envisioned. Thus, 2^{nd}-person pronouns are taken to be likely candidates for centers of attention in most instructional writing in general, which may also be deduced from their use in other non-narrative prose domains (Poesio et al., 2004b, see p. 318 for the status of 1^{st}- and 2^{nd}-person pronouns in CT). In general, we are not after a computational model for referential annotation, but rather a system with textual explanatory power. That we can couch TP in CT terms is just a way add a level of depth in terms of formality to this text linguistic framework.

Benefits of manual annotation:

1. Able to take into account propositional (e.g., clausal) referential antecedents. In other words, not limited/restricted to entity coherence only, and also includes coreferential antecedents for 'problematic' R1-pronouns (Poesio et al., 2004b, p. 319-320) such as demonstratives (and see also the 'other' class of errors in Strube & Hahn, 1999, p. 327-328)

2. Allows inclusion of TP type 3 and 'long-range' antecedents

3. Bases the assigning of TP pattern labels on a formal framework for utterance transitions

6.4.5 Aims of a case study

Centering Theory is a framework based on present-day conceptions of referential coherence in discourse, and is specifically informed by intuitions regarding the coherence of *English*. This fact neatly fits an agenda which is based on the assumption that early modern English texts may seem incoherent to a modern-day readership – at least more so than works produced more recently. This leads us to the following research questions:

- Are Early Modern English instructional texts more incoherent than present-day English works (as measured by CT-coherence)?
- Which linguistic and textual devices serve to guarantee referential coherence in Early and Late Modern English instructional writing?
- What, if any, is the relationship between degrees of (in)coherence and changes in prose conventions in the history of English?

6.5 Applying CT/TP on Modern English manuals

We have necessarily remained somewhat eclectic and expedient in terms of text selection for the current application of the combined account of CT and TP. Nevertheless, in keeping with the general topic-consistency criterion of this dissertation, a discourse topic is selected for our case study in section 6.5.1 that features in almost all text samples in the current corpus, namely the watering of the horse. Starting with the most modern samples and showing how CT/TP deals with patterns in a 21^{st}-century text sample, we then work our way backwards to illustrate where and how problems start arising.

6.5.1 Case study: Structuring the watering of horses

Davies (2009)

A first glance at the section on water in the text by Davies (2009) is notable for its frequent use of paralinguistic signs: the roughly two and a half pages on this topic contain at least three bullet-pointed lists involving the necessity of water for the bodily functions of the horse, percentages for the locations of water in the horse's body and the most common sources of water for horses. Perhaps somewhat unsurprising for a chapter entitled "Food and Biological Molecules", the author starts by briefly pointing out the necessity of water for horses and then quickly moves on to biological and biochemical properties which are relevant with respect to feeding, e.g., how water is transported

and where it is stored in the body. More practical advice concerning the watering of horses appears in the final paragraph on the subject of water, dealing with temperature regulation. It is provided here in full in example (13). In this, as well as the ensuing examples in this section, backward-looking centers (where available) are indicated in bold, with themes (C_ps) underlined and italics indicating the co-referent of the theme of the next utterance, if applicable. Square brackets indicate utterance transitions and curly brackets the thematic progression pattern.

(13) Temperature regulation

a. *Water* has excellent properties that make it very useful in temperature regulation.
[-] {TP3, hypertheme (cf. previous context)}

b. Water has *physical properties*, [*sic.*] which readily enable transfer of heat, build-up of heat and the loss of large amounts of heat through vaporisation.
[Establishment] {TP2, constant theme}

c. These basic properties act in addition to other physiological factors such as the circulation of large volumes of blood (fluid), large surface areas for evaporation on the skin and within the lungs.
[Exp. Smooth-shift] {TP1, linear theme}

d. Horses are able to divert blood to the body surfaces for *sweating* when the horse is getting too warm and also divert blood away from the body surface in cold ambient temperatures.
[Rough-shift] {TP3, hypertheme (bridging)}

e. Loss of too much *water* from the horse's body will cause dehydration and eventually death and therefore it is important for the body to readily absorb *water* from the digestive tract and maintain fluid balance by absorbing *water* via the kidneys when required.
[Rough-shift] {TP3, hypertheme (bridging)}

f. Water is lost from the horse's body via urine, faeces, sweat and evaporation from the lungs and skin.
[Exp. Smooth-shift] {TP1, linear theme}

g. Withholding water has dangerous implications causing dehydration and colic.
[Rough-shift] {TP3, hypertheme (bridging)}

h. Loss of too much water from the digestive tract such as diarrhoea can be life threatening, particularly for young foals.
[Rough-shift] {TP3, hypertheme}

i. Losses of 5-10% of bodyweight have been found in competing endurance horses.
[Rough-shift] {TP3, hypertheme}

j. *Voluntary water intake* by horses at rest in a moderate temperature environment is roughly 25-70 ml/kg/day.
[Rough-shift] {TP3, hypertheme}

k. For a 500 kg horse this equates to *12.5-35 litres per day.*
[Smooth-shift] {TP2, constant theme}

l. *This* depends on the balance between water intake from the feed and drinking and water

losses.
[Exp. Smooth-shift] {TP1, linear theme}

m. If horses are given too much protein, i.e., in excess of requirements, more water will be required as excess protein must be broken down and the excess nitrogen removed via the urine.
[Rough-shift] {TP3, hypertheme (bridging)}

n. Increased salt intake also increases water intake.
[Rough-shift] {TP3, hypertheme (bridging)}

o. A supply of fresh clean water is vital for all horses (Figure 3.1).
[Rough-shift] {TP3, hypertheme (bridging)}
(Davies, 2009)

Given these 15 utterances, the majority of transitions (9 out of 14, approx. 64.28%) are analysed as hypertheme (TP3) utterances with no direct link to the previous sentence. Only 2 sentences assumed to continue the theme of the previous utterance (TP2; transitions to (13-b) and (13-k) and another 3 (i.e., sentences (13-c) (13-f) and (13-l)) pick up a theme from the rheme of the previous utterance (TP1).[52]

First of all, there exist many inferential links between entities in these utterances that rely on indirect realisation, which is somewhat to be expected given the informational, non-narrative discourse found in the current text extract. For example, the "large surface areas for evaporation on the skin" in utterance (13-c) is in all likelihood co-referential with "the body surfaces for sweating" and "body surface" in the following utterance (13-d). In a similar way, there is a likely co-referentail link between "the circulation of large volumes of blood (fluid)" and the non-finite "to divert blood" in (13-d). It would be quite simplistic to assign C_b-status to the entity *blood* in this case, and be done with question of whether a true C_b can be found for (13-d). Based on the utterance's content, however, the most likely theme for this utterance is "Horses" at the start of this sentence, and no real C_b is available in terms of its definition as outlined in section 6.2.1.

Similarly, although the transition to (13-e) in all likelihood involves an inferential link between "sweating" in the previous utterance and the 'loss of (too much) water', we do not take this as picking up a theme from the previous utterance's rheme (i.e., linear theme TP1), but rather as a hypertheme derived from the global discourse topic in this chapter by Davies. Although 'sweating' implies 'loss of water', we take such an excessive loss of water (cf. "too much") to be linked to the general topic of temperature regulation and water intake. This view is additionally supported by the fact that in the source text, utterance (13-e) is indented and thus signals the start of a new paragraph (under the general header of "temperature regulation").

[52] Although we have provided a thematic progression pattern for the very first utterance in (13-a), we have not added this case to the overall sum for the excerpt from Davies (2009).

The reason for including the entire phrase "Voluntary water intake by horses at rest in a moderate temperature environment" in utterance (13-j), and not just its nominal head "Voluntary water intake" rests on the idea that the postnominal dependent is essential in defining this entity. That is, in contrast to the previous utterance, where 'endurance horses in competition' are mentioned, the current utterance deals with a specific form of voluntary water intake that is both conditioned on a horse being at rest, and this situation in turn obtaining in a moderate temperature environment.

Note in addition that for the consecutive utterance, (13-l), we assign thematic status to a demonstrative which would not be able to receive center status in standard CT accounts. Although we acknowledge the practical focus on noun phrase entities in centering, from a purely center-of-attention point of view there does not seem to be any reason for not allowing non-nominal antecedents. Utterances (13-e) and (13-g) similarly illustrate utterances in which the theme is not a nominal entity. Such cases would have been an outright problem for an algorithmic approach to referential coherence using Centering Theory, but they can be dealt with more easily here.

One issue that poses a practical problem with respect to the identification of a thematic entity is seen in utterance (13-k), which showcases delimitation with an entity in fronted PP clause (a 500 kg horse) vs (this).[53] Using a reordering of clauses for retrieving themes in alternative reordering seems easy, e.g., *This equates to 12.5-35 litres per day for a 500 kg horse*, thus suggesting that the ambiguity due to the preposed PP has some effect on thematicity. Such cases, however, which can be gathered under the general banner of 'Frame setting', are as yet ill-understood in terms of aboutness according to Krifka (2007) and Krifka and Musan (2012). For example, "[f]rame setting, according to Jacobs (2001), is often not clearly differentiated from aboutness topic. And Chafe (1976), who stresses the difference between the two notions, uses the term topic for precisely this function" (Krifka & Musan, 2012, p. 31). According to Krifka and Musan (2012, p. 32), such cases involve the setting up of alternative frames and picking one out of these frames for the content of the ensuing sentence. In cases of preposed locative PPs, the frame-setter is said to specify a situation dimension for the predicate in the matrix clause. Based on the examples provided by Krifka and Musan (2012), such frame setters do not provide an aboutness-topic, which we follow here by assigning thematic status to 'this'.

A somewhat similar issue to (13-k) occurs in utterance (13-m), where the retrieval of a theme is made more difficult in the context of a subordinate *if*-clause. The phrases "horses", "too much protein" or "more water" all qualify as potential themes for this utterance. If we base our analysis solely on the matrix clause, "more water" would

[53] This utterance reminds of a sentence seen in an example from Friedrichs and Palmer (2014) above, (12-c), which contains two entities in a subordinate clause which are ranked over the "heavy water"-entity in the matrix clause.

receive thematic status. Conversely, the 'topicalisation' of the conditional subordinate clause could be interpreted as changing the ranking of utterance entities in terms of thematicity.

We can reorder the conditional in (13-m) which would put thematic emphasis on "(more) water" or, as an alternative, rephrase the conditional itself (e.g., *Horses that are given too much protein, i.e., in excess of requirements, will require more water as excess protein must be broken down and the excess nitrogen removed via the urine*). On the basis of a file card metaphor for aboutness-topics (cf. Reinhart, 1981; Krifka & Musan, 2012), such a rephrased sentence of (13-m) would neatly store the information in the 'comment constituent' under a file card labeled *horses that are given too much protein*. However, this rephrasing changes the structure of both matrix as well as subordinate clause to such an extent that it is unclear whether this agnate still reflects the sentence planning which resulted the original utterance by Davies. Here, we opt for "more water" as the theme of this utterance not so much on a reordering or on the basis of ranking of themes based on the status of the (matrix) clause, but rather on the hypertheme in this section of the text. In the utterances leading up to (13-m), we take the the salient underlying topic to be the amount of water intake, which depends on situational circumstances (climate, level of exertion, etc.). In that context, the subordinate *if*-clause is regarded as a frame setter that specifies a new circumstance, and the "more water" is judged to link up with the "this" in the previous utterance, meaning here 'more water [than this]' (i.e., the amount of water intake specified under the circumstances in (13-l)).

Especially in the last 6 sentences of the extract from Davies, i.e., (13-j) through (13-o), a considerable degree of utterances are unlinked (cf. the hypertheme TP3 pattern). Nevertheless, all themes appear to be centred around the hypertheme of water intake. That is, given the conditions of moderate temperatures and a general lack of exertion, water intake depends on the size of the horse and specific forms of nutrition which requires additional water intake. Although the last utterance, (13-o), is also concerned with the necessity of water intake, its status in this context is not entirely clear; it may have to be regarded as somewhat of an afterthought.[54]

Hunter (1796)

(14) *Watering of horses*

a. WATERING OF HORSES in a proper manner, is a matter on which the preservation of their health greatly depends, particularly while they are travelling.
[-] {TP2, constant theme (from title)}

[54]Incidentally, the "Figure 3.1" that is referred to in the text is a picture of a water trough with fresh clean water (Davies, 2009, p. 31).

b. That water which is least cold and penetrating being greatly preferable, for this reason, the water of a pond or river should be made choice of, rather than that of a well or fountain; but if you are obliged to make use of the latter, its piercing quality may, and ought to be meliorated by the addition of a little warm water and bran.
[Establishment (R-s)] {TP3, hypertheme (bridging)}

c. When a horse is out on a journey, *he* should be permitted to drink moderately of the first good water you come to after seven o'clock in a summer's morning, and nine or ten in the winter, and while he is drinking draw his head up three or four times, and make him move a little betwixt each draught, after which you may continue your journey smartly, but not too fast for six or seven miles more.
[Rough-shift] {TP3, hypertheme (bridging)}

d. *A horse* should during the whole of his journey be treated in this manner, for when you call to bait on the road, and the horse is hot, you must not let him drink for some time, as it might prove of bad consequence, and when his bridle is taken off, his thirst would prevent him from eating, unless he were permitted to drink immediately, or had been properly watered on the road, which would either occasion you to stop too long at a place, or soon incapacitate your horse from travelling.
[Smooth-shift] {TP2, constant theme}
(Hunter, 1796)

The first utterance in this text fragment, (14-a), continues the theme mentioned in the title of the entry in the dictionary of farriery terms by Hunter (1796). It is therefore analysed as a constant theme progression (TP2), although no CT label is provided (note that capitalisation in this utterance reflects that in the original source). Via a bridging inference or indirect realisation (from "watering" to the general global topic water), the transition to the second utterance ((14-b)) is taken to be a hypertheme pattern TP3. Had it not been for the fact that it is the second utterance in this paragraph in the text, and is thus necessarily categorised as an Establishment transition in CT terms, it would have been analysed as a Rough-shift transition compatible with the TP3 label. The third utterance in this example is indeed analysed as such a Rough-shift, with the theme changing from a focus on water in (14-b) to the horse in (14-c). This theme is maintained in utterance (14-d), resulting in the labelling of this transition as a Smooth-shift, accompanied by a TP2 constant theme. One may note, too, the frame setting subordinate clause in (14-c), which specifies that the horse in question (i.e., the theme in (14-c) and (14-d)) is a subset of the generic horse, namely a travelling horse.

Although an excerpt of merely four utterances in (14) cannot be taken as exemplary for the utterance transitions in this text by Hunter (1796), they illustrate how the topic of providing water to the horse deviates from its introduction in the 21^{st}-century sample taken from Davies (2009) above. In addition, given that we base our utterance boundaries on punctuation markers, the two hypertheme (TP3) and two constant theme (TP2) progression patterns offer perhaps only a glimpse of the transitions which a more refined model of discourse unit boundaries may have revealed (cf. the length in words of the sentences in examples (13) and (14)). While mention of providing water to the horse

may surface in the form of isolated sentences in other entries, the four utterances in (14) represents the full treatment of the global topic of watering horses in this particular source text.

Baret (1618)

(15) Chapter 22. Of *his Watering.*

a. Concerning the watering of a running horse there is a difference betwixt some mens opinions and mine, but because I have spoken thereof heretofore, I will but touch their opinions briefly, and referre the taking or leaving of them, as (in their iudgement) they shall finde reason to perswade.
[-] {TP2, constant theme (from title)}

b. 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?
[Establishment (R-S)] {TP3, hypertheme (long range context) }

c. For (first) watering in the house is very hurtfull, for if he be anything subiect to take cold, it will be nourished by *drinking cold water*, (if he be not heated after:) because that waterish humors are enemies to a horse, and will increase unnaturall superfluities: for as standing water doth putrifie & gather filth, so will a horse gather much corruption, and excrementall humors by *drinking cold water*, if the coldnesse be not mitigated by exercise, to disperse the naturall heate into all the parts of the body.
[Exp. Smooth-shift] {TP1, linear theme}

d. And further, if *hee* should drinke at any time (in the house) after labour, before his radicall moisture hath quenched that excesse of heate gotten by labour, the receiving of cold water suddenly into his body, (and not having any exercise to heate the same) will so oversway the quantity of naturall heat, that it will benumbe his body and make him shake, and so hazard either the mortall disease of foundring in the body, or else hinder concoction, with over raw humors, so that he cannot have good digestion, for want of which, unwholsome crudities will ingender.
[Exp. Smooth-shift] {TP1, linear theme}

e. And therefore water your Horse as little as you can in the house, but let him fetch it abroad, that you may heate it in his body by galloping, and so preserve him from such dangers as may insue thereby.
[Exp. Smooth-shift] {TP1, linear theme (bridging)}
(Baret, 1618)

Although no Centering transition is identified for the first utterance in the fragment of Baret, (15-a), its theme can be derived from the chapter title and results in a constant theme pattern (TP2). For the theme of (15-b), Baret seems to be harking back to topics discussed in the prior co-text. The beginning of this utterance ("Now ...") briefly reiterates these topics and prioritises them, in addition to jointly referring to these topics in the same interrogative sentence using the quantifier phrase "all these".[55] Since the derivation of the utterance transition label and thematic progression pattern

[55]Note that in this case, the fragment by Baret continues beyond the lines provided here. The remainder of the chapter addresses the second and third of these topics.

for this sentence relies on the previous discourse, we assume an Establishment transition and a long-range context for the hypertheme pattern here (TP3). The theme of (15-c) is more easily established, and results in an Expensive Smooth-shift transition with an associated linear theme (TP1) thematic progression. Identifying a theme in the utterance in (15-d) is complicated, but we take it to be underlined phrase "the receiving of cold water suddenly into his body". Partly due to the if-conditional in the first full clause of this sentence, we take the central clause in this utterance to be "the receiving of cold water suddenly into his body [...] will so oversway the quantity of naturall heat, that it will benumbe his body and make him shake" ("if hee should drinke", etc.). Given this theme, we assume another Expensive Smooth-shift, with the theme deriving from the gerund "drinking cold water" in utterance (15-c). The identification of the transition to (15-e) as a linear theme (TP1), from "hee" in (15-d) to "your Horse" in (15-e), relies on a bridging inference between a generic (running) horse and the horse of the reader.

As the example sentences from Baret (1618) show, deriving themes from texts in our corpus becomes more complicated the further we go back in time. Whereas in the 21^{st}-century fragment most themes neatly line up with the subjects of the sentence, such a mapping of grammatical role and thematic status is far less straightforward in earlier texts. In the next section, §6.5.2, we will therefore briefly evaluate the application of the current model on the texts in this cursory pilot study, and will draw particular attention to problems encountered in using CT-based concepts in practice.

6.5.2 Evaluation

The combination of TP patterns and CT utterance transitions labels shows surprising overlap, as may be expected on the basis of the model's outline in section 6.4. Nevertheless, a summary of utterance transitions indicates that although the discourse organisation can be mapped to some extent in this way, even short PDE fragments on a single global topic may indeed be far from CT-coherent, with (13) containing nine Rough-shifts, three Expensive Smooth-shifts, one regular Smooth-shift and one Establishment transition. More generally, this points to a difficulty in applying CT 'in the real': as was already seen in section 6.2.4, corpus-based CT experiments generally struggle with the identification of coherent utterance transitions, in particular due to a lack of a co-referential relationship between entities across two consecutive utterances.[56]

The short text sample drawn from Davies (2009) illustrates that in the current pilot, demonstratives (13-l), if-adjuncts (13-m), topicalised 'frame-setting' adjuncts (13-k)

[56]The studies listed in table 6.5 report that the lack of C_bs results in proportions of unlinked transitions to be as high as 83.4% in natural text (e.g., Friedrichs and Palmer (2014); see also footnote 27 on page 201).

and embedded centers in complex NPs (13-j) may indeed be found to be problematic for the identification of centers in PDE non-narrative texts. Having been flagged up by Poesio et al. (2004b), among others, they are also shown to impact the current application of a combined CT/TP model. Although manual annotation of centers partly resolves such problems by remaining agnostic as to salience ranking and by dealing with identification of problematic centers less stringently, this pilot underscores these two crucial problems for application of the CT framework. For example, in allowing indirect realisation, bridging inferences (see examples (13-c) and (13-d), respectively) or long-range dependencies shows the inadequacy of CT.

In addition, there remains the problem of identifying themes/C_ps in older texts which do not seem to adhere to Present-Day English conventions of prose writing. In lieu of most CT-based studies, we suggest sentences are not the best instantiation of utterance, particularly for pre-modern texts. Relying on punctuation to demarcate utterances (e.g., example (14-d) from Hunter (1796)) does not do justice to cognitive shifts in center of attention which is at the heart of CT. This is particularly obvious in examples such as (14), where, if we take the sentence as the 'discourse move', progression is not being well represented by CT-coherence metrics. In addition, explicit discourse markers that serve to signal shifts in the center of attention in such earlier texts (e.g., 'concerning X', 'for X'), or a particular rhetorical relationship between prior co-text and the current discourse move ('now whereas', 'and further'), may perhaps also license the use of more severe 'incoherent' utterance transitions in terms of CT. In the same way that Friedrichs and Palmer (2014) have suggested that some text domains generally might be more permissive of CT-incoherent transitions, the use of such discourse markers may mitigate the incoherence effect of these transitions. A more appropriate response to finding incoherent transitions would thus be not to tweak parameters so as to maximise CT-coherence, but rather, to use such metrics to investigate how incoherence may be licensed under certain grammatical or textual conditions. This puts focus on CT/TP as a descriptive tool for modelling (in)coherence, but not as a model for discourse coherence in general.

Beyond correctly identifying centers, however, the status of salience marking at the level of the sentence may be a more integral issue, since it directly affects one of the central components of CT's inner mechanics, C_f-ranking. As was already argued for in previous sections of this chapter, issues surrounding C_f-ranking cause serious problems for a correct application of CT-based models in terms of their ability to gauge discourse, or even cross-utterance, coherence. Incorporating more fine-grained knowledge of how salience marking is achieved at the level of the sentence seems necessary for a correct identification of utterance transitions. Inadequate modelling of salience at the level of the sentence, especially in the context of non-canonical IS configurations, will most

likely result in an inadequate depiction of the coherence of a discourse. Not surprisingly, CT is thus unable to fully capture the connection between sentence-level IS and supra-sentential DS.

Although Centering Theory's simplicity may be part of its appeal, it could thus also be said to be its Achilles' heel: too many additional settings have to be taken into account to provide the mechanics of the theory enough substance to account for the data in an acceptable fashion. Using the present efficient but coarse approach to salience (ranking) results in coherent utterance transitions being unjustly identified as incoherent. This major weakness seems to disqualify, or at least incapacitate, CT-based models in their current form as comprehensive frameworks that can bridge the gap between sentence-level Information Structure and supra-sentential Discourse Structure.

6.6 Chapter summary

In the current theory-driven chapter, we have sought to combine a computationally-oriented, formal theory of referential coherence with a descriptive, text linguistic framework for textual progression. After a review of the literature on both theories, we have proposed a model that exploits their overlap and combines separate analyses of the transitioning between utterances. Our main aim was to derive a framework that is able to make insightful the degree of (in)coherence in a text, without *a priori* assumptions of textual coherence. Applying this framework in a case studies on the current body of texts reveals first and foremost that more work is necessary on relevant patterns of discourse organisation in English in general. Although certain patterns can be detected for how topic-switches may be signalled or how discourse coherence is established in assumed topic-coherent text fragments in different periods of Modern English, comparing the degree of perceived coherence between texts, particularly from a diachronic perspective, is a complex matter and needs further study.

Chapter 7

Conclusion

7.1 Introduction

In the present chapter we will present our conclusions as based on the main findings drawn from chapters 3 through 6.

Since the three main sections in this dissertation offer different ways to identifying developments in periodic styles of discourse and discourse organisation, section 7.2 provides an overview of general results stemming from the various core chapters of this dissertation. Specifically, in the first part of this dissertation we have explored the current set of texts using a traditional corpus-based approach for measuring the parameter of "personal involvement", an indicator of periodic style in English written prose. Part II has sought to approach such varying styles using a data-driven, quantitative methodology which offers cues as to how low-level patterns of grammar may affect perceptions of style. This part of the dissertation has also illustrated how quantitative measures may aid in assessing (dis)similarities in idiosyncratic styles of writing for the texts in our corpus. By drawing on theories of referential coherence and textual progression, Part III of the current dissertation explores how texts from different periods in the history of English vary in their use of conventions for discourse organisation in the expression of comparable subject matter.

In the third section of this chapter, we will look at these results from a somewhat wider angle (cf. section 7.3). Here, we draw connections between findings across chapter boundaries and, as a result, combine aspects obtained through different methodological vantage points. In addition, this section will connect some of our results to the wider literature. Lastly, in the final section of this chapter we provide some directions for future research (section 7.4).

7.2 General results

In chapter 3, the use of 1^{st}-person singular and plural pronouns as well as 2^{nd}-person pronouns (not specified for number) was used as an indicator of a personally involved style of prose. As a feature associated with distinct periodic differences in prose styles, the significant decrease in 1^{st}-person singular and 2^{nd}-person singular as well as plural pronouns indicates a sharp divide between texts published before the start of the 19^{th}-century and those manuals published after this date. The general lack of personal involvement is corroborated by results stemming from chapter 4: texts in the later half of our corpus prefer to suppress the addressee of a directive, with the sharpest decline observed for the most involved manner to direct the addressee, by way of imperatives or with pronouns referring to either reader or writer. However, texts towards the end of the period covered by our corpus also seem to rely less on instructive utterances in general. This suggests that either the current sub-register, or our corpus sampling, gradually moves away from instruction and towards other text type categories, with descriptive utterances a primary candidate given the secondary discourse purpose of information transfer as outlined in the description of the current sub-register of popular lore.

The exploratory statistical approach of chapter 5, which is based on frequencies of local part-of-speech bundles, is able to identify clusters of texts in our corpus. These frequencies help to identify a group of Early Modern English manuals which employ comparable grammatical features, as well as a cluster texts published after the second half of the 19^{th}-century. Texts composed and published in the 18^{th}- and early 19^{th}-centuries take up an intermediate position between these two clusters, reflecting relatively average frequencies in terms of the use for certain grammatical clusters that are most distinctive of the early and late periods. These texts appear not nominal enough to be categorised as texts of the last two centuries covered by our corpus, nor verbal enough to be classified as texts from the Early Modern period, for example in terms of the use of modals and base forms of lexical verbs.

The last core chapter of the current dissertation, chapter 6, presents a theoretical model for gauging the organisation of discourse and level of referential coherence in these manuals. This chapter provides a model which integrates a computational linguistic model for entity coherence (Centering Theory) with a text linguistic model of textual progression (Thematic Progression). Applying such a model to text fragments drawn from our corpus shows that employing contemporary theories of discourse structure for mapping prose from earlier periods of Modern English faces serious challenges.

7.3 Towards profiles of periodic prose style

The previous section has summarised some of the exploratory results in the current dissertation. What remains to be established, however, is the extent to which these separate findings offer an insight into coherent, interconnected features of developing styles of prose. In other words, how far have we come towards establishing period-specific profiles of grammars of prose, and in particular, in addition to studies such as Biber and Finegan (1989) or Adamson (2000)?

With reference to the observation by Biber and Finegan (1997/2001) that "specialist, expository registers follow a different developmental course from popular written registers" (Biber & Finegan, 1997/2001, p. 81), we argue that on the basis of the current sub-register of popular lore, it is primarily the expository rather than the specialist (or, non-popular) that may have been an important factor in this development. This would link up with the conclusion by Biber and Gray that it is the influence of information purpose and conciseness in composing rather than influences of popularisation (and colloquialisation) which are deemed more important for the trajectory of registers of English (cf. Biber & Gray, 2012, p. 326).

However, it is likely that the current text domain has seen a register shift in terms of audience: from a wide, general readership in the Early Modern era to a more select stratum of society interested in writing on horse care in contemporary society. Importantly, this shift may have additionally involved a change from a popular and wide to a more specialist discourse community – the texts by Skeavington (c1840) and Fleming (1884) are written by veterinarians, and the text by Davies (2009) seems to be written from a point of view which takes for granted an educational background with a basic knowledge of biology and chemistry. Although it is impossible to assess with a great deal of precision the audience for which such manuals attempt to cater (even explicit mentions of a particular group of addressees may not always reflect the intended audience in Early Modern English manuals; cf. Taavitsainen, 1999, p. 246), it seems that writers of horse manuals have changed from semi-professional caretakers of horses among the gentry to professionals of veterinary care and nutrition. The effects of professional training, as noted by Halliday (2004, pp. 95-96) and Biber and Finegan (1997/2001, p. 82), may in turn have affected the more popular products of their writing.

That such different perspectives on stylistic conventions, as outlined in the chapters above, may be interrelated is evidenced by some of the patterns of occurrence of features investigated. In fact, the interconnectedness of such approaches may be illustrated by the fact that Adamson (2000, p. 608) refers to the stylometric research by Milic (1967) on the use of lexical bundles in the writings of Swift. Milic finds that Swift on average

starts every third sentence with a connective, and often even a double connective (*for although, but however*) or cluster of connectives (e.g., *and therefore if notwithstanding*; Adamson, 2000, citing Milic, 1967), which is interpreted in terms of a particularly opaque use of adverbial connectors in the 18^{th}-century.

We can trace similar patterns in our current data set. For example, one may observe a correspondence between a POS trigram found to be associated with the cluster of Early Modern English texts in chapter 5: a personal pronoun, modal verb and infinitive form of a lexical verb (PNP-VM0-VVI; e.g., *you should give*). We may connect the general decline of this POS trigram, which has double digit frequencies up until the end of the 17^{th}-century, with the distribution and trajectory of the use of Direct Address directives in the early section of the corpus (cf. chapter 4). In fact, the use of such a bundle may even be related to a decline in the use of 1^{st}- and 2^{nd}-person pronouns as attested in chapter 3 (this is based on pure speculation, however, since the PNP tag may refer to 1^{st}-, 2^{nd}- as well as 3^{rd}-person pronouns. Nevertheless, declining numbers for 1^{st}- and 2^{nd}-person pronouns certainly have not increased the frequency of the PNP-VM0-VVI towards the later end of our corpus). In addition, Biber and Gray (2012) associate modals and lexical verbs with a colloquial, spoken communicative demands (or even "colloquial style"? cf. Leech et al., 2012) of registers of English, which seems to be corroborated by the general characterisation of prose styles in the early versus late sections of our corpus.

.

Note too, for example, the use of the combination of a noun (mostly singular, but also plural nouns), a punctuation marker and a coordinating conjunction (NN1-PUN-CJC) in example (14) by Hunter (1796) in chapter 6. This example attest to the fact that within utterances with a single 'theme', one may find extensive paratactic sequences in prose styles different from that of instructive-informational writing in contemporary English.

7.4 Directions for future research

Although several suggestions have already been made in various parts strewn across preceding chapters, the current section outlines some general avenues for further research. For example, the statistical techniques used in chapter 5 only scratch the surface what is possible in terms of research in stylometry, especially after the surge in machine learning techniques. Similarly, features other than personal affect as exploited in chapters 3 and 4 may provide more discriminatory power in terms of the periodic profiling of texts in our corpus.

However, we would like to draw attention to more general extensions of the approach

in this dissertation here. For example, although expanding our corpus with more texts or larger text fragments will likely be beneficial for improving the current results, we may also suggest the use of a different, and possibly more comparable, body of texts for such explorations. That is, applying the approaches used here to a parallel corpus, e.g., translations of works that have repeatedly been translated in the history of English, may yield interesting material for the results obtained here. One such source may be Boethius' *De Consolatione Philosophiæ*, which has seen various translations in English, and as early as the 9^{th}-century AD.

Such parallel corpora seem to have their own challenges in diachronic research, however, in a number of confounding factors; not least of which the influence of a Latin original. In addition, uncertainty about whether later translators were influenced by both previous translations in the target language, as well as the exposure to translations in other target languages to which the translator has access, may distort the results of such a comparison. To this may also be added the influence of a particular national tradition for philosophical diction. Nevertheless, the results of an investigation at the meso-level of discourse of such a parallel diachronic corpus could provide important cues as to the consistency of results obtained in the present dissertation. That is not to say, however, that work on popular informational-instructive writing in the history of English is no longer necessary, with the trajectory of this (sub-)register of English warranting further study yet.

Lastly, we may highlight the need for a more substantial body of theoretical work on discourse structure and discourse organisation, in particular with reference to the status of discourse structure in (the history of) English. Considerable advances seem to have been made in recent decades in terms of the interface between syntax and discourse at the sentence level. Although Los et al. (2012, p. 4) note that the proliferation of concepts and definitions in IS is a mark of the field's comparative infancy, this is most certainly true for a more rigorous approach to discourse structure – the treatment of which often only receives lip service in what are, in essence, contributions on information structure. While the pragmatic sub-field of information structure has thus seen considerable unification in terms of concepts and definitions, research into its supra-sentence level sibling discourse structure is mainly being carried out using different theoretical underpinnings and using incompatible methodologies in various sub-branches of linguistics. Examples may be found in formal, computationally-oriented studies such as Grosz and Sidner (1986), mentioned in chapter 6 above, as well as theories involving rhetorical coherence such as segmented discourse representation theory (SDRT; cf. Lascarides & Asher, 2003) and rhetorical structure theory (RST Mann & Thompson, 1988). Of course, important 'informal', discourse analytic studies may be found in abundance, such as Brown and Yule (1983), Levinsohn (2009), van Dijk

(1997), to name but a few. Specifically with reference to English, such works may be complemented by text or corpus linguistic approaches such as Biber et al. (2007), as well as invaluable historical form-to-function approaches of discourse markers, discourse particles or intra- and intersentential connectors in the history of English (e.g., Brinton, 1996; Lenker, 2010; Meurman-Solin, 2012).

We suggest that future studies of discourse structure should be well-grounded in current theories of information structure at the level of clause and sentence, but should primarily focus on patterns of intersentential connection and global co-referential linking (cf. Brinton, 2015). In our view, the most promising approach for the study of discourse organisation continues work on the dynamics between grammar and discourse or information flow as seen in for example Chafe (2001), Fries (1994), Givn (1983). It is hoped that the current study has contributed to such explorations in historical perspective by outlining different methods which, in combination, may afford a glimpse into periodic conventions for achieving such discourse flow in prose composition.

Bibliography I.

Primary sources (source texts)

Baret, M. (1618). An hipponomie or the vineyard of horsemanship. monograph.

Blundeville, T. (1565). The fower chiefyst offices belongyng to horsemanshippe: part ii - the order of dietynge of horses. monograph.

Clifford, C. (1585). The schoole of horsmanship. monograph.

Davies, Z. (2009). Introduction to horse nutrition. monograph. Chichester: Wiley-Blackwell.

Duberstein, K. & Johnson, E. L. (2009/2012). How to feed a horse: understanding the basic principles of horse nutrition. electronic.

Fleming, G. (1884). The practical horse keeper. monograph.

Gibson, W. (1721). The true method of dieting horses. monograph.

Hunter, J. (1796). A complete dictionary of farriery and horsemanship. monograph.

Kirby, J. (1823). Farriery. Chapter in Encyclopaedia Britannica, 6th Edition, edited by Charles Maclaren. Edinburgh: A. Constable and Co.

Leighton-Hardman, A. C. (1977). A guide to feeding horses and ponies. monograph. London: Pelham.

Markham, G. (1607). Cauelarice, or the english horseman. monograph.

Matheson, D. (1921). The horse: in health, accident & disease. monograph. London: C. Arthur Pearson.

Morgan, N. (1609). The perfection of horse-manship, drawne from nature; arte, and practise. monograph.

Skeavington, G. (c1840). The modern system of farriery. monograph.

Speed, A. (1697). The gentleman's compleat jockey; with the perfect horseman, and experienc'd farrier. monograph.

Bibliography II.

References

Aarts, J. & Granger, S. (1998). Tag sequences in learner corpora. In S. Granger (Ed.), *Learner english on computer* (pp. 132–141). London: Longman.

Adamson, S. (2000). Literary language. In R. Lass (Ed.), *The cambridge history of the english language* (Vol. III: 1476–1776, pp. 539–653). Cambridge University Press.

Adolph, R. (1968). *The rise of modern prose style*. Cambridge, MA/London: MIT Press.

Alcorn, R. J. (2011). *Pronouns, prepositions and probabilities: a multivariate study of old english word order* (Doctoral dissertation, University of Edinburgh).

Algeo, J. (2006). *British or american english? : a handbook of word and grammar patterns*. Studies in English Language Studies in English Language. Cambridge: Cambridge University Press.

Allen, C. (1995). *Case marking and reanalysis. grammatical relations from old to early modern english*. Oxford: Clarendon Press.

Argamon, S., Koppel, M., & Avneri, G. (1998). Routing documents according to style. In *Proceedings of the first international workshop on innovative internet information systems (iiis-98)*.

Ariel, M. (2007). A grammar in every register? the case of definite descriptions. In N. Hedberg & R. Zacharsky (Eds.), *The grammar-pragmatics interface: essays in honor of jeanette k. gundel* (Chap. 12, pp. 265–292). John Benjamins.

Atkinson, D. (2001). Scientific discourse across history: a combined multi-dimensional /rhetorical analysis of the philosophical transactions of the royal society of london. In S. Conrad & D. Biber (Eds.), *Variation in english: multi-dimensional studies* (pp. 45–65). Harlow: Pearson.

Austin, J. L. (1962). *How to do things with words*. Harvard/Oxford: Clarendon Press.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Baayen, R. H. (2014). *'languager' (r manual)*.

Baayen, R. H., van Halteren, H., & Tweedie, F. (1996). Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, *11*(3), 121–131.

Bailey, R. W. (1979a). Authorship attribution in a forensic setting. In D. E. Ager, F. E. Knowles, & M. W. A. Smith (Eds.), *Advances in computer-aided literary and linguistic research*. AMLC.

Bailey, R. W. (1979b). The future of computational stylistics. *Association for Literary and Linguistic Computing Bulletin*, *7*, 4–11.

Baron, A. & Rayson, P. (2008). Baron, a. and rayson, p. (2008). vard 2: a tool for dealing with spelling variation in historical corpora. In *Proceedings of the postgraduate conference in corpus linguistics*.

Baroni, M. & Evert, S. (2009). Statistical methods for corpus exploitation. In A. Ldeling & M. Kyt (Eds.), *Corpus linguistics: an international handbook* (Vol. 2, pp. 777–803). Berlin/New York: Walter de Gruyter.

Beaver, D. I. (2004). The optimization of discourse anaphora. *Linguistics & Philosophy*, *27*(1), 3–56.

Bech, K. (2001). *Word order patterns in old and middle english: a syntactic and pragmatic study* (Doctoral dissertation, University of Bergen).

Bcue-Bertaut, M., Kostov, B., Morin, A., & Naro, G. (2014). Rhetorical strategy in forensic speeches: multidimensional statistics-based methodology. *Journal of Classification*, *31*, 85–106.

Beh, E. J. & Lombardo, R. (2014). *Correspondence analysis: theory, practice and new strategies*. Wiley series in probability and statistics. Chichester: Wiley.

Behagel, O. (1909). Beziehungen zwischen umfang und reihenfolge von satzgliedern. *Indogermanische Forschungen*, (25), 110–142.

Bendixen, M. (1996). A practical guide to the use of correspondence analysis in marketing research. *Marketing Research Online*, *1*, 16–38.

Benzcri, J. P. (1973). *L'analyse des donnes*. Paris: Dunod.

Berg, T. (2011). The modification of compounds by attributive adjectives. *Language Sciences*, *33*(5), 725–737.

Bestgen, Y. (2014). Inadequacy of the chi-squared test to examine vocabulary differences between corpora. *Literary and Linguistic Computing*, *29*(2), 164–170.

Biber, D. (1985a). Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics*, (23), 337–360.

Biber, D. (1985b). Spoken and written textual dimensions in english: resolving the contradictory findings. *Language*, *62*(2), 384–414.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1989). A typology of english texts. *Linguistics*, *27*(1), 3–43.

Biber, D. (1992). Using computer-based text corpora to analyse the referential strategies of spoken and written texts. In J. Svartvik (Ed.), *Directions in corpus linguistics: proceedings of the nobel symposium 82* (pp. 213–252). Berlin: Mouton.

Biber, D. (1995). *Dimensions of register variation.* Cambridge: Cambridge University Press.

Biber, D. (2001). Dimensions of variation among eighteenth-century speech-based and written registers. In S. Conrad & D. Biber (Eds.), *Variation in english: multi-dimensional studies* (pp. 200–214). Harlow: Longman.

Biber, D. (2003). Compressed noun-phrase structures in newspaper discourse: the competing demans of popularization vs. economy. In J. Aitchison & D. M. Lewis (Eds.), *New media language* (pp. 169–181). London: Routledge.

Biber, D. (2006). Register: overview. In K. Brown (Ed.), *Encyclopedia of language & linguistics* (Second Edition, pp. 476–482). Oxford: Elsevier.

Biber, D. & Clark, V. (2002). Historical shifts in modification patterns with complex noun phrase structures. In T. Fanego, M. J. Lpez-Couso, & J. Prez-Guerra (Eds.), *English historical morphology: selected papers from 11 icehl, santiago de compostela, 7– 11 september, 2000* (pp. 43–66). Benjamins.

Biber, D., Connor, U., & Upton, T. (2007). *Discourse on the move: using corpus analysis to describe discourse structure.* Studies in corpus linguistics. Amsterdam/Philadelphia: John Benjamins.

Biber, D. & Conrad, S. (2009). *Register, genre and style.* Cambridge University Press.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, *25*(3), 371–405.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: investigating language structure and use.* Cambridge approaches to linguistics. Cambridge: Cambridge University Press.

Biber, D. & Finegan, E. (1989). Drift and the evolution of english style: a history of three genres. *Language*, *65*(3), 487–517.

Biber, D. & Finegan, E. (1992). The linguistic evolution of five written and speech-based english genres from the 17th to the 20th centuries. In M. Rissanen, O. Ihalainen, T. Nevalainen, & I. Taavitsainen (Eds.), *History of englishes: new methods and interpretations in historical linguistics* (pp. 684–704). Berlin: de Gruyter.

Biber, D. & Finegan, E. (1997). Diachronic relations among speech-based and written registers in english. In T. Nevalainen & L. Kahlas-Tarkka (Eds.), *To explain the present: studies in the changing english language in honour of matti rissanen* (Vol. 52, pp. 253–275). Mmoires de la Socit Nophilologique de Helsinki. Helsinki: Socit Nophilologique.

Biber, D. & Finegan, E. (1997/2001). Diachronic relations among speech-based and written registers in english. In S. Conrad & D. Biber (Eds.), *Variation in english: multi-dimensional studies*. Harlow: Longman.

Biber, D. & Gray, B. (2012). The competing demands of popularization vs. economy: written language in the age of mass literacy. In T. Nevalainen & E. C. Traugott (Eds.), *The oxford handbook of the history of english* (pp. 314–328). Oxford/New York: Oxford University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written english*. London: Longman.

Binongo, J. N. G. (2003). Who wrote the 15th book of oz? an application of multivariate analysis to authorship attribution. *Chance*, *16*(2), 9–17.

Binongo, J. N. G. & Smith, M. W. A. (1999). The application of principal component analysis to stylometry. *Literary and Linguistic Computing*, *14*(4), 445–465.

Birner, B. & Ward, G. (1998). *Information status and noncanonical word order in english*. Amsterdam: John Benjamins.

Blackwell, S. (2000). Looking up *Look*: discourse markers in the bank of english. In J. M. Kirk (Ed.), *Corpora galore: analyses and techniques in describing english* (pp. 3–16). Amsterdam & Atlanta: Rodopi.

Blum-Kulka, S. & Olshtain, E. (1989). Requests and apologies: a cross-cultural study of speech act realization patterns (ccsarp). *Applied Linguistics*, *5*(3), 196–213.

Brennan, S., Friedman, M. W., & Pollard, C. J. (1987). A centering approach of prounouns. In *Proceedings of the 25th acl, stanford, ca* (pp. 155–162).

Bresnan, J. W. (2014, August). *Frequency effects in spoken syntax: 'have' and 'be' contraction*. Plenary talk presented at ISLE-3, Zurich. Retrieved from http://www.isle3.uzh.ch/static/files/104.xml

Brinton, L. J. (1996). *Pragmatic markers in english: grammaticalization and discourse functions*. Topics in English Linguistics 19. Berlin/New York: Mouton de Gruyter.

Brinton, L. J. (2015). Historical discourse analysis. In D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), *The handbook of discourse analysis* (2nd ed., Chap. 10, pp. 222–243). Chichester: Wiley-Blackwell.

Brook O'Donnell, M., Rmer, U., & Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics*, *18*(1), 83–108.

Brown, G. & Yule, G. (1983). *Discourse analysis*. Cambridge: Cambridge University Press.

Burnley, J. D. (1986). Curial prose in england. *Speculum*, (61), 593–614.

Burrow, J. A. (1982). *Medieval writers and their work: middle english literature and its background 1100–1500*. Oxford/New York: Oxford University Press.

Burrows, J. F. (1987). *Computation into criticism: a study of jane austen's novels and an experiment in method*. Oxford: Clarendon Press.

Burrows, J. F. (1992). Computers and the study of literature. In C. S. Butler (Ed.), *Computers and written texts* (pp. 167–204). Oxford: Blackwell.

Burrows, J. F. (1993). Tiptoeing into the infinite: testing for evidence of national differences in the language of english narrative. In S. Hockey & N. Ide (Eds.), *Research in humanities computing '92*. Oxford: Oxford University Press.

Burrows, J. F. (2002). 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, *17*(3), 267–287.

Butler, C. S. (2004). Corpus studies and functional linguistic theories. *Functions of Language*, *11*(2), 147–186.

Carter, R. & Nash, W. (1990). *Seeing through language: a guide to styles of english writing*. Oxford: Basil Blackwell.

Chafe, W. L. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. N. Li (Ed.), *Subject and topic*. New York/London: Academic Press.

Chafe, W. L. (1980). *The pear stories: cognitive, cultural and linguistic aspets of narrative production*. Norwood, New Jersey: Ablex Publishers.

Chafe, W. L. (1985). Linguistic differences produced by differences between speaking and writing. In D. R. Olson, N. Torrance, & A. Hildyard (Eds.), *Literature, language and learning: the nature and consequences of reading and writing.* (pp. 105–123). Cambridge: Cambridge University Press.

Chafe, W. L. (2001). The analysis of discourse flow. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (Chap. 35, pp. 671–687). Malden/Oxford: Blackwell.

Clark, H. H. (1977). Bridging. In P. N. Johnson-Laird & P. C. Wason (Eds.), *Thinking: readings in cognitive science*. London/New York: Cambridge University Press.

Clark, H. H. & Haviland, S. (1977). Comprehension and the given-new contract. In R. Freedle (Ed.), *Discourse comprehension and production*. Norwood, New Jersey: Ablex Publishers.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM review*, *51*(4), 661–703.

Cluett, R. (1976). *Prose style and critical reading*. New humanistic research. New York/London: Teachers College Press, Columbia University.

Cockayne, T. O. (1864). *Leechdoms, wortcunning, and starcraft of early england: being a collection of documents illustrating the history of science in this country before the norman conquest*. London: Longman, Green, Longman, Roberts and Green.

Crystal, D. (2015). *Making a point: the pernickity story of english punctuation*. London: Profile Books.

Culpeper, J. & Kyt, M. (2002). Lexical bundles in early modern english dialogues: a window into the speech-related language of the past. In T. Fanego, B. Mndez-Naya, & E. Seoane (Eds.), *Sounds, words, texts and change. selected papers from 11 icehl, santiago de compostela, 7-11 september 2000* (pp. 45–63). Current Issues in Linguistic Theory 224. Amsterdam: John Benjamins.

Culpeper, J. & Kyt, M. (2010). *Early modern english dialogues: spoken interaction as writing*. Studies in English Language. Cambridge: Cambridge University Press.

Curth, L. H. (2013). *'a plaine and easie waie to remedie a horse': equine medicine in early modern england*. History of science and medicine library. Leiden/Boston: Brill.

Dane, F. (1974). Functional sentence perspective and the organization of the text. In F. Dane (Ed.), *Papers on functional sentence perspective* (pp. 106–208). The Hague: Mouton.

Di Eugenio, B. (1998). Centering in italian. In M. A. Walker, A. K. Joshi, & E. F. Prince (Eds.), *Centering theory in discourse* (pp. 115–138). Oxford: Clarendon Press.

Diller, H. (2001). Genre in linguistic and related discourses. In H. Diller & M. Grlach (Eds.), *Towards a history of english as a history of genres* (pp. 3–43). Heidelberg: Winter.

Diller, H. & Grlach, M. (2001). *Towards a history of english as a history of genres*. Anglistische Forschungen. C. Winter.

Domnguez-Rodrguez, M. V. & Rodrguez-lvarez, A. (2015). The reader is desired to observe: metacomments in the prefaces to english school grammars of the eighteenth century. *Journal of Historical Pragmatics*, *16*(1), 86–108.

Downing, A. (1995). Thematic layering and focus assignment in chaucer's *General Prologue* to *The Canterbury Tales*. In M. Ghadessy (Ed.), *Thematic development in english texts* (pp. 147–163). Open linguistics series. London/New York: Printer.

Downing, A. (2001). Thematic progression as a functional resource in analysing texts. Retrieved from http://pendientedemigracion.ucm.es/info/circulo/no5/downing.htm

Dubois, B. L. (1987). A reformulation of thematic progression typology. *Text - Interdisciplinary Journal for the Study of Discourse*, *7*(2), 89–116.

Eder, M. (2015). Does size matter? authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, *30*(2), 167–182.

Ellis, N. C. (2003). Constructions, chunking, and connectionism: the emergence of second language structure. In C. Doughtly & M. Long (Eds.), *The handbook of second language acquisition* (pp. 63–103). Malden, MA: Blackwell.

Enkvist, N. E. (1973). *Linguistic stylistics*. The Hague: Mouton.

Erman, B. & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*(1), 29–62.

Ernestus, M., van Mulken, M., & Baayen, R. H. (2006). Ridders en heiligen in tijd en ruimte: moderne stylometrische technieken toegepast of oud-franse teksten [knights and saints in time and space: modern stylometric techniques applied to old french texts]. In *Taal & tongval* (Vol. 58, pp. 70–83).

Erteschik-Shir, N. (2007). *Information structure: the syntax-discourse interface*. Syntax and Morphology. Oxford: Oxford University Press.

Fairman, T. (2000). English pauper letters 1800-34 and the english letters. In D. Barton & N. Hall (Eds.), *Letter writing as a social practice* (pp. 63–82). Amsterdam/Philadelphia: John Benjamins.

Fairman, T. (2002). 'riting these fu lines': english overseers' correspondence, 1800-1835. In *Verslagen en mededelingen van de koninklijke academie voor nederlandse taal- en letterkunde* (Vol. 3, pp. 557–73).

Fairman, T. (2003). Letters of the english labouring classes and the english language, 1800-34. In M. Dossena & C. Jones (Eds.), *Insights into late modern english* (pp. 265–82). Bern: Peter Lang.

Firbas, J. (1964). On defining the theme in functional sentence analysis. In *L'cole de prague d'aujourd'hui* (Vol. 1). Travaux Linguistiques de Prague. Academia.

Fish, S. E. (1982/1973). What is stylistics and why are they saying such terrible things about it? In *Is there a text in this class?* Cambridge, MA/London: Harvard University Press.

Fish, S. E. (1982/1979). What is stylistics and why are they saying such terrible things about it? part ii. In *Is there a text in this class?* Cambridge, MA/London: Harvard University Press.

Fleischman, S. (1990). *Tense and narrativity: from medieval performance to modern fiction*. London: Routledge.

Fludernik, M. (1993). Second person fiction: narrative *You* as addressee and/or protagonist. *Arbeiten aus Anglistik und Amerikanistik*, *18*(2), 217–247.

Fludernik, M. (1995). Pronouns of address and 'odd' third person forms: the mechanics of involvement in fiction. In K. Green (Ed.), *New essays in deixis: discourse, narrative, literature* (pp. 99–129). Amsterdam & Atlanta: Rodopi.

Forsyth, R. S. (1995). *Stylistic structures: a computational approach to text classification* (Doctoral dissertation, University of Nottingham, Nottingham).

Fox, B. A. (1987). *Discourse structure and anaphora: written and conversational english*. Cambridge: Cambridge University Press.

Francis, G. (1989). Thematic selection and distribution in written discourse. *Word*, *40*, 201–221.

Friedrichs, A. & Palmer, A. (2014). Centering theory in natural text: large-scale corpus study. In J. Ruppenhofer & G. Faa (Eds.), *Proceedings of konvens* (pp. 137–144). Hildesheim, Germany.

Fries, C. C. (1952). *The structure of english: an introduction to the construction of english sentences*. New York: Harcourt, Brace.

Fries, P. H. (1981). On the status of theme in english: arguments from discourse. *Forum Linguisticum*, *6*, 1–38.

Fries, P. H. (1994). On theme, rheme and discourse goals. In M. Coulthard (Ed.), *Advances in written text analysis* (pp. 229–249). London: Routledge.

Fries, P. H. (1995). A personal view of theme. In M. Ghadessy (Ed.), *Thematic development in english texts* (pp. 1–19). Open linguistics series. London/New York: Printer.

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on computational linguistics* (pp. 611–617). Geneva.

Ghadessy, M. (Ed.). (1995). *Thematic development in english texts*. Open linguistics series. London: Pinter.

Givn, T. (1983). *Topic continuity in discourse: a quantitative cross-language study*. Amsterdam: Benjamins.

Givn, T. (1990). *Syntax: a functional-typological introduction*. Amsterdam/Philadelphia: Benjamins.

Givn, T. (1992). The grammar of referential coherence as mental processing instructions. *Linguistics*, *30*(1), 5–55.

Givn, T. (2001). *Syntax: an introduction* (2nd ed.). Amsterdam/Philadelphia: John Benjamins.

Glynn, D. & Robinson, J. (Eds.). (2014). *Corpus methods for semantics: quantitative studies in polysemy and synonymy*. Human Cognitive Processing. John Benjamins.

Gordon, I. A. (1966). *The movement of english prose*. English Language Series. London: Longman.

Gordon, P. C., Grosz, B. J., & Gillion, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, *17*, 311–348.

Grlach, M. (1992). Text-types and language history: the cookery recipe. In M. Rissanen, O. Ihalainen, T. Nevalainen, & I. Taavitsainen (Eds.), *History of englishes: new*

methods and interpretations in historical linguistics. Berlin/New York: Mouton de Gruyter.

Graham, N., Hirst, G., & Marthi, B. (2005). Segmenting documents by stylistic character. *Language Engineering*, *11*(4), 397–415.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Greenacre, M. J. (1988). Clustering the rows and columns of a contingency table. *Journal of Classification*, *5*(1), 39–51.

Greenacre, M. J. (2007). *Correspondence analysis in practice* (2nd edition). Boca Raton: Chapman & Hall/CRC.

Gregory, M. L. & Michaelis, L. A. (2001). Topicalization and left-dislocation: a functional opposition revisited. *Journal of Pragmatics*, (33), 1665–1706.

Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Snchez & M. Almela (Eds.), *A mosaic of corpus linguistics: selected approaches* (pp. 269–291). Frankfurt am Main: Peter Lang.

Gries, S. T. (2014). Frequency tables, effect sizes, and explorations. In D. Glynn & J. Robinson (Eds.), *Corpus methods for semantics: quantitative studies in polysemy and synonymy* (pp. 365–389). Amsterdam/Philadelphia: John Benjamins.

Gries, S. T., Newman, J., & Shaoul, C. (2011). N-grams and the clustering of registers. *ELR Journal*, *5*(1).

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 38th acl, cambridge, ma* (pp. 44–50).

Grosz, B. J. & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, *12*(3), 175–204.

Grosz, B. J. & Sidner, C. L. (1998). Lost intuitions and forgotten intentions. In M. A. Walker, A. K. Joshi, & E. F. Prince (Eds.), *Centering theory in discourse* (pp. 39–54). Oxford: Clarendon Press.

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, *21*(2), 203–225.

Grothendieck, G. (n.d.). *'gsubfn' (r manual)*.

Grund, P. (2011). The science of pronominal usage: he and it in coreference to inanimate entities in late middle english texts on alchemy. *Journal of English Linguistics*, *39*(4), 335–358.

Hahn, U. & Strube, M. (1997). Centering in-the-large: computing referential discourse segments. In *Proceedings of acl 97 / eacl 97* (pp. 104–111).

Halliday, M. A. K. (1967). Notes on transitivity and theme in english: part 2. *Journal of Linguistics*, *3*(2), 199–244.

Halliday, M. A. K. (1970). Language structure and language function. In J. Lyons (Ed.), *New horizons in linguistics* (pp. 140–164). Harmondsworth: Penguin.

Halliday, M. A. K. (1981). Linguistic function and literary style: an inquiry into the langague of william golding's *The Inheritors*. In D. C. Freeman (Ed.), *Essays in modern stylistics* (pp. 325–360). London: Methuen.

Halliday, M. A. K. (1985/2004). *An introduction to functional grammar* (4th ed.) (C. M. I. M. Matthiessen, Ed.). London: Edward Arnold.

Halliday, M. A. K. (1988). On the language of physical science. In M. Ghadessy (Ed.), *Registers of written english: situational factors and linguistic features*. Pinter.

Halliday, M. A. K. (2003 [1990]). New ways of meaning: the challenge to applied linguistics. In J. Webster (Ed.), *On language and linguistics* (Vol. 3 in the Collected Works of M. A. K. Halliday, pp. 139–174). London: Continuum.

Halliday, M. A. K. (2004). *The language of science* (J. Webster, Ed.). Collected works of M.A.K. Halliday: v. 5. London: Continuum.

Halliday, M. A. K. & Matthiessen, C. M. I. M. (1999). *Construing experience through meaning: a language-based approach to cognition.* Open linguistics series. London/New York: Cassell.

Hasan, R. & Fries, P. H. (Eds.). (1995). *On subject and theme: a discourse functional perspective*. Current Issues in Linguistic Theory. Amsterdam/Philadelphia: John Benjamins.

Hausenblas, K. (1964). On the characterization and classification of discourses. In *L' cole de prague d'aujourd'hui* (Vol. 1, *67–84*). Travaux Linguistiques de Prague. Academia, Editions de l'Academie Tchescoslovaque des Sciences.

Hilpert, M. & Gries, S. T. (2009). Assessing frequency changes in multistage diachronic corpora: applications for historical corpus linguistics and the study of. *Literary and Linguistic Computing*, *24*(4), 385–401.

Hiltunen, T. & Tyrrk, J. (2012, October). Normalising and tagging early modern english medical writing (1500 - 1700): a pilot study. In *Presentation at finsse*. Joensuu, Finland.

Hirst, G. & Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, *22*(4), 405–417.

Hobbs, J. R. (1985). *On the coherence and structure of discourse* (tech. rep. No. CSLI-85-37). Stanford University, Center for the Study of Language and Information. Stanford, California.

Holmes, D. (1985). The analysis of literary style – a review. *Journal of the Royal Statistical Society. Series A (General)*, *148*(4), 328–341.

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, *13*(3), 111–117.

Holmes, D. I. & Kardos, J. (2003). Who was the author? an introduction to stylometry. *Chance*, *16*(2), 5–8.

Hoover, D. L. (2003). Multivariate analysis and the study of style variation. *Literary and Linguistic Computing*, *18*(4), 341–360.

Hornik, K., Buchta, C., Hothorn, T., Karatzoglou, A., Meyer, D., & Zeileis, A. (2014). *'rweka' (r manual).*

Horobin, S. & Smith, J. J. (1999). A database of middle english spelling. *Literary and Linguistic Computing*, *14*(3), 359–374.

Huddleston, R. (1988). Constituency, multi-functionality and grammaticalization in halliday's functional grammar. *24*(1), 137–174.

Huddleston, R. (2002). Clause type and illocutionary force. In R. Huddleston & G. K. Pullum (Eds.), *The cambridge grammar of the english langauge* (Chap. 10). Cambridge: Cambridge University Press.

Huddleston, R. D. (1971). *The sentence in written english: a syntactic study based on an analysis of scientific texts*. Cambridge studies in linguistics. London: Cambridge University Press.

Huddleston, R. D. & Pullum, G. K. (Eds.). (2002). *The cambridge grammar of the english langauge*. Cambridge: Cambridge University Press.

Hudson, S. B. (1988). *The structure of discourse and anaphor resolution: the discourse center and the roles of nouns and pronouns* (Doctoral dissertation, University of Rochester).

Hurewitz, F. (1998). A quantitative look at discourse coherence. In M. A. Walker, A. K. Joshi, & E. F. Prince (Eds.), *Centering theory in discourse* (pp. 273–292). Oxford: Clarendon Press.

Hyland, K. (2001). Bringing in the reader: addressee features in academic articles. *Written Communication*, *18*(4), 549–574.

Hyland, K. (2002a). Authority and invisibility: authorial identity in academic writing. *Journal of Pragmatics*, *34*, 1091–1112.

Hyland, K. (2002b). Directives: argument and engagement in academic writing. *Applied Linguistics*, *23*(2), 215–239.

Hyland, K. (2005). Stance and engagement: a model of interaction in academic discourse. *Discourse Studies*, *7*(2), 173–192.

Irmer, M. (2009). *Bridging inferences in discourse interpretation* (Doctoral dissertation, Universitt Leipzig).

Jackendoff, R. (1997). *The architecture of the language faculty*. Cambridge, MA: MIT Press.

Jacobs, A. & Jucker, A. H. (1995). The historical perspective in pragmatics. In A. H. Jucker (Ed.), *Historical pragmatics: pragmatic developments in the history of english* (pp. 1–33). Amsterdam: John Benjamins.

Jauss, H. R. (1979). The alterity and modernity of medieval literature. *New Literary History*, *10*.

Jockers, M. L. (2013). *Macroanalysis: digital methods and literary history*. Topics in the digital humanities. Urbana: University of Illinois Press.

Jucker, A. H., Schneider, G., Taavitsainen, I., & Breustedt, B. (2008). Fishing for compliments: precision and recall in corpus-linguistic compliment research. In A. H. Jucker & I. Taavitsainen (Eds.), *Speech acts in the history of english*. Amsterdam/Philadelphia: John Benjamins.

Juola, P. & Baayen, R. H. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, *20*(Suppl. Issue), 59–67.

Kameyama, M. (1986). A property-sharing constraint in centering. In *Proceedings of the acl-86* (pp. 200–206).

Kameyama, M. (1998). Intra-sentential centering: a case study. In M. A. Walker, A. K. Joshi, & E. F. Prince (Eds.), *Centering theory in discourse* (pp. 89–112). Oxford: Clarendon Press.

Karamanis, N. (2003). *Entity coherence for descriptive text structuring* (Doctoral dissertation, University of Edinburgh, Edinburgh).

Kastovsky, D. (1982). Word formation: a functional view. *Folia Linguistica*, *16*, 181–198.

Kastovsky, D. (2006). Vocabulary. In R. Hogg & D. Denison (Eds.), *A history of the english language* (pp. 199–270). Cambridge Books Online. Cambridge University Press. Retrieved from http://dx.doi.org/10.1017/CBO9780511791154.005

Kehler, A. (2004). Discourse coherence. In *The handbook of pragmatics*. Blackwell.

Keiser, G. R. (1999). Practical books for the gentleman. In L. Hellinga & J. B. Trapp (Eds.), *The cambridge history of the book in britain* (Chap. 23, pp. 470–494). Cambridge: Cambridge University Press.

Kibble, R. (2001, December). A reformulation of rule 2 of centering theory. *Computational Linguistics*, *27*(4), 579–587.

Kohnen, T. (2001a). On defining text types within historical linguistics: the case of petitions/statutes. *European Journal of English Studies*, *5*(2), 197–203.

Kohnen, T. (2001b). Review of: leslie k. arnovick. *Diachronic Pragmatics: Seven Case Studies in English Illocutionary Development.* amsterdam/philadelphia, 1999. *Journal of Historical Pragmatics*, (2), 321–329.

Kohnen, T. (2002). Towards a history of english directives. In A. Fischer, G. Tottie, & H. M. Lehmann (Eds.), *Text types and corpora: studies in honour of udo fries.* Tbingen: Gunter Narr Verlag.

Kohnen, T. (2004). Methodological problems in corpus-based historical pragmatics. the case of english directives. In K. Aijmer & B. Altenberg (Eds.), *Advances in corpus linguistics: papers from the 23rd international conference on english language research on computerized corpora (icame 23), gteborg 22-26 may 2002.* Amsterdam/New York: Rodopi.

Kohnen, T. (2006). Variability of form as a methodological problem in historical corpus analysis: the case of modal expressions in directive speech acts. In C. Mair & R. Heuberger (Eds.), *Corpora and the history of english* (pp. 221–233). Heidelberg: Winter.

Kohnen, T. (2007). Text types and the methodology of diachronic speech act analysis. In S. M. Fitzmaurice & I. Taavitsainen (Eds.), *Methods in historical pragmatics* (52). Topics in English Linguistics [TiEL]. Berlin/New York: De Gruyter Mouton.

Kohnen, T. (2008a). Directives in old english: beyond politeness? In A. H. Jucker & I. Taavitsainen (Eds.), *Speech acts in the history of english.* Amsterdam/Philadelphia: John Benjamins.

Kohnen, T. (2008b). Tracing directives through text and time: towards a methodology of a corpus-based diachronic speech-act analysis. In A. H. Jucker & I. Taavitsainen (Eds.), *Speech acts in the history of english.* Amsterdam/Philadelphia: John Benjamins.

Kohnen, T. (2010). Genre in linguistic traditions: systemic functional and corpus linguistics. In A. S. Bawarshi & M. J. Reiff (Eds.), *Genre: an introduction to history, theory, research, and pedagogy* (Chap. 3). West Lafayette, Indiana: Parlor Press.

Komen, E. R. (2013). *Finding focus: a study of the historical development of focus in english* (Doctoral dissertation, Radboud University Nijmegen).

Kopaczyk, J. (2013). *The legal language of scottish burghs: standardization and lexical bundles (1380-1560).* Oxford: Oxford University Press.

Kopaczyk, J. & Jucker, A. H. (Eds.). (2013). Communities of practice in the history of english, Amsterdam/Philadelphia: John Benjamins.

Koppel, M., Akiva, N., & Dagan, I. (2003). A corpus-independent feature set for style-based text categorization. In *Ijcai-2003 work- shop on computational approaches to text style and synthesis.* Acapulco, Mexico.

Koppel, M., Argamon, S., & Shimoni, A. R. (2003). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *17*(4), 401–412.

Koppel, M. & Schler, J. (2003). Exploiting stylistic idiosyncracies for authorship attribution. In *Proceedings of the icaij03 workshop on computational approaches to style analysis and synthesis* (pp. 69–72).

Kornai, A. (2008). *Mathematical linguistics.* Advanced information and knowledge processing. London: Springer.

Kress, G. (1988). Textual matters: the social effectiveness of style. In D. Birch & M. O'Toole (Eds.), *Functions of style* (pp. 126–141). Pinter.

Krifka, M. (2007). Basic notions of information structure. In C. Fry, G. Fanselow, & M. Krifka (Eds.), *Interdisciplinary studies on information structure* (Vol. 6, pp. 13–56). Working papers of the SFB 632. Potsdam: Universittsverlag.

Krifka, M. & Musan, R. (2012). Information structure: overview and linguistic issues. In M. Krifka & R. Musan (Eds.), *The expression of information structure* (Chap. 1, 5, pp. 1–44). The Expression of Cognitive Categories (ECC). Berlin: De Gruyter Mouton.

Kruijff-Korbayov, I. & Steedman, M. (2003). Discourse and information structure. *Journal of Logic, Language and Information*, *12*(3), 249–259.

Lambrecht, K. (1994). *Information structure and sentence form : topic, focus, and the mental representations of discourse referents.* Cambridge studies in linguistics: 71. Cambridge : Cambridge University Press, 1994. Retrieved from http://ezproxy.lib.ed.ac.uk/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat00234a&AN=edinb.546221&site=eds-live

Lascarides, A. & Asher, N. (2003). *Logics of conversation.* Cambridge: Cambridge University Press.

Lau, P. (2015). *Semantic change and politeness: a study of verbs of commanding* (Master's thesis, University of Edinburgh, Edinburgh).

Lebart, L., Salem, A., & Berry, L. (1998). *Exploring textual data.* Text, speech and language technology. Boston/London: Kluwer Academic.

Lee, D. Y. (2001). Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the bnc jungle. *Language Learning & Technology*, *5*(3), 37–72.

Leech, G. & Short, M. (2007). *Style in fiction: an introduction to english fictional prose* (2nd ed.). Harlow: Pearson/Longman.

Leech, G., Smith, N., & Rayson, P. (2012). English style on the move: variation and change in stylistic norms in the twentieth century. *Language and Computers*, *76*(1), 69–98.

Lenker, U. (2010). *Argument and rhetoric : adverbial connectors in the history of english.* Topics in English linguistics. Berlin/New York: Mouton de Gruyter.

Lenker, U. (2011). A focus on adverbial connectors: connecting, partitioning and focusing attention in the history of english. Retrieved August 6, 2014, from http://www.helsinki.fi/varieng/journal/volumes/08/lenker/

Levelt, W. J. (1989). *Speaking: from intention to articulation*. Cambridge, MA/London: MIT Press.

Levinsohn, S. H. (2009). *Self-instruction materials on narrative discourse analysis*. SIL International.

Lijffijt, J., Nevalainen, T., Sily, T., Papapetrou, P., Puolamki, K., & Mannila, H. (2016). Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, *31*(2), 374–397.

Longacre, R. E. (1996). *The grammar of discourse*. Topics in language and linguistics. New York: Plenum Press.

Los, B. (2009). The consequences of the loss of verb-second in english: information structure and syntax in interaction. *English Language and Linguistics*, *13*(1), 97–125.

Los, B. (2012). The loss of verb-second and the switch from bounded to unbounded systems. In A. Meurman-Solin, M. J. Lpez-Couso, & B. Los (Eds.), *Information structure and syntactic change in the history of english*. Oxford studies in the history of English. New York: Oxford University Press.

Los, B., Lpez-Couso, M. J., & Meurman-Solin, A. (2012). The loss of verb-second and the switch from bounded to unbounded systems. In A. Meurman-Solin, M. J. Lpez-Couso, & B. Los (Eds.), *Information structure and syntactic change in the history of english* (Chap. 1). Oxford studies in the history of English. New York: Oxford University Press.

Mair, C., Hundt, M., Leech, G., & Smith, N. (2002). Short term diachronic shifts in part-of-speech frequencies: a comparison of the tagged lob and f-lob corpora. *International Journal of Corpus Linguistics*, *7*(2), 245–264.

Mann, W. C. & Thompson, S. A. (1988). Rhetorical structure theory: toward a functional theory of text organization. *Text*, *8*(3), 243–281.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, (13).

Martin, J. (1985). *Factual writing: exploring and challenging social reality*. Victoria, Australia: Deakin University Press.

Martin, J. (1992). *English text*. John Benjamins.

Martnez Insua, A. E. (2011). When discourse matches syntax: on meta-informative centering theory and discourse coherence in the recent history of english. *International Journal of English Studies*, *11*(2), 97–117.

Marttila, V. (2011). New arguments for new audiences: an corpus-based analysis of interpersonal strategies in early modern english medical recipes. In I. Taavitsainen & P. Pahta (Eds.), *Medical writing in early modern english* (pp. 135–157). Cambridge University Press.

Mathesius, V. (1928/64). On linguistic characterology with illustrations from modern english. In J. Vachek (Ed.), *A prague school reader in linguistics (reprinted from the hague, actes du premier congres international du linguistes)*. Bloomington: Indiana University Press.

Mathesius, V. (1939). On so-called functional sentence perspective. *Slovo a slovensnost*, (7), 169–180.

Matthiessen, C. M. I. M. & Halliday, M. A. K. (2014). *An introduction to functional grammar* (4th ed.). New York: Routledge.

Meurman-Solin, A. (2012). The connectives *And, For, But,* and *Only* as clause and discourse type indicators in 16th- and 17th-century epistolary prose. In A. Meurman-Solin, M. J.-C. Lpez-Couso, & B. Los (Eds.), *Information structure and syntactic change in the history of english*. Oxford studies in the history of English. Oxford: Oxford University Press.

Milic, L. (1967). *A quantitative approach to the style of jonathan swift*. The Hague: Mouton.

Milner, M. (2013). The physics of holy oats: vernacular knowledge, qualities and remedy in fifteenth-century england. *Journal of Medieval and Early Modern Studies*, *43*(2), 219–245.

Miltsakaki, E. & Kukich, K. (2000). The role of centering theory's rough-shift in the teaching and evaluation of writing skills. In *Proceedings of the 38th annual meeting on association for computational linguistics* (pp. 408–415). Association for Computational Linguistics.

Miltsakaki, E. & Kukich, K. (2004, March). Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, *10*, 25–55.

Mitchell, B. (1985). *Old english syntax*. Oxford: Clarendon Press.

Mitkov, R. (2010). Discourse processing. In A. Clark, C. Fox, & S. Lappin (Eds.), (Chap. 21, pp. 599–629). Wiley-Blackwell.

Moessner, L. (2001). Genre, text type, style, register: a terminological maze? *European Journal of English Studies*, *5*(2), 131–138.

Molony, V. & Warwick, C. M. (2013). Jeremiah kirby, author of 'farriery' in the 1806-1823 editions of the encyclopaedia britannica. *Veterinary History*, *16*(4), 353–360.

Mosteller, F. & Wallace, D. L. (1964). *Applied bayesian and classical inference: the case of the federalist papers*. Reading, MA: Addison-Wesley.

Murtagh, F. (2005). *Correspondence analysis and data coding with java and r*. CRC Press.

Nenadic, O. & Greenacre, M. J. (2014). *'ca' (r manual)*.

Nenadic, O. & Greenacre, M. J. (2007, February). Correspondence analysis in r, with two- and three-dimensional graphics: the ca package. *Journal of Statistical Software*, *20*(3), 1–13.

Nevalainen, T. (2006). Mapping change in tudor english. In L. Mugglestone (Ed.), *The oxford history of english* (pp. 178–211). Oxford: Oxford University Press.

Newmeyer, F. J. (2001). The prague school and north american functionalist approaches to syntax. *Journal of Linguistics*, *37*(1), 101–126.

Nunberg, G., Briscoe, T., & Huddleston, R. (2002). Punctuation. In R. Huddleston & G. K. Pullum (Eds.), *The cambridge grammar of the english langauge* (Chap. 12). Cambridge University Press.

Nwogu, K. & Bloor, T. (1991). Thematic progression in professional and popular medical texts. In E. Ventola (Ed.), *Functional and systemic linguistics: approaches and uses* (55, pp. 369–384). Trends in Linguistics. Studies and monographs.

Oakes, M. P. (2014). *Literary detective work on the computer*. Amsterdam/Philadelphia: John Benjamins.

Odenstedt, B. (1973). *The boke of marchalsi: a 15th century treatise on horse-breeding and veterinary medicine: edited from ms. harley 6398* (Doctoral dissertation, Stockholm University, Stockholm).

Oksanen, J. (2014). Cluster analysis: tutorial with r. Retrieved from http://cc.oulu.fi/~jarioksa/opetus/metodi/sessio3.pdf

Pahta, P. (2001). Creating a new genre: contextual dimensions in the production and transmission of early scientific writing. *European Journal of English Studies*, *5*(2), 205–220.

Pahta, P. & Taavitsainen, I. (1995). Authorities in medieval medical practices. *The New Courant*, (3).

Pahta, P. & Taavitsainen, I. (2014). Medical writing in early modern english.

Parkes, M. B. (1992). *Pause and effect: an introduction to the history of punctuation in the west*. Aldershot: Scolar Press.

Passonneau, R. J. (1998). Interaction of discourse structre with explicitness of discourse anaphoric noun phrases. In M. A. Walker, A. K. Joshi, & E. F. Prince (Eds.), *Centering theory in discourse* (pp. 327–358). Oxford: Clarendon Press.

Passonneau, R. J. & Litman, D. (1993). Feasibility of automated discourse segmentation. In *Proceedings of 31st annual meeting of the acl* (pp. 148–155). Columbus, OH.

Payne, J. & Huddleston, R. (2002). Nouns and noun phrases. In R. Huddleston & G. K. Pullum (Eds.), *The cambridge grammar of the english langauge*. Cambridge University Press.

Pearson, J., Stevenson, R., & Poesio, M. (2001). The effects of animacy, thematic role, and surface position on the focusing of entities in discourse. In M. Poesio (Ed.), *Proceedings of the first sempro*.

Peikola, M., Skaffari, J., & Tanskanen, S. (2009). *Instructional writing in english: studies in honour of risto hiltunen*. Pragmatics & Beyond New Series. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Perret, M. (1988). *Le signe et la mention: adverbes embrayeurs ci, a, la, iluec en moyen franais (xive-xve sicles)*. Genve: Droz.

Poesio, M., Cheng, H., Hitzeman, J., Stevenson, R., & Di Eugenio, B. (2002). *A corpus-based evaluation of centering theory*. Department of Computer Science, University of Essex.

Poesio, M., Stevenson, R., Eugenio, B. D., & Hitzeman, J. (2004a). *Centering: a parametric theory and its instantiations*. University of Essex.

Poesio, M., Stevenson, R., Eugenio, B. D., & Hitzeman, J. (2004b). Centering: a parametric theory and its instantiations. *Computational Linguistics*, *30*(3), 309–363.

Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical pragmatics* (pp. 223–255). New York: Academic Press.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the english language*. Harlow: Longman.

R Core Team. (2014). R: a language and environment for statistical computing. Retrieved February 1, 2015, from %7Bhttp://www.R-project.org/%7D

Rayson, P., Archer, D., Baron, A., Culpeper, J., & Smith, N. (2007). Tagging the bard: evaluating the accuracy of a modern pos tagger on early modern english corpora. In *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

Reinhart, T. (1981). Pragmatics and linguistics: an analysis of sentence topics. *Philosophica*, *27*, 53–94.

Reiter, E. & Dale, R. T. (2000). *Building natural language generation systems*. Cambridge/New York: Cambridge University Press.

Rissanen, M. (2000). Syntax. In R. Lass (Ed.), *The cambridge history of the english language* (Vol. III: 1476–1776, pp. 187–331). Cambridge: Cambridge University Press.

Rohdenburg, G. & Schlter, J. (2009). *One language, two grammars? differences between british and american english.* Studies in English Language Studies in English Language. Cambridge: Cambridge University Press.

Rudman, J. (2006). Authorship attribution: statistical and computational methods. In K. Brown (Ed.), *Encyclopedia of language & linguistics* (Second Edition, pp. 611–617). Oxford: Elsevier.

Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: a pain in the neck for nlp. In *Proceedings of the third international conference on intelligent text processing and computational linguistics (cicling 2002)* (pp. 1–15). Mexico City.

Sily, T., Nevalainen, T., & Siirtola, H. (2011). Variation in noun and pronoun frequencies in a sociohistorical corpus of english. *Literary and Linguistic Computing*, *26*(2), 167–188.

Sanders, T. & Maat, H. P. (2006). Cohesion and coherence: linguistic approaches. In K. Brown (Ed.), *Encyclopedia of language & linguistics* (Second Edition, pp. 591–595). Oxford: Elsevier.

Sandig, B. & Selting, M. (1997). Discourse styles. In T. A. van Dijk (Ed.), *Discourse as structure and process* (Vol. 1, pp. 9–43). Discourse Studies. A multidisciplinary introduction. London: Sage.

Santini, M. (2004). A shallow approach to syntactic feature extraction for genre classification. In *Proceedings of the 7th annual colloquium for the uk special interest group for computational linguistics.*

Sapir, E. (1921). *Language: an introduction to the study of speech.* New York: Harcourt, Brace and World.

Searle, J. R. (1975). Indirect speech acts. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics volume 3: speech acts* (pp. 59–82). New York: Academic Press.

Searle, J. R. (1976). A classification of illocutionary acts. *Language in Society*, *5*(1), 1–23. Reprint of: Searle, John R. (1975). 'A Taxonomy of Illocutionary Acts'. In: *Language, Mind, and Knowledge.* Ed. by K. Gnderson. Vol. 7. Minnesota studies in the philosophy of science. Minneapolis: University of Minnesota Press.

Sfinhufvud, A. C. (1978). *A late middle english treatise on horses* (Acta Universitatis Stockholmiensis, Stockholm studies in English 47, Stockholm University).

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

Sinclair, J. (2004). Corpus and text: basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: a guide to good practice.* Oxford: Arts and Humanities Data Service.

Smith, C. S. (2003). *Modes of discourse: the local structure of texts.* Cambridge: Cambridge University Press.

Smith, J. J. (2006). From middle to early modern english. In L. Mugglestone (Ed.), *The oxford history of english* (pp. 120–146). Oxford/New York: Oxford University Press.

Smith, J. J. (2012). Middle english: syntax. In L. Brinton & A. Bergs (Eds.), *English historical linguistics: an international handbook* (Chap. 28, Vol. 1, pp. 435–450). Handbcher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK) 34/1. Berlin: Mouton de Gruyter.

Smolensky, P. & Prince, A. (1983 [2004]). *Optimality theory: constraint interaction in generative grammar*. Malden, Mass: Blackwell.

Speyer, A. (2007). Die bedeutung der centering theory fr fragen der vorfeldbesetzung im deutschen [the relevance of centering theory for questions of vorfeld-positioning in german]. *Zeitschrift fr Sprachwissenschaft*, *26*(1), 83–115.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanties*, *35*, 193–214.

Stamou, C. (2008). Stylochronometry: stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, *23*(2), 181–199.

Stankiewicz, E. (1991). The concept of structure in contemporary linguistics. In L. R. Waugh & S. Rudy (Eds.), *New vistas in grammar: invariance and variation* (Vol. 49). Current Issues in Linguistic Theory. John Benjamins.

Stetter, C. (1991). Text und struktur. hat die sprechakttheorie eine historische dimension? In D. Busse (Ed.), *Diachrone semantik und pragmatik. untersuchungen zur erklrung und beschreibung des sprachwandels* (113, pp. 67–81). Reihe Germanische Linguistik. Tbingen: Niemeyer.

Stewart, L. L. (2006). Computational stylistics. In K. Brown (Ed.), *Encyclopedia of language & linguistics* (Second Edition, pp. 769–775). Oxford: Elsevier.

Stirling, L. & Huddleston, R. D. (2002). Deixis and anaphora. In R. Huddleston & G. K. Pullum (Eds.), *The cambridge grammar of the english langauge* (Chap. 17). Cambridge: Cambridge University Press.

Strube, M. & Hahn, U. (1996). Functional centering. In *Proceedings 34th annual conference of the association for computational linguistics (acl '96)* (pp. 270–277). Santa Cruz.

Strube, M. & Hahn, U. (1999). Functional centering: grounding referential coherence in information structure. *Computational Linguistics*, *25*(3), 309–344.

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

Stubbs, M. (1996). *Text and corpus analysis: computer assisted studies of language and culture*. Oxford: Blackwell.

Stubbs, M. & Barth, I. (2003). Using recurrent phrases as text-type discriminators: a quantitative method and some findings. *Functions of Language*, *10*(1), 61.

Swales, J. (1990). *Genre analysis: english in academic and research settings*. Cambridge: Cambridge University Press.

Taavitsainen, I. (2009). Authority and instruction in two sixteenth-century medical dialogues. In M. Peikola, J. Skaffari, & S. Tanskanen (Eds.), *Instructional writing in english: studies in honour of risto hiltunen* (Vol. 189). Pragmatics & Beyond New Series. Amsterdam/Philadelphia: John Benjamins.

Taavitsainen, I. (1993). Genre/subgenre styles in late middle english texts? In M. Rissanen, M. Kyt, & M. Palander-Collin (Eds.), *Early english in the computer age* (pp. 171–200). Berlin: Mouton de Gruyter.

Taavitsainen, I. (1994). On the evolution of scientific writings from 1375 to 1675: repertoire of emotive features. In F. Fernndez, M. Fuster, & J. J. Calvo (Eds.), *English historical linguistics 1992: papers from the 7th international conference on english historical linguistics, valencia, 22-26 september 1992* (pp. 329–342). Amsterdam/Philadelphia: John Benjamins.

Taavitsainen, I. (1997). Genre conventions: personal affect in fiction and non-fiction in early modern english. In M. Rissanen, M. Kyt, & K. Heikkonen (Eds.), *English in transition: corpus-based studies in linguistic variation and genre styles* (23). Topics in English Linguistics. Mouton de Gruyter.

Taavitsainen, I. (1999). Dialogues in late medieval and early modern english medical writing. In A. H. Jucker, G. Fritz, & F. Lebsanft (Eds.), *Historical dialogue analysis*. John Benjamins.

Taavitsainen, I. (2000). Metadiscursive practices and the evolution of early english medical writing 1375-1550. In J. M. Kirk (Ed.), *Corpora galore: analyses and techniques in describing english* (pp. 191–207). Amsterdam & Atlanta: Rodopi.

Taavitsainen, I. (2001a). Changing conventions of writing: the dynamics of genres, text types, and text traditions. *European Journal of English Studies*, *5*(2), 139–150.

Taavitsainen, I. (2001b). Language history and the scientific register. In H. Diller & M. Grlach (Eds.), *Towards a history of english as a history of genres* (Vol. 5, *2*, pp. 139–150). Heidelberg: C. Winter.

Taavitsainen, I. (2004). Genres of secular instruction: a linguistic history of useful entertainment. *Miscelnea: a Journal of English and American Studies*, (29), 75–94.

Taavitsainen, I. & Pahta, P. (2004). *Medical and scientific writing in late medieval english*. Studies in English language. Cambridge University Press.

Taavitsainen, I. & Pahta, P. (1995). Scientific 'thought-styles' in discourse structure: changing patterns in a historical perspective. In B. Warvik, S. Tanskanen, & R.

Hiltunen (Eds.), *Organization in discourse. proceedings from the turku conference.* (Vol. 14, pp. 519–529). Anglicana Turkuensia.

Taboada, M. & Zabala, L. (2008). Deciding on units of analysis within centering theory. *Corpus linguistics and linguistic theory.*

Tallentire, D. R. (1972). *An appraisal of methods and models in computational stylistics, with particular reference to author attribution* (Doctoral dissertation, University of Cambridge, Cambridge).

Tannen, D. (1982). Oral and literate strategies in spoken and written narratives. *Language*, *58*(1), 1–21.

Tanskanen, S., Skaffari, J., & Peikola, M. (2009). Approaching instructional writing in english. In M. Peikola, J. Skaffari, & S. Tanskanen (Eds.), *Instructional writing in english: studies in honour of risto hiltunen* (Vol. 189). Pragmatics & Beyond New Series. Amsterdam/Philadelphia: John Benjamins.

Taylor, A. & Pintzuk, S. (2012). Rethinking the ov/vo alternation in old english: the effect of complexity, grammatical weight, and information status. In T. Nevalainen & E. C. Traugott (Eds.), *The oxford handbook of the history of english* (pp. 835–845). Oxford/New York: Oxford University Press.

Thomas, S. (1999). Thematic networks and text types. *ASp*, (139-47).

Traugott, E. C. (1972). *A history of english syntax: a transformational approach to the history of english sentence structure.* New York: Holt, Rinehart and Winston.

Traugott, E. C. (1992). Syntax. In R. M. Hogg (Ed.), *The cambridge history of the english language* (Vol. I: The Beginnings to 1066, pp. 168–289). Cambridge: Cambridge University Press.

Traugott, E. C. (2004). Historical pragmatics. In L. R. Horn & G. Ward (Eds.), *Handbook of pragmatics* (pp. 538–561). Oxford: Blackwell.

Traugott, E. C. & Dasher, R. B. (2002). *Regularity in semantic change.* Cambridge studies in linguistics 96. Cambridge: Cambridge University Press.

Trost, P. (1962). Subjekt a predikt [subject and predicate]. In *Slavica pragensia* (Vol. 4, pp. 267–270). Universita Karlova.

Truswell, R. (2011). Relatives with a leftward island in early modern english. *Natural Language & Linguistic Theory*, *29*, 291–332.

Tummers, J., Speelman, D., & Geeraerts, D. (2012). Multiple correspondence analysis as heuristic tool to unveil confounding variables in corpus linguistics. In *Proceedings of the 11th international conference on the statistical analysis of textual data* (pp. 926–936). Lige: Presses Universitaires de Louvain.

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

Turan, U. (1998). Ranking forward-looking centers in turkish: universal and langauge-specific properties. In M. A. Walker, A. K. Joshi, & E. F. Prince (Eds.), *Centering theory in discourse*. Oxford: Clarendon Press.

Tyrrk, J. (2013). Exploring-part-of-speech profils and authorship attribution in early modern medical texts. In A. H. Jucker, D. Landert, A. Seiler, & N. Studer-Joho (Eds.), *Meaning in the history of english: words and texts in context* (Vol. 148). Studies in Language Companion Series. Amsterdam/Philadelphia: John Benjamins.

Vallduv, E. (to appear). Information structure. In M. Aloni & P. Dekker (Eds.), *The cambridge handbook of formal semantics*. Cambridge University Press.

van Dijk, T. A. (1997). The study of discourse. In T. A. van Dijk (Ed.), *Discourse studies: a multidisciplinary introduction* (Vol. 1: Discourse as Structure and Process, pp. 1–34). London: SAGE Publications.

van Kemenade, A. (2012). Rethinking the loss of verb-second. In T. Nevalainen & E. C. Traugott (Eds.), *The oxford handbook of the history of english* (pp. 822–834). Oxford/New York: Oxford University Press.

van Kemenade, A. & Los, B. (2006). Discourse adverbs and clausal syntax in old and middle english. In A. van Kemenade & B. Los (Eds.), *The handbook of the history of english* (pp. 224–248). Oxford: Blackwell.

Walker, M. A. (1989). Evaluating discourse processing algorithms. In *Proceedings of acl-89, vancouver, canada* (pp. 251–261).

Walker, M. A. (1993). Initial contexts and shifting centers. In *Proceedings of the workshop on centering*. University of Pennsylvania. Philadelphia.

Walker, M. A. (1996). Limited attention and discourse structure. *Computational Linguistics*, *22*(2), 255–264.

Walker, M. A. (1998). Centering, anaphora resolution, and discourse structure. In M. A. Walker, A. K. Joshi, & E. F. Prince (Eds.), *Centering theory in discourse*. Oxford: Clarendon Press.

Walker, M. A. (2000). *Toward a model of the interaction of centering with global discourse structure*.

Walker, M. A., Iida, M., & Cote, S. (1994). Japanese discourse and the process of centering. *Computational Linguistics*, *20*(2), 193–232.

Walker, M. A., Joshi, A. K., & Prince, E. F. (1998). Centering in naturally occurring discourse: an overview. In M. A. Walker, A. K. Joshi, & E. F. Prince (Eds.), *Centering theory in discourse*. Oxford: Clarendon Press.

Ward, G. (1988). *The semantics and pragmatics of preposing*. New York: Garland.

Ward, G., Birner, B., & Huddleston, R. D. (2002). Information packaging. In R. Huddleston & G. K. Pullum (Eds.), *The cambridge grammar of the english langauge* (Chap. 16). Cambridge University Press.

Weinert, R. & Miller, J. E. (E. (1998). *Spontaneous spoken language: syntax and discourse*. Oxford: Oxford : Clarendon Press.

Werlich, E. (1976). *A text grammar of english*. UTB Anglistik. Quelle & Meyer.

Wodarczyk, A. & Wodarczyk, H. (Eds.). (2013). *Meta-informative centering in utterances*. Studies in Language Companion Series. John Benjamins.

Zhao, Y. & Zobel, J. (2005). Effective and scalable authorship attribution using function words. *Information Retrieval Technology. Lecture Notes in Computer Science*, (3689), 174–189.

Appendix A

Part-of-Speech Tagset (CLAWS-5)

Table A.1: UCREL CLAWS-5 tagset

Tag†	Description
AJ0	adjective (unmarked) (e.g. GOOD, OLD)
AJC	comparative adjective (e.g. BETTER, OLDER)
AJS	superlative adjective (e.g. BEST, OLDEST)
AT0	article (e.g. THE, A, AN)
AV0	adverb (unmarked) (e.g. OFTEN, WELL, LONGER, FURTHEST)
AVP	adverb particle (e.g. UP, OFF, OUT)
AVQ	wh-adverb (e.g. WHEN, HOW, WHY)
CJC	coordinating conjunction (e.g. AND, OR)
CJS	subordinating conjunction (e.g. ALTHOUGH, WHEN)
CJT	the conjunction THAT
CRD	cardinal numeral (e.g. 3, FIFTY-FIVE, 6609) (excl. ONE)
DPS	possessive determiner form (e.g. YOUR, THEIR)
DT0	general determiner (e.g. THESE, SOME)
DTQ	wh-determiner (e.g. WHOSE, WHICH)
EX0	existential THERE
ITJ	interjection or other isolate (e.g. OH, YES, MHM)
NN0	noun (neutral for number) (e.g. AIRCRAFT, DATA)
NN1	singular noun (e.g. PENCIL, GOOSE)
NN2	plural noun (e.g. PENCILS, GEESE)
NP0	proper noun (e.g. LONDON, MICHAEL, MARS)
NULL	the null tag (for items not to be tagged)
ORD	ordinal (e.g. SIXTH, 77TH, LAST)
PNI	indefinite pronoun (e.g. NONE, EVERYTHING)
PNP	personal pronoun (e.g. YOU, THEM, OURS)
PNQ	wh-pronoun (e.g. WHO, WHOEVER)
PNX	reflexive pronoun (e.g. ITSELF, OURSELVES)
POS	the possessive (or genitive morpheme) ‘S or ’
PRF	the preposition OF
PRP	preposition (except for OF) (e.g. FOR, ABOVE, TO)
PUN	punctuation - general mark (i.e. . ! , : ; - ? ...)

Tag	Description
TO0	infinitive marker TO
UNC	"unclassified" items which are not words of the English lexicon
VBB	the "base forms" of the verb "BE" (except the infinitive), i.e. AM, ARE
VBD	past form of the verb "BE", i.e. WAS, WERE
VBG	-ing form of the verb "BE", i.e. BEING
VBI	infinitive of the verb "BE"
VBN	past participle of the verb "BE", i.e. BEEN
VBZ	-s form of the verb "BE", i.e. IS, 'S
VDB	base form of the verb "DO" (except the infinitive)
VDD	past form of the verb "DO, i.e. DID
VDG	-ing form of the verb "DO", i.e. DOING
VDI	infinitive of the verb "DO"
VDN	past participle of the verb "DO", i.e. DONE
VDZ	-s form of the verb "DO", i.e. DOES
VHB	base form of the verb "HAVE" (except the infinitive), i.e. HAVE
VHD	past tense form of the verb "HAVE", i.e. HAD, 'D
VHG	-ing form of the verb "HAVE", i.e. HAVING
VHI	infinitive of the verb "HAVE"
VHN	past participle of the verb "HAVE", i.e. HAD
VHZ	-s form of the verb "HAVE", i.e. HAS, 'S
VM0	modal auxiliary verb (e.g. CAN, COULD, WILL, 'LL)
VVB	base form of lexical verb (except the infinitive)(e.g. TAKE, LIVE)
VVD	past tense form of lexical verb (e.g. TOOK, LIVED)
VVG	-ing form of lexical verb (e.g. TAKING, LIVING)
VVI	infinitive of lexical verb
VVN	past participle form of lex. verb (e.g. TAKEN, LIVED)
VVZ	-s form of lexical verb (e.g. TAKES, LIVES)
XX0	the negative NOT or N'T
ZZ0	alphabetical symbol (e.g. A, B, c, d)

† Not including original CLAWS-5 quotation and bracketing tags

Taken from: `http://ucrel.lancs.ac.uk/claws5tags.html`

Appendix B

Most frequent POS categories in the corpus

Table B.1: Absolute frequencies, percentages and standardised (per 1,000) scores for the 15 most frequent POS categories (n= 53) in Blundeville (1565)

1565 Blu	NN1	PUN	AV0	PRP	CJC	AT0	AJ0	PNP	NN2	VVI	DT0	CJS	TO0	VVN	VBI
n	622	595	400	341	288	282	266	251	177	146	133	123	115	112	103
%	12.91	12.34	8.30	7.08	5.98	5.90	5.52	5.21	3.67	3.03	2.76	2.55	2.39	2.32	2.14
n/1,000	129.05	123.45	82.98	70.74	59.75	58.51	55.19	52.07	36.72	30.29	27.59	25.51	23.86	23.24	21.37

Table B.2: Absolute frequencies, percentages and standardised (per 1,000) scores for the 15 most frequent POS categories (n= 58) in Clifford (1585)

1585 Cli	NN1	PUN	PNP	AV0	PRP	CJC	VVI	AT0	AJ0	DPS	VVB	DT0	CJS	VM0	NN2
n	742	600	576	438	408	302	253	246	245	199	184	169	163	156	136
%	12.62	10.20	9.79	7.45	6.94	5.13	4.30	4.18	4.17	3.38	3.13	2.87	2.77	2.65	2.31
n/1,000	126.15	102.01	97.93	74.46	69.36	51.34	43.01	41.82	41.65	33.83	31.28	28.73	27.71	26.52	23.12

Table B.3: Absolute frequencies, percentages and standardised (per 1,000) scores for the 15 most frequent POS categories (n= 53) in Markham (1607)

1607 Mar	NN1	PUN	PNP	AV0	PRP	CJC	AT0	AJ0	VVI	DT0	NN2	DPS	VM0	CJS	VVB
n	691	559	398	381	365	304	300	240	199	156	152	145	142	133	122
%	13.53	10.94	7.79	7.46	7.14	5.95	5.87	4.70	3.90	3.05	2.98	2.84	2.80	2.60	2.39
n/1,000	135.25	109.41	77.90	74.57	71.44	59.50	58.72	46.98	38.95	30.53	29.75	28.38	27.80	26.03	23.88

Table B.4: Absolute frequencies, percentages and standardised (per 1,000) scores for the 15 most frequent POS categories (n= 53) in Baret (1618)

1618 Bar	NN1	PUN	PRP	AV0	PNP	CJC	AT0	AJ0	VVI	CJS	VM0	DT0	VVB	DPS	NN2
n	696	546	343	341	319	294	284	250	187	169	149	147	132	113	112
%	14.17	11.12	6.98	6.94	6.49	5.99	5.78	5.09	3.81	3.44	3.03	2.99	2.69	2.3	2.28
n/1,000	141.69	111.16	69.83	69.42	64.94	59.85	57.82	50.90	38.07	34.41	30.33	29.93	26.87	23.00	22.80

Table B.5: Absolute frequencies, percentages and standardised (per 1,000) scores for the 15 most frequent POS categories (n= 51) in Speed (1697)

1697 Spe	PUN	NN1	PRP	PNP	AT0	VVB	CJC	AJ0	AV0	NN2	CJS	DT0	VVI	VVN	DPS
n	677	645	388	367	338	311	301	268	258	198	139	114	99	98	96
%	13.49	12.85	7.73	7.31	6.73	6.20	6.00	5.34	5.14	3.95	2.80	2.27	1.97	1.95	1.91
n/1,000	134.89	128.51	77.31	73.12	67.34	61.96	59.97	53.40	51.40	39.45	27.69	22.71	19.73	19.53	19.13

Table B.6: Absolute frequencies, percentages and standardised (per 1,000) scores for the 15 most frequent POS categories (n= 53) in Gibson (1721)

1721 Gib	PUN	NN1	PRP	AV0	AJ0	AT0	CJC	NN2	DT0	PNP	PRF	CJS	VVN	DPS	VBB
n	989	851	642	567	548	465	413	373	370	334	241	230	197	147	140
%	12.25	10.54	7.95	7.02	6.79	5.76	5.12	4.62	4.58	4.14	2.99	2.85	2.44	1.82	1.73
n/1,000	122.52	105.43	79.53	70.24	67.89	57.61	51.16	46.21	45.84	41.38	29.86	28.49	24.41	18.21	17.34

Table B.7: Absolute frequencies, percentages and standardised (per 1,000) scores for the 15 most frequent POS categories (n= 55) in Hunter (1796)

1796 Hun	NN1	PUN	PRP	AT0	AV0	AJ0	CJC	PNP	NN2	CJS	VVN	PRF	VVB	VVI	DT0
n	651	590	390	355	355	340	302	251	175	137	137	131	121	105	102
%	12.81	11.61	7.67	6.99	6.99	6.69	5.94	4.94	3.44	2.70	2.70	2.58	2.38	2.07	2.01
n/1,000	128.12	116.12	76.76	69.87	69.87	66.92	59.44	49.40	34.44	26.96	26.96	25.78	23.81	20.67	20.07

Table B.8: Absolute frequencies, percentages and standardised (per 1,000) scores for the 15 most frequent POS categories (n= 51) in Kirby (1823)

1823 Kir	NN1	PUN	PRP	AT0	AV0	AJ0	CJC	NN2	VVN	PRF	PNP	DT0	CJS	VVI	VM0
n	701	573	418	397	348	327	235	226	189	174	173	153	121	108	94
%	13.71	11.21	8.18	7.77	6.81	6.40	4.60	4.42	3.70	3.40	3.38	2.99	2.37	2.11	1.84
n/1,000	137.10	112.07	81.75	77.64	68.06	63.95	45.96	44.20	36.96	34.03	33.84	29.92	23.67	21.12	18.38

Table B.9: Absolute frequencies, percentages and standardised (per 1,000) scores for the 15 most frequent POS categories (n= 55) in Skeavington (c1840)

1840 Ske	NN1	PUN	PRP	AT0	AJ0	AV0	CJC	PNP	NN2	VM0	VVN	DT0	VVI	CJS	PRF
n	685	630	404	352	318	244	244	215	196	151	140	139	138	123	115
%	13.52	12.43	7.97	6.95	6.28	4.82	4.82	4.24	3.87	2.98	2.76	2.74	2.72	2.43	2.27
n/1,000	135.18	124.33	79.73	69.47	62.76	48.15	48.15	42.43	38.68	29.80	27.63	27.43	27.24	24.27	22.70

Table B.10: Absolute frequencies, percentages and standardised (per 1,000) scores for the 15 most frequent POS categories (n= 52) in Fleming (1884)

1884 Fle	NN1	PUN	AT0	PRP	AJ0	AV0	NN2	CJC	PRF	VVN	DT0	CJS	VM0	VBZ	PNP
n	780	510	425	413	349	229	226	210	201	192	137	131	117	112	93
%	15.98	10.45	8.71	8.46	7.15	4.69	4.63	4.30	4.12	3.93	2.81	2.68	2.40	2.29	1.91
n/1,000	159.77	104.47	87.05	84.60	71.49	46.91	46.29	43.02	41.17	39.33	28.06	26.83	23.97	22.94	19.05

Table B.11: Absolute frequencies, percentages and standardised (per 1,000) scores for the 15 most frequent POS categories (n= 53) in Matheson (1921)

1921 Mat	NN1	PUN	AT0	PRP	AJ0	AV0	NN2	CJC	PRF	VVN	DT0	PNP	VBZ	CJS	VVI
n	761	496	422	414	377	303	295	215	189	164	151	129	123	101	101
%	14.93	9.73	8.28	8.12	7.40	5.95	5.79	4.22	3.71	3.22	2.96	2.53	2.41	1.98	1.98
n/1,000	149.30	97.31	82.79	81.22	73.97	59.45	57.88	42.18	37.08	32.18	29.63	25.31	24.13	19.82	19.82

Table B.12: Absolute frequencies, percentages and standardised (per 1,000) scores for the 15 most frequent POS categories (n= 50) in Leighton-Hardman (1977)

1977 Lei	NN1	AT0	PRP	PUN	AJ0	NN2	AV0	CJC	VVN	DT0	PRF	VM0	VVI	CJS	VBI
n	774	411	380	380	373	323	294	183	155	138	138	130	121	112	94
%	15.77	8.38	7.74	7.74	7.60	6.58	5.99	3.73	3.16	2.81	2.81	2.65	2.47	2.28	1.92
n/1,000	157.73	83.76	77.44	77.44	76.01	65.82	59.91	37.29	31.59	28.12	28.12	26.49	24.66	22.82	19.16

Table B.13: Absolute frequencies, percentages and standardised (per 1,000) scores for the 15 most frequent POS categories (n= 47) in Davies (2009)

2009 Dav	NN1	AJ0	PRP	PUN	NN2	AV0	AT0	CJC	VVN	PRF	DT0	VM0	VVI	VVB	VBZ
n	920	405	390	385	377	270	235	215	139	115	113	94	94	84	82
%	19.79	8.71	8.39	8.28	8.11	5.81	5.05	4.62	2.99	2.47	2.43	2.02	2.02	1.81	1.76
n/1,000	197.85	87.10	83.87	82.80	81.08	58.06	50.54	46.24	29.90	24.73	24.30	20.22	20.22	18.06	17.63

Appendix C

Correspondence analysis – Additional analyses

C.1 Correspondence analysis without PUN tags

Given an earlier suggestion concerning the potentially undesirable distortion caused by punctuation markers, we considered a look at the present data set of text samples and POS trigrams without those POS trigrams containing PUN tags. It turned out that of the 7,305 trigrams types used in the current corpus, 5,841 different trigrams (types) remained after discarding POS trigrams with PUN tags. Of this number of types, the proportion of hapaxes was approximately 48.11% for hapax legomena (n = 2,810) and 15.53% for dis legomena (907 types).

These trigrams without punctuation markers cover 35,330 observations (tokens). Recalling that the total number of observations in the total data set is 51,974, one-third of the tokens in the data set thus appears to contain a PUN tag ((1 - 35,330/51,774)*100 $\approx$ 32.02%). In turn, these tokens with punctuation markers make up roughly one-fifth of the types found in the complete corpus: (1- 5,841/7,305)*100 $\approx$ 20.04%, with the other four-fifths consisting of POS trigram types without PUN tags.[1]

For ease of comparison with the results in section 5.3.2, we select the 283 most frequent types of 'no PUN' trigrams, which together cover 50% (n = 17,680) of the 'no PUN' tokens, for a correspondence analysis. From the scree plot, it can be seen that the first three dimensions do not account for the same amount of inertia (57.6%) as in the correspondence analysis shown above in section 5.3.3 (which was 74.1% for the

[1] This suggests that the punctuation markers occur with relatively large frequencies, whereas the non-punctuation trigrams occur with more variety and lower frequencies. This should not be entirely surprising, of course, as the number of possible combinations (types) of three non-PUN tags is numerous in comparison, whereas the number of underlying tokens is only twice the size of the set of types containing a punctuation marker.

Table C.1: Scree plot for correspondence analysis ('no PUN' trigrams)

Dimension	value	%	cum%	scree plot
1	0.184983	35.0	35.0	*********
2	0.064652	12.2	47.2	***
3	0.055020	10.4	57.6	***
4	0.039717	7.5	65.1	**
5	0.030556	5.8	70.9	*
6	0.029851	5.6	76.5	*
7	0.027058	5.1	81.6	*
8	0.024007	4.5	86.2	*
9	0.020701	3.9	90.1	*
10	0.019826	3.7	93.8	*
11	0.017575	3.3	97.1	*
12	0.015159	2.9	100.0	*
Total inertia	0.529104	100.0		

first three dimensions), but that it is somewhat comparable to the inertia accounted for by the first three dimensions in section 5.3.2 (60.5% there). When we compare the absolute inertias in the original correspondence analysis of 50% of tokens and the current set, the inertias in both solutions are respectively 0.493153 for the solution with punctuation trigrams and 0.529104 for the solution based on non-punctuation tag trigrams.

When we take a look at the plots, the same basic pattern is observed in the relative position of text samples as in the solution in section 5.3.2. Some slight differences may be observed, however. First of all, the cloud of POS trigram tags and text samples is centred somewhat more around the origin (and as a result, the limits of the axis are reduced somewhat). In addition, the text by Davies (2009) is now positioned somewhat closer to that of Leighton-Hardman (1977), and both are positioned somewhat closer to the origin on the horizontal axis, but somewhat lower on the vertical axis in figure C.1.

With respect to dimension 3, the relative position between points in figure C.2a are the same as before, but their coordinates on this axis are inversed. Where previously Speed (1697) had a negative coordinate on dimension 3 while the 16^{th}- and other 17^{th}-century manuals had positive coordinates, in figure C.2a this latter group has negative coordinates on dimension 3 and Speed appears above the horizontal axis (coordinates on dimension 1 all the while appear to remain the same). The same is true for texts on the right hand side of the vertical axis, with positive coordinates on dimension 1 and inversed coordinates on dimension 3 (cf. the position of Davies).

The same effect is also found when we plot dimensions 2 and 3 in figure C.2b: coordinates of data points on dimension 2 remain the same (i.e., their vertical positioning), while the coordinates on dimension 3 have changed signs. In addition, it is

borne out again that the text samples by Davies and Leighton-Hardman appear closer together. This not only suggests that these texts are more similar when we compare their most frequent POS trigrams without PUN, but also that the absence of trigrams containing punctuation markers has an effect on the interpretation of dimension 2 in particular.

In sum, although there are minor variations in a correspondence analysis on POS trigrams devoid of punctuation markers, the general pattern in terms of the distances between text samples remains the same.

The results presented here indicate that using tags without PUN does not show a markedly better solution to the problem. Given that basically the same pattern is observed in the data even when we take out the PUN tag trigrams, the general clusters of text samples must be quite robust. Despite taking out one-third of the trigram tokens, and specifically, those that could supposedly have distorted our results, we see the same basic pattern in figure C.1 as before in figure 5.4. At least in terms of statistical interpretation, then, the results in this section do not appear to invalidate the methodology and findings presented in section 5.3.2.

With respect to textual interpretation, additional analyses might reveal underlying patterns in the data set. These were not carried out here, however, since it was deemed that the data set used above, including both tags for punctuation markers as well as other POS categories, offers a richer perspective of the data in general.

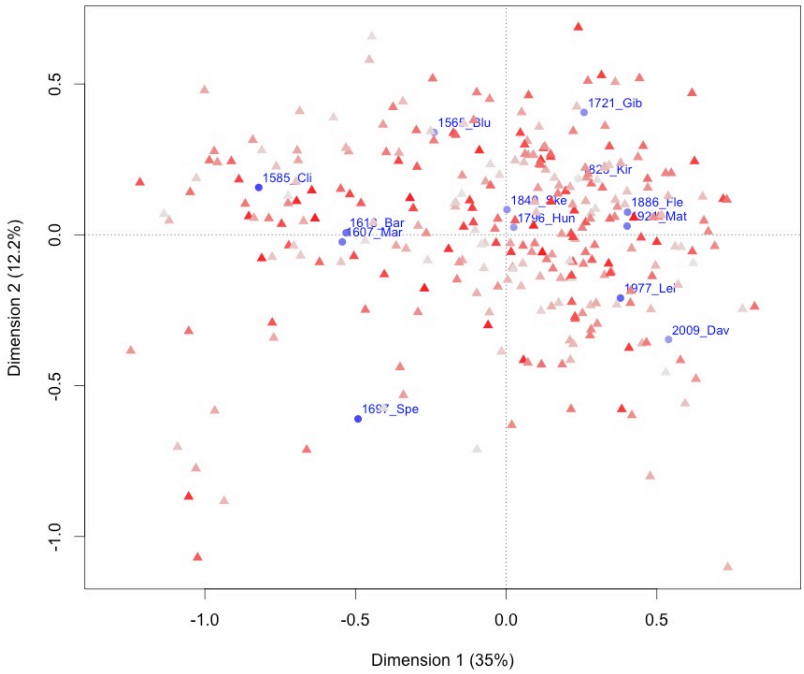
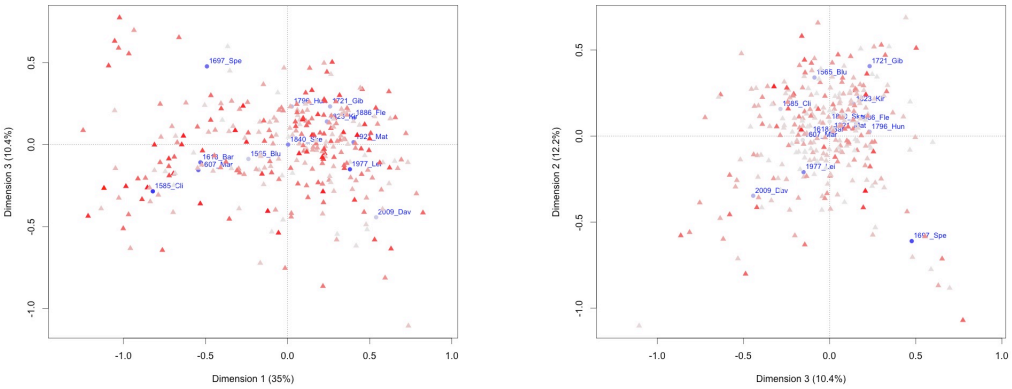


Figure C.1: Symmetric plot of trigrams and sources (no PUN, 50% of tokens)



(a) Symmetric plot, 100^{+} obs., Dims 1 & 3 (b) Symmetric plot, 100^{+} obs., Dims 2 & 3

Figure C.2: Symmetric plot of the third dimension (no PUN, 50% of tokens)

C.2 Correspondence analysis without outliers (text samples)

The plots provided in sections 5.3.2 and 5.3.3 suggested that the text sample by Speed (1697) is an outlier in the plot. Correspondence analysis offers a useful way to deal with such data by treating outliers as supplementary points. This procedure involves the computation of correspondences between rows and columns without taking the contributions of these supplementary variables into account (i.e., these rows or columns do not play a role in determining the total inertia of the solution; they have coordinates but no mass). After total inertia of the solution is established, the relative position of these points can be used to they display them in the solution that is based on the contributions of other remaining texts (cf. Baayen, 2008; Greenacre, 2007; Nenadic & Greenacre, 2007). For ease of visual inspection, we carry out the CA on the POS trigrams with 100^{+} observations. The results below are therefore best compared to those in §5.3.3.

Bendixen suggests to consider rows or columns as supplementary points when they have both high coordinate values and high contributions on a particular axis (Bendixen, 1996, p. 33). When we take a look at the solutions for the correspondence analyses in sections 5.3.2 and 5.3.3, the manuals Speed as well as Davies (2009) may be considered as potential outliers.[2] For example, the contribution of Speed on dimension 2 is particularly marked, i.e. contributions of 597 in the solution in 5.3.2 and 650 in 5.3.3, together with high (negative) coordinates. In comparison, the manuals that follow Speed in terms of contributions on dimension 2 are Davies on the first CA (ctr = 146) and Gibson (ctr = 98) in the second. In turn, the manual by Davies (2009) has an exceptionally large contribution on dimension 3: 386 for the first CA reported (the next manual, Clifford, has a contribution of 208 on this dimension), and a contribution of 463 on the second CA (followed by Gibson with a contribution of 113).

The first thing to be noticed in the scree plot in table C.2, particularly in comparison to the results obtained in the model that includes Speed as well as Davies (cf. table 5.4), is that the retention in the solution for the first three dimensions is similar: 74.0% here versus 74.1% in the model that includes the outlying text samples. However, if we compare the eigenvalues for each specific dimension, as well as how much they account for the total inertia, there are considerable differences: the total inertia here is 0.215512 (versus 0.295120 before). In addition, the three dimensions with significant eigenvalues account differently for their share of total inertia. The first dimension accounts for a larger share of the inertia here than before (53.7% versus 44.8%), but the contribution

[2]Note that there may also be a number of columns (POS trigrams) that may be considered as outliers in the current plots, especially when there is an indication of outliers in the rows. As our main focus is on detecting patterns in the relative clustering of text samples, however, we will not devote attention to POS trigram outliers here.

to the retention for the second and third dimensions is lower than in the corresponding model that includes the outliers: 13.1% versus 19.2% (dimension 2) and 7.2% versus 10.1% (dimension 3). As these outliers were associated with particularly dimension 2 (in the case of Speed) and dimension 3 (in the case of Davies) in section 5.3.3, this finding corroborates the notion that dimensions 2 and 3 could partly be described as the axes 'Speed versus not-Speed' and 'Davies versus not-Davies' with respect to the positioning of text samples in the biplots. In turn, in the current dimensional reduction the first dimension is comparatively more important, leading to a greater asymmetry between the three principal axes for this solution.

Table C.2: Scree plot for correspondence analysis (100^{+} observations, with supplementary rows)

Dimension	value	%	cum%	scree plot
1	0.115693	53.7	53.7	**********
2	0.028286	13.1	66.8	***
3	0.015500	7.2	74.0	**
4	0.013491	6.3	80.3	**
5	0.011573	5.4	85.6	*
6	0.008598	4.0	89.6	*
7	0.007711	3.6	93.2	*
8	0.006265	2.9	96.1	*
9	0.004628	2.1	98.3	*
10	0.003766	1.7	100.0	
Total inertia	0.215512	100.0		

The plots in figure 5.7 show a different picture than those obtained above. First of all, the texts by Speed and Davies appear inside the general cluster of text samples and POS trigrams. That their angle to the both axis is somewhat similar to before indicates that these texts 'depend' on both axis in more or less the same way, but it can be observed that their distance to the origin is reduced considerably. In addition, it is observed that these text samples do not have a corresponding data point or are sometimes indicated by empty circles, which is an effect of their having no mass.

The general pattern that emerges from an inspection of dimensions 1 and 2 is that, in addition to fewer extremes towards the negative region of dimension 2, the texts by Clifford, Baret and Markham appear below the horizontal axis in figure C.3. The new correspondence analysis thus causes the 16^{th}- and 17^{th}-century manuals to appear as less of a cluster and with larger distances between samples, particularly along dimension 2. The text samples by Hunter and Skeavington appear near the origin as before, but the text by Gibson now appears further from the origin on dimension 2 (its coordinate on dimension 1 remains largely the same). Not nearly as many differences are observed on the right hand side of the vertical axis, although the position of Davies is, of course, different. It is thus particularly in the early half of the corpus that the removal of

outliers is seen to have an effect in the plot of dimensions 1 and 2.

When we look at the cloud of points from the top, cf. figure C.4a, it is particularly noticeable that there is an early cluster of Clifford, Markham and Baret on the left, and a late periodic cluster on the right hand side of the plot, consisting of the manuals by Leighton-Hardman, Matheson, Fleming and the supplementary point of Davies. It is particularly the 19^{th}-century manuals that seem dispersed across the plot of dimensions 1 and 3, however. With Fleming appearing in this late cluster, another 19^{th}-century manual by Kirby occurs in conjunction with the 18^{th}-century manuals by Gibson and Hunter. The third 19^{th}-century manual, by Skeavington, appears somewhat in between this central cluster and a group of two texts by Speed (as supplementary point) and Blundeville in the bottom left quadrant of the plot. The vicinity of these three manuals is rather inexplicable in terms of their publication dates: Blundeville and Speed are nearly a century and a half apart, and the same in turn goes for Speed and Skeavington. It is however particularly the similarity in coordinates on dimension 3 which seems to join these texts.

If we then look at these three texts in a plot of dimensions 2 and 3 (cf. figure C.4b), it can be seen that it is particularly the second dimension that helps to differentiate the manuals by Blundeville, Skeavington and Speed. Although the relative position of texts on is difficult to account for, temporal considerations do not seem to play any role here, and it is potentially the text by Gibson (1721) that may account for part of this distortion given its high coordinate (and a contribution of 620) on dimension 2.

In sum, the current CA improves our visualisation in some respects but creates new problems in terms of potential outliers and inexplicable clustering of texts. Given the higher asymmetry between dimensions in the current solution, the analysis in section 5.3.3 seems as adequate a depiction of the data as the results presented here.

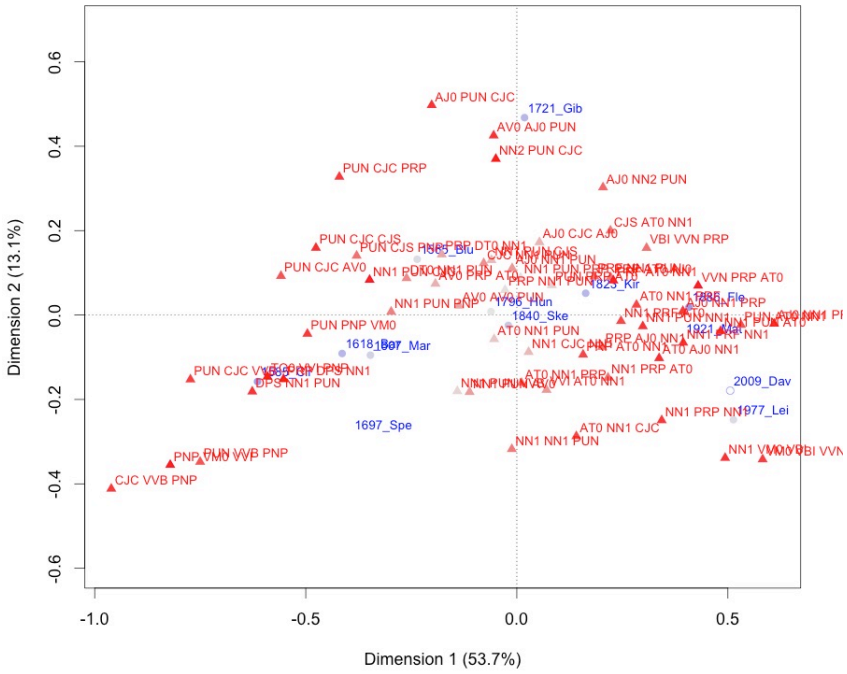
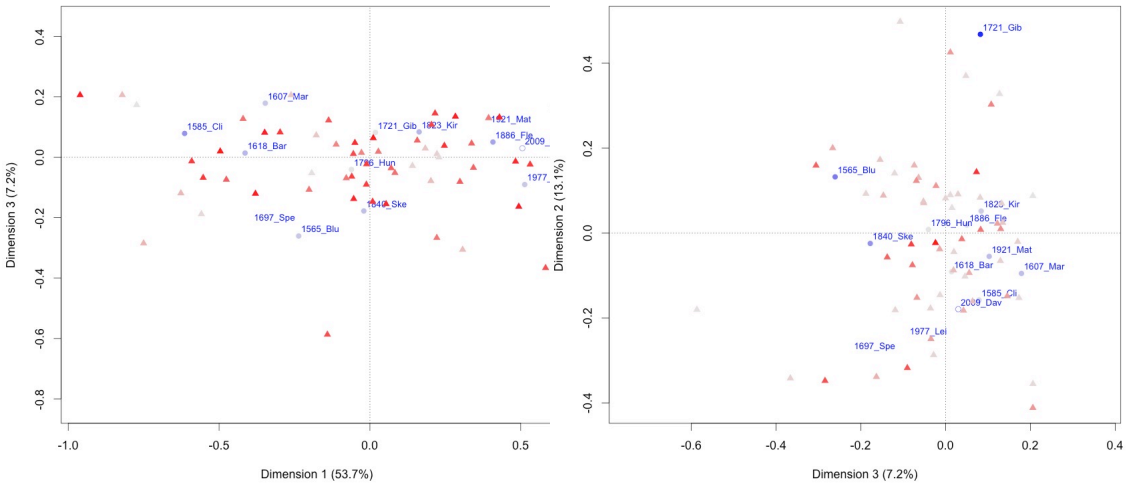


Figure C.3: Symmetric plot of trigrams and sources (100^+ observations, supplementary variables: Speed, Davies)



(a) Symmeric plot, 100^+ obs., Dims 1 & 3 (b) Symmetric plot, 100^+ obs., Dims 2 & 3

Figure C.4: Symmetric plot of the third dimension (100^+ obs., supp.: Speed, Davies)