Pushing Forward the Boundaries of Corpus-Oriented Historical Sociolinguistics

Terttu Nevalainen

1. Introduction

Tracing the history of English in sociolinguistic terms, we could follow Scott and Machan (1992), and begin at the beginning. They say:

> It is clear that the efficient cause of the beginning of what we call the English language was arguably a sociolinguistic phenomenon: the invasion of England in the 5th and 6th centuries by Germanic tribes who brought with them their own culture, customs and language. The society that these tribes initiated in England [...] necessarily placed demands on communication different from those experienced on the Continent. These exigencies concomitantly shaped the form and function of the dialects of Old English.
>
> (Scott and Machan 1992: 19)

Much of the research on linguistic variability in Anglo-Saxon England indeed comes under traditional Old English dialectology, which makes the best use of the fragmentary textual evidence available. It is only from the Late Middle English period onwards that data sources become ample and varied enough to relate linguistic variation to speaker variables such as regional background, social status, and gender.

One way to gain access to the social context of earlier English is by means of corpora, structured collections of texts which have been rendered into a computer-readable form for ease of data retrieval. In this paper I discuss data retrieved from a historical sociolinguistic corpus which allows users to delimit their searches according to the writer's social rank, domicile, sex, education, and other extralinguistic factors.[1] Features like these are normally searchable

[1] This article is a revised version of the presentation that I gave at the meeting of the English Research Association of Hiroshima in May 2008 in conjunction with my visit to Japan funded

in corpora and databases compiled for present-day sociolinguistic research.

My focus falls on the Renaissance period from the late 15th to the end of the 16th century. Three issues are addressed: (1) regional variation in ongoing processes of language change, (2) individual variation during these processes, and (3) comparison of these two, analyzing the degree to which people in close contact with each other participate in simultaneous processes of change. I hope to show how these two perspectives, macro and micro levels of analysis, can usefully complement each other (see, e.g., Nevalainen and Raumolin-Brunberg 2003: 16–25; Palander-Collin, Nevala and Nurmi 2009).

The processes to be discussed include the generalization across the literate population of the present-tense indicative plural form *are* of the verb *be* and of the 3rd-person present-tense singular indicative suffix *-s*, variation between the past-tense forms *was* and *were* with plural subjects, and the replacement of multiple negation (i.e. negative concord) by single negation accompanied by non-assertive forms. Negative concord and the use of *was* with plural subjects are common in non-standard and casual speech today, whereas the indicative plural *be* and third-person singular *-th* have practically disappeared even from traditional regional varieties of British English.

2. Historical Corpora

Several parameters can be used to distinguish between corpora. The basic distinctions that can be made include the time period covered (synchronic *vs.* diachronic corpora), size (big *vs.* small corpora) and number of genres included (multigenre *vs.* single-genre corpora), and degree of structuring and annotation (flat *vs.* annotated corpora).

A good example of a diachronic corpus is the Helsinki Corpus of English Texts (HC). It may be characterized as a diachronic multigenre corpus. It is diachronic in that it covers almost a thousand years of the history of English from the 8th to the 18th century. Its original version consists of 1.5 million words, and the Early Modern English section is represented by about half a million words. By many standards this counts as a small corpus. However, one million words seems to be the norm for heuristic corpora used for basic fact-

by the Japan Society for the Promotion of Science. I would like to thank Professor Akiyuki Jimura for inviting me to address the meeting, and all those present for an interesting discussion of a number of issues raised by this topic.

finding and pilot studies. One million is the size of the classic Brown University Corpus of American English, the texts in which all predate to 1961. The components of the International Corpus of English (ICE) consisting of regional varieties are also one million words each.

The HC consists of a number of genres and therefore qualifies as a multigenre corpus. Most genres included in the HC were composed for public distribution and consumption. In the Renaissance period such genres include biographies, handbooks, philosophical and educational treatises, and histories. Whenever possible, non-public writings such as private diaries and personal letters were also included. Language composed for oral delivery, as in sermons and plays, was similarly sampled, as were texts originally produced in the spoken medium, such as trial proceedings.

Corpora such as the HC which consists of many genres can provide information on *usage-based* differences in language variation and change. In sociolinguistically-oriented work, this information is supplemented by corpora that enable the study of *user-based* variation. Personal letters, in particular, supply material produced by individuals whose life-histories have been recorded for posterity. Material like this allows the correlation of language variation with speaker variables. One such resource is the Corpus of Early English Correspondence (CEEC). One of the original aims of the CEEC project was to provide real-time data on processes of morphosyntactic change in English between 1410 and 1680, which is the period covered by the 1998 version of the corpus. This diachronic corpus currently covers the period from c. 1400 to 1800, as the original version of the corpus has been extended to cover the 18th century. This complete version now runs up to some 5 million words.[2]

Consisting of personal letters, the CEEC is a single-genre corpus. Personal correspondence is the only genre to provide a whole range of authentic communication from the early 15th century to the present day. As discussed in Nevalainen and Raumolin-Brunberg (2003), the writers in the CEEC have been classified according to various sociolinguistic parameters, including domicile. Four geographical regions receive special attention: London, the Court, East Anglia, and the North. Writers living in counties north of the Chester-Humber

[2] Both the HC and the original version of the CEEC have been released in three different formats: as a text-only version, and as grammatically annotated tagged and parsed versions. See the online Corpus Resource Database (CoRD) at <http://www.helsinki.fi/varieng/CoRD/corpora>.

line, i.e. north of Lincolnshire, count as Northerners, and those resident in Norfolk and Suffolk as East Anglians. People who lived in London (mostly in the City of London, though Southwark is also represented) are entered as Londoners.

The Court refers to those writers, mostly resident at Westminster, who were courtiers or belonged to the royal household, or worked as high-ranking government officials at Court. Including the Court as a category of its own makes it possible, among other things, to examine the effect of overt prestige on language variation and change. It was, after all, the language of the 'better-brought-up sort' of London, and of the Royal Court in particular, that George Puttenham recommended as a model for aspiring poets in his *The Arte of English Poesie* (1589).

3. Studying Variation and Change in Renaissance English

Using a selection of materials like the HC, we can trace back patterns of variation and processes of change in English grammar over the centuries. A distinction can be made between language changes emanating from the official written end of the genre spectrum, such as statutes, and those that were first manifested in more informal, colloquial texts, such as comedy.

By contrast, corpora such as the CEEC can help us describe the language of individuals, and of groups of individuals classified according to their social rank, gender, domicile, and level of education, for example. If we are interested in processes of linguistic change, this information will enable us to answer various questions to do with the social embedding of these processes. It also makes it possible to trace back the sociolinguistic origins of ongoing changes and contributes to our understanding of the pathways of diffusion of language change throughout the language community in a given period of time.

In this section I study the macro- and micro-level uses of the CEEC by contrasting the language of one notable individual, Queen Elizabeth I, with that of her immediate Court entourage in the last two decades of the 16th century. To relate their idiolectal variation to contemporary regional trends at large, I compare their usage with general findings from the Court, the City of London, East Anglia, and the North of England.

In the CEEC, the holograph material by Elizabeth Tudor (1533-1603) spans almost fifty years from 1548 to 1596. The recipients of her 33 letters include Lord Protector Somerset, King Edward VI and King James VI of

Scotland; James in fact receives most of them in the 1580s and 90s. The courtiers and public figures that I use as my comparative material include Robert Dudley, the Earl of Leicester (1532–1588), the Queen's suitor and favourite; William Cecil (1520–1598), her advisor and chief minister; Francis Walsingham (1532–1590), her principal secretary and spymaster; as well as Robert Cecil (1563–1612), who followed his father in the Queen's service.

The four linguistic features I focus on were all undergoing change in the 16th century. They consist of the following alternative expressions:

- be vs. are in present indicative plural of be:
  they be good vs. *they are good*

- *-th* vs. -s in third-person singular present indicative:
  it moveth vs. *it moves*

- *was* vs. *were* with plural subjects:
  we was there vs. *we were there*

- multiple negation vs. single negation followed by non-assertive forms:
  you never go nowhere vs. *you never go anywhere*

Some illustrations of Elizabeth's use of these forms are shown in bold type in (1) and (2), where her usage proves quite variable: she uses both *-s* and *-th* and *are* and *be*.

(1) My deare brother, As ther is naught that **bredes** more for-thinking repentance and agrived thoughtes than good turnes to harme the giuers ayde, so **hathe no** bonde **euer** tied more honorable mynds, than the shewes of any acquital by grateful acknowelegement in plain actions; for wordes **be** leues and dides the fruites. (A 1591 FO ELIZABETH1 65)

(2) Right deare brother, the strangenes of harde accidens that **ar** arrived here, of unloked for, or unsuspected, attemps in Skotland, euen by some suche as lately issued out of our lande, **constraineth** me, as wel for the care we have of your person as of the discharge of our owne honor and consciense, (A

1585 FO ELIZABETH1 23)

Elizabeth's use of single negation followed by a non-assertive form is also illustrated in (1), and an example of her use of multiple negation in (3), which is an instance of a co-ordinate structure.

(3) and first, for my promis made of reciproke usage in all amicable maner, I trust I **nether** haue, **nor neuer** shall, make fraction of in the lest scruple; (A 1586 FO ELIZABETH1 33)

The Queen's variation between the past-tense forms was and were with plural subjects is shown in (4) and (5).

(4) **My wordes wer** thes: "I heare say the offending lordz hopes by ther frindz to skape ther paine; (A 1593 FO ELIZABETH1 91)

(5) And for the lettar, **some wordes and fourme was** suche as fitted not our two frindeships, as Randol also can shewe you, but I haue sent you a lettar that I am sure containes all you desired in spetiall wordes. I trust it shal content you; (A 1586 FO ELIZABETH1 34)

The processes of change are discussed below, and the numerical data are presented in the Appendix.[3]

3.1. *Be* vs. *Are*

Let us begin by considering forms that originated in Northern English dialects, first the replacement of the Southern indicative plural *be* by the Northern *are*. The process was nearing completion at Court towards the end of the 16th century, as shown in Figure 1. As can be seen, *are* had not been a typical Court form in the earlier part of the 16th century but the usage there had caught up with the City of London by the mid-century.

[3] More information on these and other regional processes can be found in Nevalainen and Raumolin-Brunberg (2003, Ch. 8). For a discussion of the aggregate figures on *be/are*, see Nevalainen (2000), for *was/were* Nevalainen (2006a) and negative concord Nevalainen (2006b).

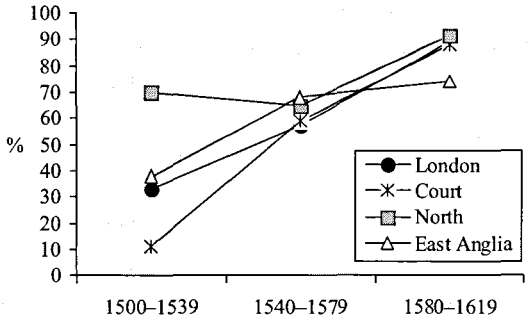

Figure 1. The replacement of *be* by *are* in the indicative plural. Regional distribution of *are*. CEEC 1998 and Supplement.

Figure 2 compares the Queen with some of her well-known courtiers. It suggests that the Queen is in fact the most conservative among them: she uses the incoming form only some 50% of the time in her letters to James VI. Robert Dudley, the Earl of Leicester, comes closest to her usage; born in 1532, he was also very close to her age. Robert Cecil employs more the incoming form, but, born in 1563, he is more than a generation younger than the Queen. Age cannot, however, explain the differences because William Cecil and Francis Walsingham, who use the incoming form most of the time, were either older than the Queen (Cecil) or close to her age (Walsingham).

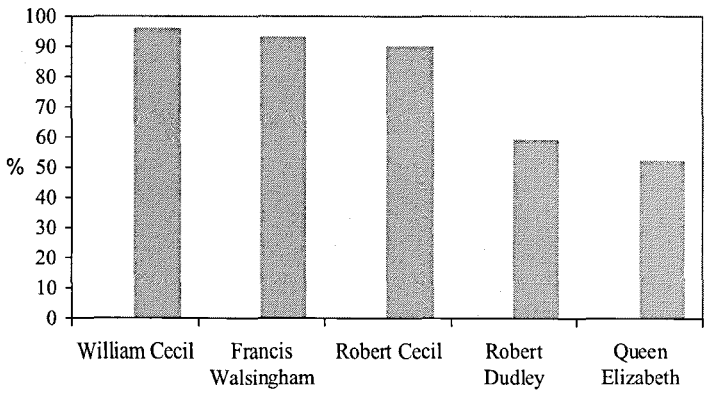

Figure 2. The use of *are* (%), 1580–99. CEEC 1998.

One possible reason for the range of variation in the four men's letters might be scribal influence. While Queen Elizabeth's letters are all holograph, written by her in their entirety, the same is true of the great majority, but not all, of the four courtiers' correspondence. However, no obvious pattern emerges if we compare their holographs with secretarial letters: both show similar variable usage. The differences between the two groups, the Queen and the Earl of Essex as opposed to the two Cecils and Walshingham, might therefore reflect a change spreading from below in social terms. After zero incidence in King Henry VIII's time, the Court usage of *are* had become variable during the 16th century. With respect to this feature the Court had not been a trendsetter, but had rather followed the practice spreading from the outside, the nobility and royalty being the slowest to adopt the incoming form.

3.2. Verbal *-th* vs. *-s*

Moving on to another feature of Northern origins, the third-person *-s*, the picture looks rather different although there are also similarities. Figure 3 shows that, as in the case of *are*, it was Londoners that promoted the form more readily than those resident at Court. In the late Elizabethan period, the London writers in the CEEC include, for example, Philip Henslow and his theatrical circle writing around 1600.

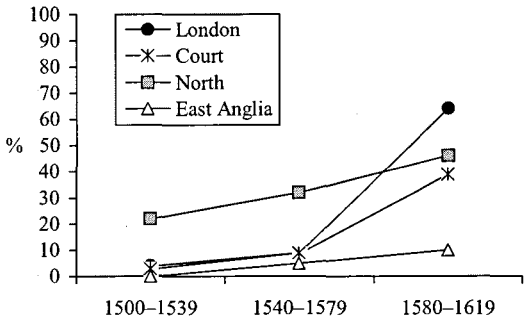

Figure 3. The replacement of *-th* by *-s* in verbs other than *have* and *do*. Regional distribution of *-s*. CEEC 1998 and Supplement.

Figure 3 shows that the diffusion of *-s* was not as far advanced as that of *are* at the turn of the seventeenth century. In fact, with the average rate of 39%, it had barely reached mid-range at Court. However, as indicated by Figure 4, Queen Elizabeth uses the incoming feature half of the time, which is clearly more than the Court average towards the end of the 16th century. Robert Dudley and Robert Cecil both also use it more frequently than was customary at Court. By contrast, Francis Walsingham and the older Cecil hardly ever employ the incoming form.

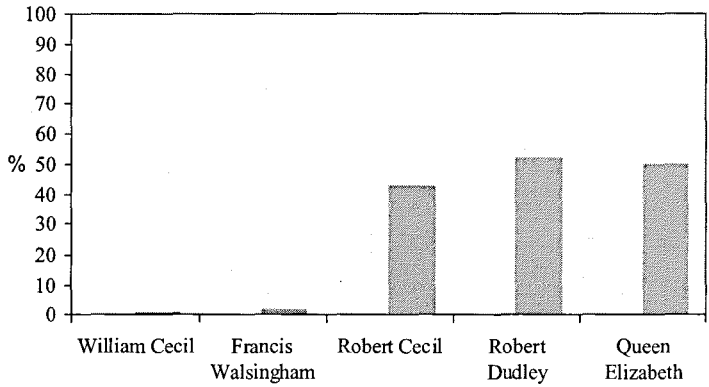

Figure 4. The use of *-s* (%), 1580–99, excluding *have* and *do*.

With respect to this change, only 30% of the writers in the CEEC had a variable grammar in the third-person singular in this period; that is, most people used either the recessive or the incoming form and only 30% were like the Queen, Robert Dudley and Robert Cecil, who used both of them concurrently. This means that this change may have been spreading from above the level of social awareness.

Seeing the non-use of the incoming form both in London and at Court in the mid-16th century, a person's age was also likely to play a role in this change: William Cecil was older than the other male writers studied, and hardly participated in this change at all. Age cannot fully explain why Walshingham, but not Elizabeth and Dudley, ignored the incoming form. If the change spread from above the level of social awareness, the reason may lie in the relatively formal nature of his secretarial correspondence. A similar social

evaluation clearly did not, however, affect the use of the concurrently incoming Northern form *are*.

3.3. *Was* vs. *Were* with Plural Subjects

My next case illustrates another feature that was variable but failed to spread in the South in the written language of personal correspondence. Like *are* and verbal *-s*, the use of *was* instead of *were* with plural subjects was favoured in the North between the late 15th and mid-16th centuries but, unlike them, it did not pick up either in East Anglia or in the capital region. These developments are presented in Figure 5.

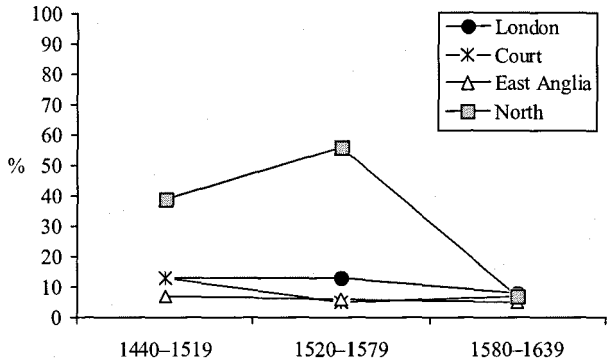

Figure 5. The use of *was* vs. *were* with plural subjects. Regional distribution of *was*. CEEC 1998 and Supplement.

The general Court usage and that of the Queen and her courtiers are quite uniform, as shown by Figure 6. Although this feature is not as common in the language as the previous two, the differences between the North and the rest of the country in the first two periods are statistically significant. No such difference can be detected among the individuals in Figure 6.

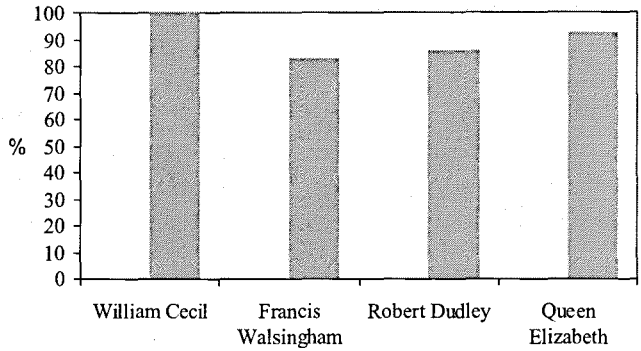

Figure 6. The use of *were* (%), 1580–99 (R. Cecil excluded because N < 10).

3.4. Multiple Negation

As far as the disappearance of multiple negation and the rise of single negation with non-assertive indefinites are concerned, Elizabeth's usage is again well in keeping with that of her Court. Figure 7 indicates that this process is already attested at Court in King Henry's time and even earlier. In her letters Elizabeth uses the incoming non-assertive forms such as *any* with negation in almost 90% of the cases, as suggested in Figure 8.

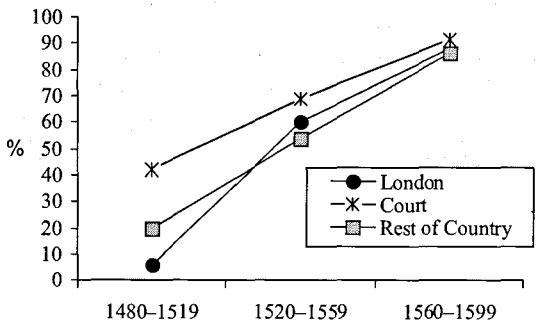

Figure 7. The replacement of multiple by single negation. Regional distribution of single negation. CEEC 1998 and Supplement; male writers.

Previous work shows that the demise of multiple negation is connected with the individual's level of education and employment (Nevalainen 2006b). It is therefore not surprising that the differences in Figure 8 are not statistically significant. The process is also linguistically conditioned: if co-ordinate cases such as *neither ... nor* in example (3), where multiple negation lingered on much later than in other contexts, were excluded from the data, Queen Elizabeth's letters would show no instances of negative concord.

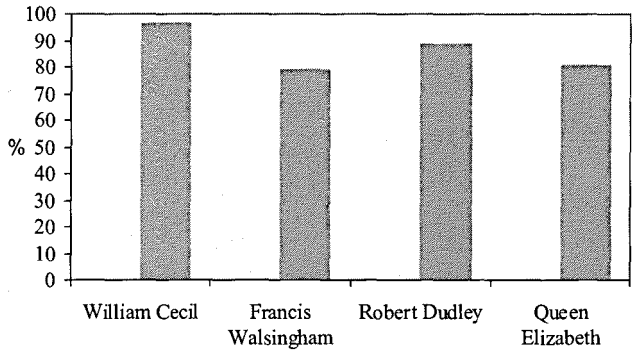

Figure 8. The use of single negation (%), 1580–99 (R. Cecil excluded because N < 10).

4. Conclusion

I would like to conclude by suggesting that the examination of Tudor Court correspondence has revealed to us variable usage, variable both synchronically and diachronically. At the individual level, statistically significant variation was detected with respect to the two features with Northern origin that divided Queen Elizabeth and her courtiers, viz. the indicative *are* and, especially, verbal *-s*. The Queen herself emerges as a linguistic innovator at Court in the case of *-s*, but is, at the same time, found lagging somewhat behind her chief ministers in her use of *are*.

There is no denying that the Royal Court had a role to play in the nation-wide diffusion of the changes that took place in Tudor English. In the 16th century, the Court formed a centre of linguistic focusing, and was instrumental in transmitting Southern influences to the rest of the country. As these linguistic practices were not prescribed, however, they could be challenged

and even overridden. We have seen this happen with the plural *be* and the third-person *-th*, which gave way to their originally Northern counterparts *are* and *-s*, first in the City and, after some resistance, also at Court.

Apart from the variation we have detected, the Court could also present a unified front both in aggregate and at the level of individuals. Innovations were promoted by those professionally involved in running the government and its various functions on a day-to-day basis. Processes like the disappearance of multiple negation belong to this group, which originally appear to have had closer links with the written language than with the spoken idiom. Not unexpectedly, uniformity also characterized those variable features that failed to spread to the South in general, such as the use of *was* with plural subjects, which may have been related to the Northern Subject Rule at this point in time.

University of Helsinki

Appendix. Relative frequencies (%) and variable totals (N) of the items discussed.

Writer	Frequency of *-s* (N)*	Frequency of *are* (N)*	Frequency of single neg. (N)	Frequency of *were* (N)
William Cecil	1% (164)	96% (52)	98% (46)	100% (28)
Francis Walshingham	2% (130)	93% (40)	79% (19)	83% (12)
Robert Cecil	43% (47)	90% (19)	100% (9)	100% (8)
Robert Dudley	52% (122)	59% (92)	90% (62)	86% (42)
Queen Elizabeth	50% (82)	52% (21)	81% (21)	93% (15)

(*using the Chi-square test, differences of the frequency distributions are statistically significant at the .001 level)

References

CEEC = The Corpus of Early English Correspondence (1998). Compiled by Terttu Nevalainen (Project leader), Helena Raumolin-Brunberg (Senior scholar); Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin at the Department of English, University of Helsinki.

HC = The Helsinki Corpus of English Texts (1991). Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka,

Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English), at the Department of English, University of Helsinki.

Nevalainen, Terttu (2000) "Processes of Supralocalisation and the Rise of Standard English in the Early Modern Period." In: Ricardo Bermúdez-Otero, David Denison, Richard M. Hogg and C. B. McCully, eds., *Generative Theory and Corpus Studies: A Dialogue from 10 ICEHL*, 329–71. Berlin and New York: Mouton de Gruyter.

Nevalainen, Terttu (2006a) "Vernacular Universals? The Case of Plural Was in Early Modern English." In: Terttu Nevalainen, Juhani Klemola and Mikko Laitinen, eds., *Types of Variation: Diachronic, Dialectal and Typological Interfaces* (Studies in Language Companion Series 76), 351–69. Amsterdam and Philadelphia: John Benjamins.

Nevalainen, Terttu (2006b) "Negative Concord as an English 'Vernacular Universal': Social History and Linguistic Typology," *Journal of English Linguistics* 34/3, 257–78.

Nevalainen, Terttu and Helena Raumolin-Brunberg (2003) *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. London: Pearson Education.

Palander-Collin, Minna, Minna Nevala and Arja Nurmi (2009) "The Language of Daily Life in the History of English: Studying How Macro Meets Micro." In: Arja Nurmi, Minna Nevala and Minna Palander-Collin, eds., *The Language of Daily Life in England (1400–1800)*, 1–23. Amsterdam and Philadelphia: John Benjamins.

Puttenham, George (1589) *The Arte of English Poesie*. London: Richard Field.

Scott, Charles T. and Tim William Machan (1992) "Introduction: Sociolinguistics, Language Change, and the History of English." In: Tim William Machan and Charles T. Scott, eds., *English in Its Social Contexts: Essays in Historical Sociolinguistics*, 3–27. New York: Oxford University Press.