Language Variation and Change, **20** (2008), 155–185. Printed in the U.S.A.
© 2008 Cambridge University Press 0954-3945/08 $16.00
doi: 10.1017/S0954394508000045

Sociolinguistic patterns in grammaticalization: *He, they*, and *those* in human indefinite reference

MIKKO LAITINEN
University of Helsinki

ABSTRACT

Sociolinguistic patterns in language change are largely based on generalizations from linguistic variables consisting of lexemes or morphemes. This article takes a diachronic, corpus-based approach to the diffusion of a change in a more extensive morphosemantic function consisting of several linguistic subsystems. It focuses on the pronoun variants *he, they*, and *those* used for human indefiniteness in two contexts: (a) epicene anaphoric uses with indefinite pronouns and (b) cataphoric personal references. The quantitative corpus analyses show that the pronoun selection in Early and Late Modern English developed a greater tendency to use one pronoun type over the other in both contexts. The main data come from the *Corpus of Early English Correspondence* and its *Extension*. Statistical analyses compare the observed correlations of the pronouns with a set of social, external variables and language-internal factors. This article concludes that it is possible to establish sociolinguistic patterns in larger shifts if we account for the closely related internal developments in the language.

In his discussion of sociolinguistic variation, Labov (2001) suggests that it would be surprising if abstract levels of grammar, such as large-scale typological changes in a language, exhibited sociolinguistic patterns found in socially evaluated variables. At the same time, he suggests that stable sociolinguistic variables, which may persist over several generations of speakers, are typically "located at an intermediate level of structure . . . [like] the use of allophones or allomorphs in a particular structural context" (2001:83).

This article concentrates on these two proposals and examines them using diachronic corpus data. I have two underlying themes that can be outlined as follows. (a) Tackling Labov's notion of sociolinguistic patterns in large-scale changes, I concentrate on one abstract morphosemantic function in English, human indefinite reference in the pronouns *he*, *they*, and *those*. I suggest that the use of the pronominal forms with which this function was coded through time constitutes a change in the typological tendencies in English, and I examine the extent to which one can establish sociolinguistic patterns in its diachronic development. My interest is social diffusion of the variants, which are elaborated

The research reported here was supported by the Academy of Finland Centre of Excellence funding for the Research Unit for Variation, Contacts and Change in English at the Department of English, University of Helsinki.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

below. (b) As regards stable variation, my aim is to show how balanced corpus data allow us to pinpoint stable sociolinguistic variables that are words or even parts of larger linguistic structures, not "intermediate level" elements, as suggested by Labov (2001:83).

The linguistic variable whose diffusion is modeled deals with a well-known linguistic puzzle in English not previously studied from this perspective, the selection of pronoun number, singular or plural, in an epicene (human indefinite) reference in three centuries of Early and Late Modern English (1500–1800) (Baron, 1986; Bodine, 1990). Balanced corpus data show that a systematic analysis of the epicene function must account for pronoun number realization in two contexts, both of them with the pronoun variants sharing the same property, namely singularity *versus* plurality. The contexts with their dependent variants are shown below by two contemporaries from the 16th century.[1] Extracts (1) and (2) illustrate anaphoric references to human indefinite pronouns (hereafter *anaphora*) with the variants *he* and *they*.[2] Extracts (3)–(5) show the variants *he, they*, and *those* in cataphoric personal reference (*cataphora*), functioning as the antecedents of the following relative clauses.

(1) and as well my selfe as eny other to whom his Highnes thowght the thing to seme other wise, he vsed in his other busynes, abyding (of his abundant goodnes) neuer the lesse graciouse lord vnto **eny man**, nor neuer was willing to put **eny man** in ruffle or trowble of **his** conscience. (Thomas More, 1534; MORE, 496)

(2) and in the meane while pray for your Grace, it may lyke your Highnes of your accustumed benignite somewhat to tendre my pore honestie and never suffre by the meane of such a bill put forth agaynst me, **eny man** to take occasion here after agaynst the treuth to slawndre me; which thyng shold yit by the perell of **theire** awne soulys do **theym selfe** more hurt than me, which shall, I trust, settle myn harte, (Thomas More, 1534; MORE, 491)

(3) Sumwhat I shal knowe more by the Quenes facion, which I shal advertise incontinently, and spare noo more the Kinges Majesties purse in postes nowe thenne other doo in ther feate, being everyman disposed to spende with the uttermost, besides **them that** waste—but I staye here for feare of entryng the matier of victualles, etc. (Stephen Gardiner, 1545; GARDINER, 158)

(4) I have lyved soo honnourable a lyfe by the Kinges Majesties goodnesse, duryng whose life, which I trust shal be lenger then myne, I feare not thiese fonde malicious folyes. But whenne **those that** nowe be yonge shal, with the fraylete of yought, wyne a contempte of religion and conceyve an other opinion of God thenne is in dede trewe, what is like to ensue therof? (Stephen Gardiner, 1545; GARDINER, 161)

(5) I entreate this offense oonly worldlye, bycause the capacitie of thoffenders seemeth to stretche noo further. And **he that** regardith not **his** obedience to **his** prince regardith not moch **his** obedience to God and His truth, which **he** hath offended in the other. (Stephen Gardiner, 1545; GARDINER, 132)

Covering the two contexts for the same function ensures that one no longer deals with individual morphemes or words, but with abstract morphosemantic phenomena. My aim is to show that anaphora and cataphora indeed constitute

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

semantic functions in which pronoun number type can vary considerably.[3] Diachronic corpus data suggest that the pronoun frequencies underwent a restructuring in which one variant type came to dominate over the other. While establishing sociolinguistic patterns in this development and comparing the sociolinguistic processes in these two contexts, I aim to illustrate that there are linguistic elements involved that remained constant for several centuries.

I will first elaborate on the anaphoric and cataphoric contexts and the dependent pronoun variants in the section entitled "Anaphoric and cataphoric contexts." The corpora used and the overall data are presented in the section entitled "The diachronic developments in the data." The empirical part then focuses on cross-tabulating the pronoun variants with a set of external social factors, such as gender and region. Data frequencies permitting, I will make use of a statistical multivariate analysis using the variable rule program (VARBRUL) (Rand & Sankoff, 1990). These VARBRUL analyses allow us to assess the effect of various sociolinguistic variables, compare them with language-internal factors, and model the social diffusion of variants in these contexts.

ANAPHORIC AND CATAPHORIC CONTEXTS

The dependent linguistic variant concerns the selection of pronoun number in a communicative function that expresses an unknown, indefinite reference. As (1)–(5) show, the empirical part of this article involves a heterogeneous set of contexts. To conceptualize the research question, Figure 1 illustrates the two deictic categories, anaphora and cataphora, in which the pronouns are used.[4] The anaphoric cases (1) and (2) concern the selection of the personal pronouns *he* and *they* with various antecedents, most typically grammatically singular compound indefinite pronouns, the underlined forms in Figure 1. The antecedents must be unknown because the linguistic interest lies in whether writers use a singular pronoun or a plural one in cases with no overt number distinction. The dependent variants in cataphora, *he*/*they*/*those* in (3)–(5), function as the antecedents of the following relative clauses (Huddleston & Pullum, 2002:1036–1037). They are similar to the indefinite pronoun antecedents in anaphora, as the entire proposition must be indefinite, and the pronouns do not have a real-world referent.[5] In both reference types, the interest lies in the writers' selection of pronouns, either *he* or they in anaphora, or *he*/*they*/*those* in cataphora.

Because the dependent pronoun variants in both anaphora and cataphora in Figure 1 can be used for similar unknown, indefinite uses, my argument is that these linguistic contexts, with their pronoun variants, are in fact two sides of the same semantic construction. This argument is similar to the way constructionists see syntactic constructions (Goldberg, 2003). Indefinite expressions are used for a particular communicative function, which may not be strictly predictable from their overt linguistic components but, when interpreted in context, create patterns of form and function used for a propositional content.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

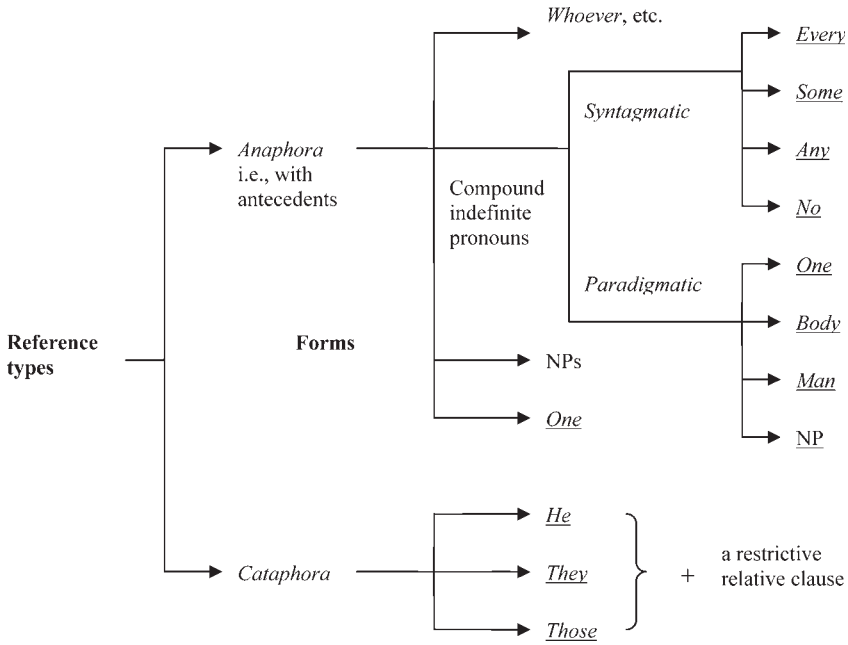


FIGURE 1. Reference types for human indefinite function.

My empirical approach to the variant distribution in these two contexts shows that they underwent a grammaticalization process that affected the frequencies of occurrence of these pronouns (Laitinen, 2006; forthcoming). Their paths of development are markedly similar, but not identical, allowing us to make various generalizations. First, one variant type, the plural, has increased its frequency in anaphora and emerged as the main variant in present-day English (Baranowski, 2002). There is, however, limited corpus-based research to describe this process in language history (a notable exception appears in Dekeyser, 1975). Second, a major restructuring took place in cataphora in Early Modern English, when the demonstrative *those* replaced the personal pronoun variants, *he* and *they*. Previous scholarship is again limited: Rydén (1966) provides a detailed summary of the synchrony in the early 16th century.[6] Jespersen comes up with an observation from Shakespeare, showing that the structure was in transition in the Elizabethan period.

(6) God giue **them** wisedome **that** haue it: & **those that** are fooles, let them vse their talents. (*Twelfth Night* I.v.14, as quoted in Jespersen, 1949:III.5.15)

Third, because language change is a social process, "taking place in the activities of human users of language in social and situational context" (J. Milroy, 2005:2), we face the task of correlating our linguistic variants with sociolinguistic variables to acquire a more precise picture of language change. The main task, therefore,

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

includes modeling how these changes spread, concentrating on language external social factors.[7]

THE DIACHRONIC DEVELOPMENTS IN THE DATA

The fact that pronoun selection in anaphora, in particular, is a normative question in present-day English has occasionally blurred diachronic approaches to these pronouns, and scholars have not been able to detect the semantic connections between anaphora and cataphora. Tables 1 and 2 summarize the general findings to provide a bird's-eye view of the diachronic evolution in these pronouns in this sense. They show the overall trend observed in the *Corpus of Early English Correspondence* and its *Extension* (the CEEC) (Raumolin-Brunberg & Nevalainen, 2007).[8] The results for the two contexts are presented separately to show the divergence in their diachronic developments.

Table 1 shows that the development of the overtly plural *they* in anaphora has been slow. Contrary to many studies of social diffusion, the change was not completed during the period examined here. According to the concept of the S-shaped curve, the development covers the midrange (between 36 and 65%) and closes near completion (between 66 and 85%) (Labov, 1994:65–83; Nevalainen & Raumolin-Brunberg, 2003:54–55). A chi-square test confirms that the development toward the plural was statistically highly significant ($p < .005$). To put it briefly, the early 16th century pronoun realizations of anaphoric pronouns to indefinite pronouns include significantly more singular third-person *he* uses than those of the late 18th century. The change in anaphora was not completed, as both variants continue to be in use in present-day English.[9] The slow shift toward the plural has continued after the 18th century (Laitinen, forthcoming).

Table 2 shows the chronological development leading to the emergence of the demonstrative over the personal pronoun variants. The personal pronouns *he* and *they* are the major variants in cataphora in the 16th century, and *those* gradually replaces them in the course of two centuries. The chi-square values indicate that the change is highly significant ($p < .001$).

TABLE 1. *The anaphoric variants* he *and* they *in the CEEC*

	He	%	*They*	%	Total
1500–1549	77	51	74	49	151
1550–1599	82	40	124	60	206
1600–1649	104	46	124	54	228
1650–1699	68	30	160	70	228
1700–1749	42	40	63	60	105
1750–1800	44	33	88	67	132
					1050

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

TABLE 2. *The cataphoric variants* he, they, *and* those *in the CEEC*

	He	%	*They*	%	*Those*	%	Total
1500–1549	31	33	39	42	23	25	93
1550–1599	41	32	55	43	32	25	128
1600–1649	50	26	66	34	78	40	194
1650–1699	37	20	49	27	96	53	182
1700–1749	3	5	10	18	42	76	55
1750–1800	19	14	12	9	101	77	132
							784

Tables 1 and 2 illustrate that the paths of development of our dependent pronoun variants are chronologically different. The change in cataphora nears completion towards the end of the time frame, whereas the development in anaphora is considerably slower. The pronouns in these two contexts, however, share a common denominator. The emerging variants are perceptually more normal when used in the epicene sense, sharing a larger behavioral potential and involving less structural conditioning, that is, fewer forms for one function (Croft, 2003:95–99). Comrie (1986) claims that our tendency to use some forms more frequently than others is ultimately psychological, so that linguistically less formal constructions correspond more closely to our conceptualizations of real-world phenomena (see also Hawkins, 2004:51). To put it simply, using *he* in an anaphoric reference to an indefinite entity whose sex (and linguistic number) is inconsequential or indeterminate is less likely than that of a nongendered variant. Similarly, using *those* in cataphora is more expected, because it is the nonanaphoric (and nongendered) variant.

The corpora used in the next sections provide us with balanced data sets that are stylistically and socially stratified, so that we are able to contextualize the processes of language change and focus on the spread of the emerging variants. Indeed, language-external embedding in grammaticalization is largely unexplored. Nevalainen (2004) notes that "little is known about the social side of real-time process of language change in general and grammaticalisation processes in particular that only some very broad empirical generalisations may be advanced." In what follows, I show how the emerging variants, *they* in anaphora and *those* in cataphora, spread socially in the correspondence material, concentrating particularly on the role of extralinguistic factors, and the pronoun results are organized according to these independent variables. I then make use of multivariate analysis, a method common in both synchronic and diachronic sociolinguistics (Nevalainen, 2006; Tagliamonte, 1998).

GENDER

Exploring pronoun selection and the writers' gender is motivating, because one of the dependent variants in both anaphora and cataphora is the gender-specific *he*.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

One could therefore expect to find differences between men and women in adopting the pronouns. At a more detailed level, the pronouns in anaphora offer an interesting field of research. As seen in Figure 1, the paradigmatic level in the antecedents concerns three types of nominal elements; of these *-man* declined in Early Modern English, and the other two, *-body* and *-one*, emerged and continue to be used today. Raumolin-Brunberg and Kahlas-Tarkka (1997) have shown that this indefinite pronoun development was brought about by grammaticalization.

The Early Modern correspondence data for men and women is skewed, as women's letters comprise about 20% of CEEC (Nevalainen & Raumolin-Brunberg, 2003:45–49). Similarly, the number of relevant occurrences of the pronoun variables reflects the social representativeness of the data. The total number of pronouns examined in anaphora is 1050, of which 186 (18%) are from women. In cataphora, the women's total is 162, 21% of the total of 784.

The rationale of this section lies in sociolinguistic studies on gender as an independent variable. It has been established that gender differentiation in language use must be combined with our knowledge of the social awareness of linguistic change. Labov's Gender Paradox predicts that in changes from above, women typically adopt prestige forms more quickly than men (diffusion). In changes from below, women tend to use higher proportions of innovative forms (actuation) (2001:293). Recent historical sociolinguistic studies have established that in the late medieval and early modern periods, the role of women was more complex. Nevalainen and Raumolin-Brunberg (2003:131) argue that "late medieval and early modern Englishwomen did not promote language changes that emanated from . . . out-side their own spheres of 'being.'" Another aspect of the topic is prestige, the role of grammars and language planning; present-day synchronic research has found that females not only favor prestige forms, but in fact create them (Milroy & Milroy, 1993:65; Nevalainen, 2000a).

Gender in anaphora

In the early modern period, the situation with the pronouns in anaphora was complex. Neither of the variants can be said to have correctness values, because the first normative statements of the singular *versus* plural anaphoric pronouns date from the middle to the end of the 18th century (Bodine, 1990; Sundby et al., 1991:106–113). The corpus data suggest that some writers were at least partly conscious of this construction, as may be the case in (7) and (8), showing an overtly singular antecedent, a generic *one* in this case, and a plural pronoun immediately following it.

(7) When **one** does not know what to say **they** talk of the weather, the sudden change will give me an opportunity to make my letter a little longer (Peter Wentworth, 1708; WENTWORTH, 68)

(8) By the example of Sampson that **one** might and (if **they** be not too foolish to live in this world) must speake riddles to **theyr** frends and try the truth of offred love (Arabella Stuart, 1603; STUART, 130)

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

Figure 2 shows the gender differences in anaphora, indicating that women have used the plural variant, *they*, more than consistently throughout the three centuries of Early and Late Modern English, so that they have in fact heavily disfavored *he*. The masculine *he* has never been such a viable option in indefinite anaphora for female writers as it has been for men.

The variable differences between men and women are statistically highly significant in all but the fourth 50-year period (1650–1699) ($p < .01$). The results indicate that gender was a factor in the pronoun selection in anaphora. Moreover, the patterns of gender variation of this variable remained stable for three centuries, and the epicene uses of *they* have been socially stratified along the male and female distinction. In this respect, this is a clear case of a stable variable in sociolinguistics (Chambers, 2003:121–126; Raumolin-Brunberg, 2002). An essential difference from Labov's definition of stable variation is that the gender differentiation in the pronoun variants in anaphora deals with the word level of a language. In fact, the uses of epicene *he* and *they* must essentially be understood as parts of a larger structure in which the third-person pronouns are anaphorically linked to their antecedents. Huddleston and Pullum's (2002:1453–1461) definition of anaphoric relations suggests that anaphors are interpreted and determined via the interpretation of their antecedents. My suggestion is that the overall shift toward the plural provides evidence of a larger evolution in English pronouns in this sense, which is linked with other closely related changes (i.e., the indefinite pronouns) and is therefore evinced only in larger contexts.

The patterns of pronoun variation accounting for the various antecedents with which the variants are used are as important as the overall data (cf. Figure 1). Nevalainen and Raumolin-Brunberg (2003) have shown that men tended to retain the disappearing *-man* forms in correspondence later than women. I set out to test whether differences exist between third-person pronouns used to refer to the *-man* forms, such as (9) and (10):

(9) In the loft at the Wattergat corner, is 43 flecis of ende wull and xvj flecis of clift wull. If ***anny man*** lacke anny geleynges, I praie you selle ***them***; ye may let the packer shot owt the sourtes when niead is. (John Johnson, 1545; JOHNSON, 541)

(10) wch (pardon the Preacher) is this reflection; "That ***every Man*** must be finaly happy, if ***he*** has not by ***his*** folly and ill conduct forfeited all title to be so; or otherwise <u>no one</u> can be finaly miserable, whilst it is consistent with infinite goodness to make <u>him</u> otherwise." (Spencer Cowper, 1745; COWPER, 53)[10]

Figure 3 shows the singular and plural pronoun variants by men, illustrating that *he* was used more frequently with the *-man* indefinites than the plural *they*. It also shows that the share of *-man* compound indefinites decreases toward the end of the time period explored here; this decrease does not affect the proportional distribution of the pronoun variants, because the *-man* compounds tend to be used more with the singular masculine anaphoric pronouns than the plural one. Women's data are excluded simply because there are only sporadic occurrences

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

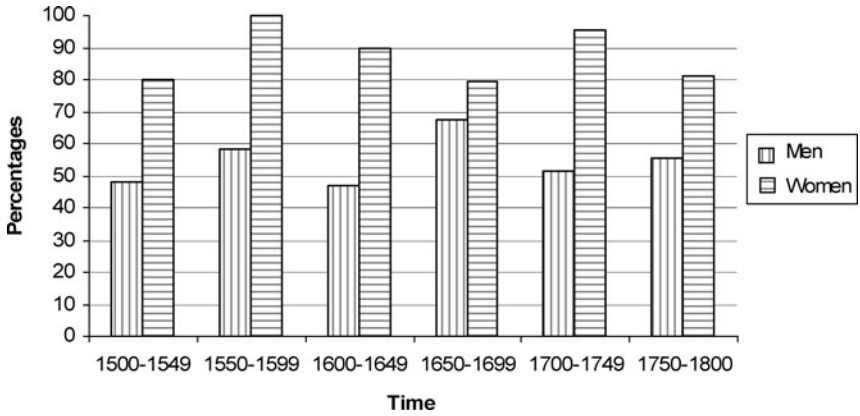


FIGURE 2. Gender variation in the uses of *they* in anaphora.

of *-man* compounds followed by anaphoric personal pronouns in the women's subcorpus. All are borderline cases in terms of their indefiniteness, as illustrated in (11)–(13).[11]

(11) you haue set no loer prise then 21 but hee doubts ***no man*** will giue so for the mony ***hee*** hath receued all as hee saith (Susanne Collett, 1626; FERRAR, 243)

(12) when I complained of your coming there contrary to your solemn Promise—that your Visits were not to *me* but to my *daughter*; would ***any Man*** after so saying expect me to admit ***him*** more? (Hester Lynch Piozzi, 1792; PIOZZI II, 53)

(13) You cannot expect, so thinking, so assured, that I should maintain a Correspondence with ***any Man***, when ***he*** neglects ***his*** part of it, tho′ that Man should be yourself. (Eliza Draper, 1769; DRAPER, 92)

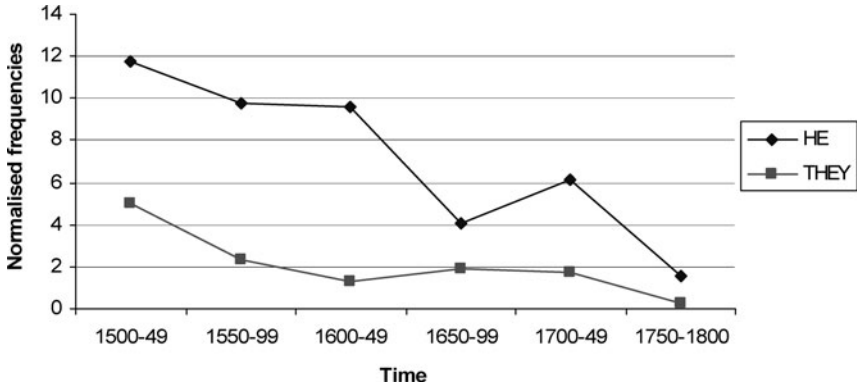


FIGURE 3. Normalized frequencies of anaphoric *he* and *they* with *-man* compounds in the CEEC (per 100,000 words).

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

The results confirm that the singular *he* with *-man* compounds correlated with the writers' gender. Even before the rise of prescriptive grammars, and long before the debates on sexism and gender equality of the late 20th century, epicene pronoun use must have been an issue recognized by some writers.

The results with *-one* and *-body* compounds corroborate the findings in *-man*. Figure 4, showing the proportion of *he*, illustrates that men tend to use both variants with *-one* and *-body* compounds in the 17th and 18th centuries.[12]

Examples (14) and (15) are from Arthur Capel. In the women's data, the plural variant is favored at statistically significant levels (16). As can be observed in Figure 4, women only occasionally resort to *he* with *-one* and *-body* antecedents.

(14) wherefore if Sir Cyrill Wyche will bring over ***any one*** to bee for that use [i.e., a secretary], I will see to do something for East . . . so as I conceive it may be best for Sir Cyrill Wyche when hee comes upon ye place to see how he likes ***them***. (Arthur Capel, 1676; ESSEX, 73)

(15) but such was ye unhappiness of that time as ye Treasury was quite empty, and ye insurrections in severall parts made ***every one*** who had money look to ***himself*** and none could bee found to lend to ye publick (Arthur Capel, 1676; ESSEX, 57)

(16) but in Earnest if I once finde ***any body*** faulty towards mee, ***they*** loose mee for ever . . . I dare sweare there is ***noebody*** soe proude of ***theire*** witt as to keep it upon such termes, but would bee glad after ***they*** had indured it a while (Dorothy Osborne, 1653; OSBORNE, 18)

In sum, the corpus data suggest stable gender differences in anaphora. The plural remains above 80% in women's data for more than a dozen generations of speakers, making gender stratification in anaphora a convincing candidate for a stable gender variable.

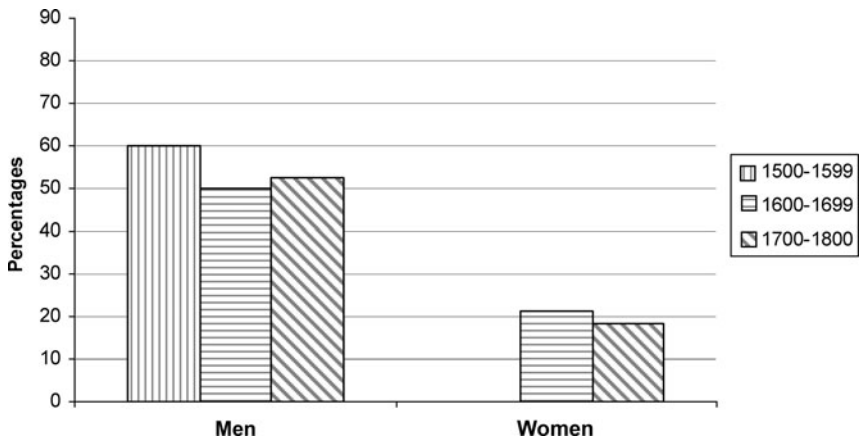


FIGURE 4. Percentages of the anaphoric *he* with *-one* and *-body* compounds in the CEEC.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

Gender in cataphora

The corpus results for men and women in cataphora, displayed in Figure 5, show the spread of *those* over the other two variants. They indicate that the change toward the demonstrative has nearly been completed in three centuries of English personal correspondence and is statistically highly significant for both men and women ($p < .001$).

The pattern of the incoming variant contrasts the gender differentiation exhibited in earlier studies in Early Modern English (Nevalainen & Raumolin-Brunberg, 2003:110–132). No clear gender affiliation can be detected in the spread of *those* in cataphoric personal reference; the fact that Figure 5 shows men leading the change in its early stages is not statistically significant. The figure also shows that gender affiliation remains consistent throughout the three centuries, as no clear crossover points of statistical significance are found.

Considering the Gender Paradox mentioned above, I have not yet determined the direction of change in the spread of *those*. I have suggested elsewhere that the comparisons between personal correspondence and the Bible (the *Authorized Version* of 1611[13]) indicate that the change toward the demonstrative *those* is from below (Laitinen, 2006). Table 3 shows the variant distribution in the early modern section 1 (1500–1570) from the *Helsinki Corpus*[14] (Kytö, 1996), which suggests that the texts more typical of the oral mode of communication tend to show higher proportions of the innovative *those* than the more literate genres or the *Tyndale Bible* of 1534. *Those* first spread to texts closer to the spoken idiom, confirming the direction of change from below in Labovian terms (1994:78).[15] Table 3 displays the percentages; the absolute number for the oral genres is 31, 81 for the literate, but for the Bible the figure is 68.

The cataphora results so far are in contrast with the data in anaphora in the previous section, as the innovative variant *those* diffuses along similar paths for both men and women.[16] More importantly, the results go against various earlier

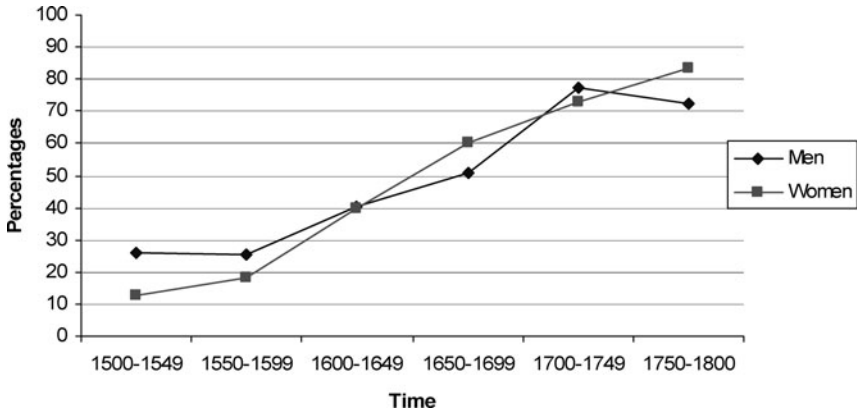


FIGURE 5. Gender variation in the emergence of cataphoric personal *those* in the CEEC.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

TABLE 3. *Cataphoric* he, they, *and* those *in the early modern section E1 (1500–1570) of the Helsinki Corpus*

	Oral Genres			Literate Genres			Bible (Tyndale)		
	He	*They*	*Those*	*He*	*They*	*Those*	*He*	*They*	*Those*
E1, 1500–1570	22%	61%	17%	41%	56%	4%	59%	40%	1%

studies in gender differentiation, because one cannot discern the more active role of women, as suggested by Labov (2001:292–293). My initial suggestion based on these corpus results is that the process was largely driven by language-internal factors, such as the tendency toward a cognitively more unambiguous variant, as argued in the section entitled "The diachronic developments in the data," and the propensity to avoid personal pronoun forms that require postmodification, as proposed by Denison (2005). In sociolinguistic terms, it is plausible that the innovative form never rose into consciousness, unlike the development toward *they* in anaphora, which attracted attention in normative grammars in the 18th century (Bodine, 1990).

Gendered variant patterns emerge when we examine the entire variable. Figure 6 shows the three centuries of cataphoric personal reference with all three variants displayed. It illustrates that cataphoric variants for men were largely interchangeable in the 16th century, up to the point when *those* began to spread.

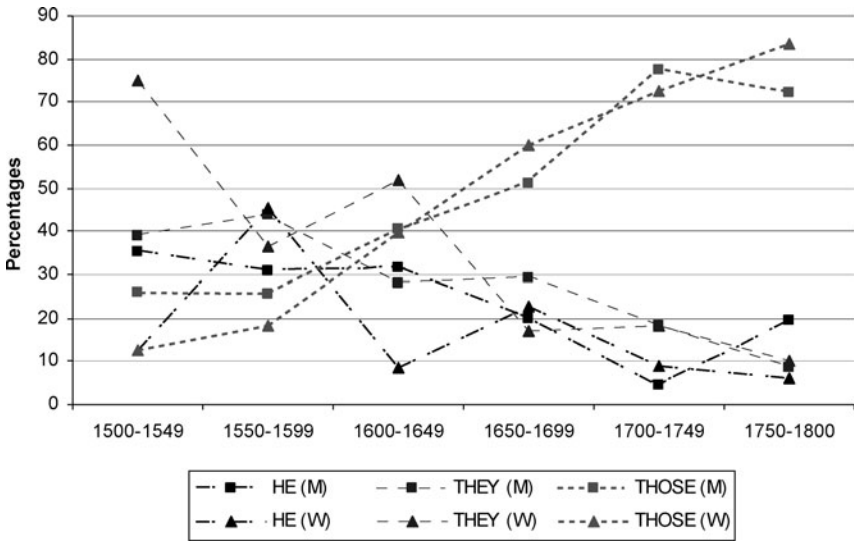


FIGURE 6. The three cataphoric pronoun variants in men's and women's material in the CEEC.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

Women's data, more infrequent in terms of tokens, indicate that they preferred the *they* variant in cataphora, which was then replaced by *those* in the 17th century.

A microlevel examination of the variants shows that women who were the leaders of change in anaphora, using the plural *they* consistently, do not necessarily participate in the change toward *those* in cataphora. A case in point is Arabella Stuart, who has 22 examples of *they* in anaphora and uses the same form in the majority of her cataphoric cases. She has 13 occurrences of *they*, see (17) and (18), and only 2 instances of the innovative *those*.

(17) But do not deceive your selfe so much to thinck I either have or will confesse my pure and innocent selfe guilty of love till you deserve that extraordinary trust (which **they who** for the saving of **theyr** soule submitt **them selves** to Auriculer confession, have all the assurance one mortall man can give or an other possibly require) many waies (Arabella Stuart, 1603; STUART, 165–166)

(18) Theare is no trust in man whose breath is in his nosthrilles. And the day will comm when **they thatt** judge shall be judged, and he that now keepeth theyr counsell and seemeth to winke at iniquity (Arabella Stuart, 1603; STUART, 157)

Discussion of the role of gender

The findings in the two previous sections not only add to our knowledge of gender as a social variable but also provide new evidence of stable variation and help us distinguish some linguistic characteristics that tend to promote stability. The conclusions are based on a mixed set of results because the corpus data show that the variant use in anaphora was heavily conditioned by gender in early and late modern England (1500–1800), whereas the development of *those* in cataphora seems to have proceeded independently of the gender of the writers. In anaphoric contexts, women were on the vanguard of the slow long-term development, as they led the stronger tendency toward the plural variant. As illustrated in Figure 2, the pronoun use in anaphora was stratified along gender lines throughout the period examined, and women strongly disfavored the masculine *he* in anaphoric epicene uses. These results of stable variation in anaphora suggest that one should find similar gendered patterns promoting linguistic stability in those sites in grammar where gender is displayed or where gender is indexed in lexemes or morphemes.

As mentioned earlier, the period examined partially corresponds with the first normative statements in English, which proscribed the break in number concord and promoted the new high prestige form, *he*. The following quote is from Murray's grammar of 1795 (see also Kirby, 1746:56, 80).

> Rule V. Pronouns must always agree with their antecedents, and the nouns for which they stand, in gender, number, and person; . . . Of this rule there are many violations to be met with; a few of which may be sufficient to put the learner on his guard. "Each of the sexes should keep within its particular bounds, and content themselves with the

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

> advantages of their particular districts." "Can any one, on their entrance into the world, be fully secure that they shall not be deceived?" "on his entrance," and "that he shall." "Let each esteem others better than themselves;" "than himself" (Murray, 1795:95–96).

One might hypothesize that the slow advancement of the plural illustrated in Figure 2 might have led to it rising in the level of social awareness, which in Weinreich et al.'s (1968) terminology leads to "the establishment of a social stereotype." The results in anaphora show that it was the men's variant that the grammarians selected as the acceptable epicene pronoun, but the variant promoted by women eventually became the prestige variant in the 20th century. Studies focusing on the role of gender in anaphora in present-day English could provide new information not only on prestige forms but also on gendered patterns in stable variables.

Finally, the corpus results indicate that we can use sociolinguistic, correlational methods in our study of larger morphosemantic or even typological changes. This approach includes using balanced corpus material, which allows us access to sociolinguistic information. It also requires carefully defining the linguistic functions and forms that are correlated with robust social categories, such as the gender of the writers. This approach also means integrating the internal dynamism of change, evidenced in cataphora or the variant distributions with the three paradigmatic alternatives *-one*, *-body*, and *-man*, for instance, with external development such as social categories. This type of analysis results in an integrated model of linguistic performance, which allows a scholar to examine the possible interfaces between these two levels. Such an approach in synchronic linguistics has been suggested by Lesley Milroy (2004) in studies of social dynamics of sound change. In the following sections, I will extend my correlations of the dependent variables to cover additional social categories.

REGISTER

Shifting the attention from interspeaker to intraspeaker variation (Schilling-Estes, 2004), I will discuss the role of situation of use, or register, in pronoun selection. The data in Tables 1 and 2 above were sorted in terms of tenor, writer-addressee relations, according to the model proposed by Halliday (2001:631).

My research questions in this section are twofold, depending on the pronominal context. The first one, which concerns the anaphoric structures, tests whether one can find patterns of variation in *he/they* in Early Modern English correspondence that are similar to those in present-day English registers. I have proposed elsewhere that the informal text types increase the writers' tendency to use the plural, whereas formal texts show a statistically significant increase in the singular (Laitinen, 2002). This section dynamizes the present-day situation, exploring the origins of such differences in Early and Late Modern English. The data are sorted to see whether it is possible to determine the direction of change of the emerging variant *they*.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

The second question focuses on the spread of the innovative *those* variant in cataphora in terms of its textual variation. As previously noted, the Helsinki Corpus data in Table 3 supports the idea of its emergence from below. Labov (1994) predicts that most linguistic changes originate in the vernacular, informal varieties of language, only to spread later to the more formal ones. The corpus used here has been coded for sender-addressee relations, enabling research on tenor (Nevalainen & Raumolin-Brunberg, 2003; Palander-Collin, 1999). A detailed discussion of the parameter coding in the CEEC appears in Nevalainen and Raumolin-Brunberg (1996:39–54). I divided the letter material into two subcorpora, and the dependent variable was then cross-tabulated with these binary categories. My aim is to explore whether Labov's (1994) widely tested observation could be detected in statistically significant pronoun variation if we compared letters written to family members and close friends, representing the more informal situations, with those addressed to all other recipients, that is, the more formal situations of use.

Register variation in anaphora

Figure 7 illustrates the proportion of the plural *they* variant in the CEEC divided according to the situation of use. It depicts the plural pronoun variant in letters to family members and close friends (coded FF) and to other recipients (O). No statistically significant differences in the anaphoric pronoun selection can be detected up to the end of the 18th century. The variant frequencies fluctuate, but in general, the letters written to family members and friends contain roughly equal numbers of the singular and plural pronouns as the other recipients. When one uses the writer-addressee relationship as a binary category in the examination, the figure allows one to conclude that register does not determine the anaphoric pronoun use in Early Modern English. There is, however, one exception that might help us to predict the sociohistorical development of the variant.

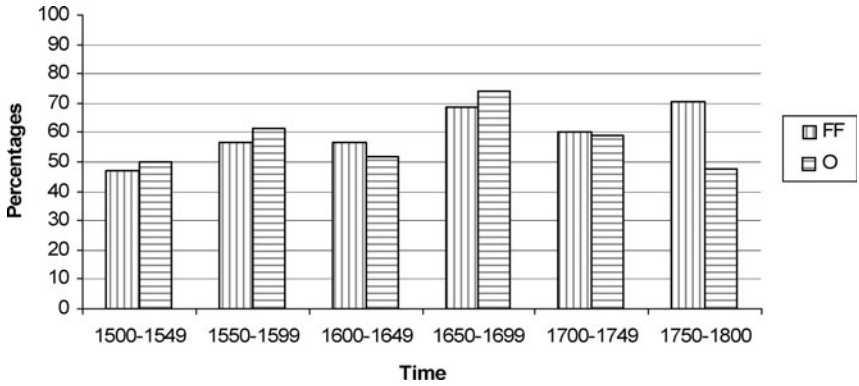


FIGURE 7. The proportions of anaphoric *they* according to the situation of use in the CEEC.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

The corpus results depicted in Figure 7 initially suggest that the epicene *they* emerged to social awareness in the middle to late 18th century, because the last 50-year period shows register differentiation of statistically significant levels ($p < .05$).[17] This remains an initial hypothesis and requires further research. However, if *they* indeed rose above the level of consciousness as an "incorrect" variant, it could have become restricted more to communicative situations characterized by casual style, as it is often defined in present-day synchronic sociolinguistics (Labov, 2001:95*ff.*). In turn, it would have been avoided in more careful and formal situations that, according to Finegan and Biber (1994), tend to contain more frequent uses of elaborate variants, such as *he*. The corpus results in this section strongly suggest that the origin of this textual distribution is found in the late 18th century.

Register variation in cataphora

Figure 8 presents the data on the incoming cataphoric *those* in the two registers, which shows that the situation of use was not a determining factor in the spread of the demonstrative and the eventual demise of the personal pronouns *he* and *they* in cataphoric personal reference. The 16th-century data on the new and vigorous stage initially suggests that *those* was used more in the letters to other addressees (O), but the differences in the absolute numbers are not statistically significant. The change towards *those* proceeds at equal speed irrespective of register.

The fact that no register variation in terms of tenor, writer-addressee relations, is detected does not necessarily mean that no textual variation existed during the three centuries during which *those* came to replace the other variants. As mentioned above, the Bible and more literate genres present us with an entirely different view of the variant selection. Taking a broader view of the various external factors considered, my suggestion is that the data in Figure 8 only confirms our

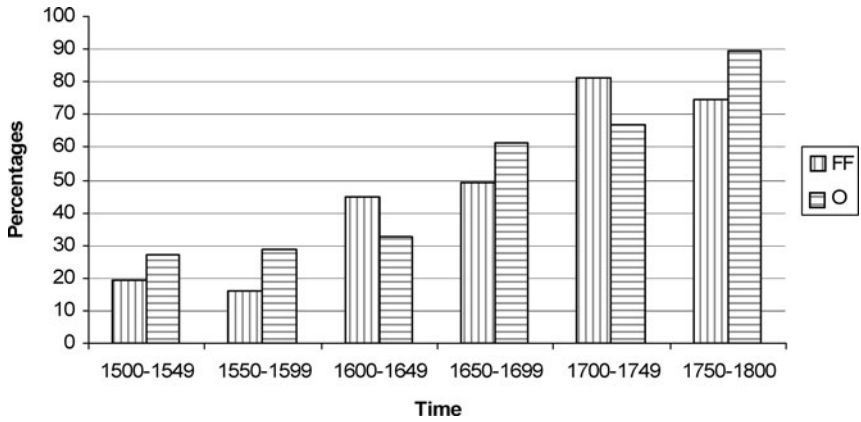


FIGURE 8. The proportions of cataphoric *those* according to the situation of use in the CEEC.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

initial proposition about the rise of *those* in cataphora. The change probably never rose to awareness and was to some extent driven by language-internal factors.

Considerable microlevel variation is found in the corpus evidence, but there are few writers whose variant use is differentiated along register lines. Most of the informants are writers with mixed variants. Some notable exceptions include Nathaniel Bacon in the 16th century, who uses *he* with a relative clause in nonfamily or non-close friend letters, (19), and *they* mainly in family correspondence, as in (20) and (21).

(19) that either you do upholde & maynteyne the right you put over to me or otherwise make restitucion of that which hath ben taken. Yt is a better tyme to complayne now then to have complayned when **he who** {did you the injury} [dealt thus with you] should have ben your judge. And in my opynion (Nathaniel Bacon, 1591; BACON III, 114–115)

(20) and if it maie [be] compassed without law it shalbe a follie to enterprise the same by suit. You write further that this will make others afrayde what waie to take with the litle they have. The office of trust is a burden to **them who** will perfourme the same honestlie, and I begg not for the place, and I shalbe content **they who** do laie such an office upon mee do know that I take **them** beholding unto mee (Nathaniel Bacon, 1592; BACON III, 219)

(21) For I knowe very well it will be for her benefit especially, & therby for mine also. Yt is no newe thinge that **thei which** love **ther** children best, beinge about these yeares that my wife is, seake to place them for a time in some good place from them. (Nathaniel Bacon, 1570S; BACON I, 12)

In the 17th century, the illustrations from John Jones are good examples of microlevel variation in terms of register. He uses the incoming form *those* mostly in his letters relating to more formal relationships. Jones, a puritan and a member of Cromwell's inner circle, wrote to a large group of addressees, such as a fellow colonel and the member of the Council of State, Philip Jones, in (22).

(22) The suddennesse of the late change, with the occasion and somewhat of the formality of it, doth amuse many precious S'^{ts} in this Lande, and it cannot be rationally expected that there should be soe sudden and cleare an aprobation in it in **those that** feare the Lord, as there were in them y^{t} were p'sent upon the place, and know the grounds and Reasons of this alterac~on. (John Jones, 1653; JONES, 219)

In more informal situations, Jones resorts to the old *he* and *they* forms.

(23) I am not sorry for y^{e} Welch Petic~on, the more the Saints are tryed, the more their luster will appeare, it is good to observe what hands promote it, who **they be that** subscribe it, and to take exact notice what temper **they** are of, and what kind of ministry **they** are and would promote. (John Jones, 1652; JONES, 210)

(24) If a stone or piece of Earth were placed in the braine and at the heart, how could and heavy would the Speritt be, even such are the affaires of the world

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

(although honest and necessary.) And therefore happy are **they that are** least incumbred with them (John Jones, 1652; JONES, 211–212)

(25) The Lord is pleased to begin to gather a people here at Dublin into a Church fellowshipp, the foundac~ons off 2 Congregac~ons are already laid, y^e^ number of either not many, but **he that** soweth in this vineyard, will give y^e^ encrease. (John Jones, 1651; JONES, 191)

Morgan Lloyd, the recipient of (23) and (24), was Jones's friend in Wales, as was William Stane in (25), a doctor to who Jones reported private matters concerning the life and selected political events in Dublin.

In this section, I have reported the first diachronic results the pronominal variants in cataphoric personal reference addressing variation in writer-addressee relationships. As little previous research on this structure exists, the textual distribution of the variants offers us an interesting field for future research, such as mapping the variant uses in various text types, such as legal texts, pamphlets, and scientific writings.

REGION

The next issue to be considered is the regional variation in the pronoun variables, with the intention of discovering the possible areal origin of the emerging *they* in anaphora and *those* in cataphora. The previous sections established that there was a slow long-term shift toward *they* in anaphora and a much quicker movement of cataphoric *those* ousting the personal-pronoun variants. The diachronic shift in anaphora in particular included the development in which one variant eventually rose to awareness, as *he* was selected as the prestige form in the 18th century. The question considered in this section is whether we can pinpoint an area from which the new prestige form in anaphora originates and to recognize any area in which the developing form in cataphora was used at statistically significant levels before other regions.

As I analyze low-frequency items in the study, no very detailed analysis of regional variation can be achieved. This concerns supralocalization of local linguistic items in particular, as suggested by Nevalainen (2000b), and the geolinguistic approaches by Hernández-Campoy and Conde-Silvestre (2005). The rationale for this section lies in the fact that no earlier diachronic studies accounting for regional variation of the pronouns in indefinite sense exist. There is limited previous scholarship on present-day English; Baranowski (2002) has shown that American writers tend to choose the anaphoric *they* in fewer frequencies than British writers do. Variation is attested in the antecedents in anaphora shown in Figure 1 above. Biber et al. demonstrate that *-one* forms tend to be more common in British English and *-body* in American English (1999:351–353). Taking a diachronic approach to the antecedents in this study, Nevalainen and Raumolin-Brunberg (2003:181) suggest a pattern in which the *-body* forms predominate in London, the Royal Court, and East Anglia, but not

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

in the North, where *-one* compounds are favored. The *-one* indefinites supralocalize in the mid-17th century, but those of *-body* do not advance to the North. Previous scholarship on cataphora is more limited, as the first account of regional variation in diffusion of innovative pronoun antecedent forms are attempted here. The scholarship has mainly focused on the shifts in the relative pronoun paradigms and less so on the geographic distribution of the variants (Dekeyser, 1984; Nevalainen, 1987; Rydén, 1966). The most notable exception is Raumolin-Brunberg (2000), concentrating on the spread of *which* over the Northern form *the which*.

For this study of regional pronoun distribution, I have established a three-level distinction that accounts for regional linguistic variability in early modern England proposed in previous historical sociolinguistic studies (Nevalainen & Raumolin-Brunberg, 2003:159–167). First, the threefold distinction here combines the residents of the capital area of London as one group, comprising merchants, lower gentry and occasional noblemen and women in the city. Second, members of the royal household, courtiers, government officials, and diplomats, many of whom lived in Westminster, have been grouped as representatives of the royal court. Contemporary language commentaries of the 16th and 17th centuries show that people of the "better sort" in London and particularly at court enjoyed a position of prestige, which could well be the source of the incoming pronominal forms. The most notable of these opinions is George Puttenham's advice to aspiring poets (1589). Third, in order to allow statistical generalizations based on my relatively low-frequency dependent variables, informants from other areas in England comprise the third group. In essence, the three-level distinction used here collapses both geographic and social distinctions to overcome limitations in the overall variant frequency. With the development of even larger historical data sources, it should be interesting to investigate whether pronoun variant distributions and frequencies exhibit similar regional distinctions that have been attested to operate on the north-south axis (Bailey et al., 1989; Nevalainen & Raumolin-Brunberg, 2003:177–180). The figures in this section show the variant distribution across two centuries, because the informants' geographic distribution in the CEEC's eighteenth century *Extension* at the time of the corpus queries was not strictly comparable to the earlier time periods in which the merchant letters in London, in particular, allow a flexible study of their language (see Nevalainen & Raumolin-Brunberg, 2003:157–184).

Figure 9 shows that, in the 16th century, the anaphoric *they* variant was not heavily favored in London, the court, or the other areas. The statistical differences in the absolute numbers between London, which shows a higher proportion of *they* in the early 16th century, and the court and the other areas are not significant. With a few fluctuations in the frequencies brought about by the small number of tokens at particular periods in certain areas, the 200 years of correspondence result in the situation in which *they* is on the verge of becoming the major variant in anaphora. The absolute figures show that in the early 16th-century material roughly half of the anaphoric cases are in the singular, whereas

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

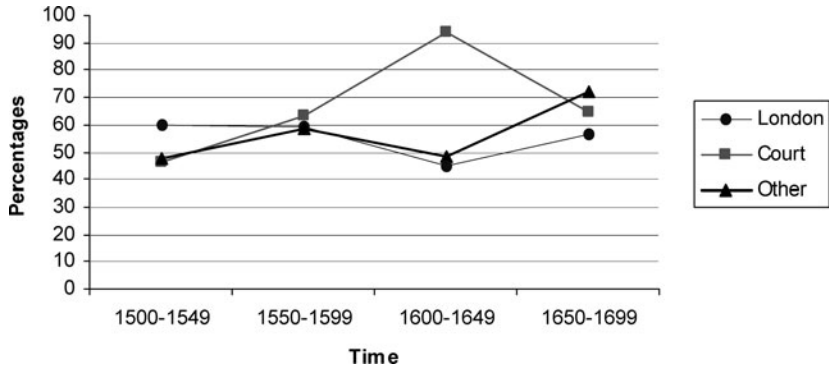


FIGURE 9. The proportion of anaphoric *they* in three regions in the CEEC.

the proportion has been reduced to around 30% by the end of the 17th century. The only statistically significant difference in the areal distribution in Figure 9 is found in the other areas. The difference between the two centuries there is statistically highly significant ($p < .005$). The data from the court seem to follow the trend of the other areas, but low absolute frequencies skew the statistical significance analysis. The peak in the material from the court in the early 17th century is skewed by one individual, Arabella Stuart, whose examples, all in the plural, constitute the majority of the data from the court at that particular period.[18]

We are only beginning to understand the anaphora variant distribution. This research indicates that not only is more data needed, but that the possible source of the middle-to-late-18th-century proscription of *they* in normative grammars was the fact that the plural *they* in anaphora was dominant in areas other than London, in addition to its predominance in women's writing, as has already been shown.

The results in Figure 10 indicate that *those* was practically nonexistent in correspondence originating from London in the 15th century. The proportions in letters from the court and the other areas are higher; London only reaches these two areas when *those* reaches the midrange (from 36 to 65%) of the S-curve. The change toward the demonstrative variant proceeds with equal speed in the court and other areas. No statistical significances are attested between them in any of the periods. When the change is at its most rapid in the 17th century, the end of the century in particular, the data indicates that London lags behind. If we elaborate on this detail, an interesting feature of the cataphoric variants is revealed when only the *they* variant is examined. The reason why *those* spreads at unequal speed in London is shown in Figure 11, which indicates that London retains the personal-pronoun variant *they* longer than the rest of the country. The demonstrative only becomes the major variant in London in the following century, whereas at the court *they* loses its position earlier, and in the other areas even prior to that.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

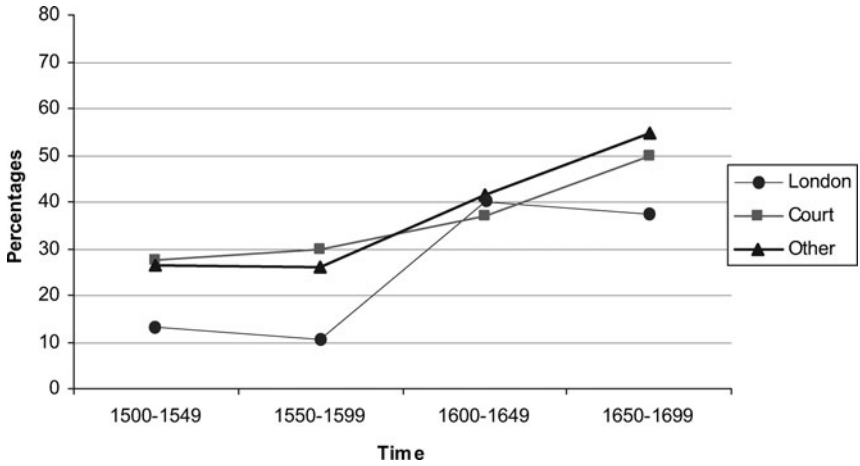


FIGURE 10. The proportion of cataphoric *those* in the three regions in the CEEC.

All in all, the three sections have shown that the paths of pronoun diffusion in terms of extralinguistic independent variants were highly dissimilar, depending on the context. The role of women was highlighted in the anaphoric *they* variant, whereas no statistically significant distribution of gender stratification was observed in cataphora. The data on register variation suggested that sender-addressee relations played a role in anaphora only after *they* had risen above the level of consciousness, but in cataphora, explanations of the emergence of *those* must be sought elsewhere. These results also show statistical significance in certain aspects of regional variation. To gain a fuller picture of the variant diffusion, the next section turns our attention to measuring the interplay of some of the external factors discussed in the previous sections.

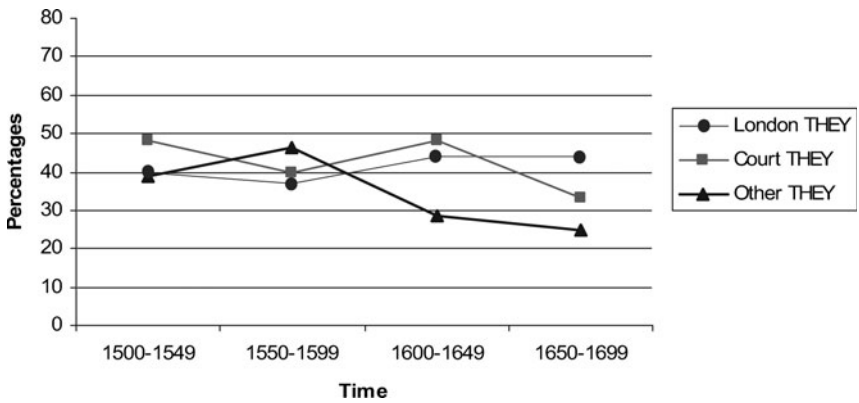


FIGURE 11. The proportion of cataphoric *they* in the three regions in the CEEC.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

MULTIVARIATE ANALYSIS

To highlight the role of abstract morphosemantic functions that are expressed using the various pronoun forms, the last part of my article focuses on statistical analyses in which the observed variation in language external factors is compared with the data from one language-internal factor. The corpora used offer sufficient data for studying the interplay between external factors, and the statistical methods allow us to include language-internal factors in the analysis. These include the type of antecedent in anaphoric contexts and the relative pronoun form in cataphora. No direct comparisons between the internal and external variables are made, and the factor weights of the statistical multivariate analysis are given independent of each other. The multivariate analyses are carried out to estimate the relative weight of independent variables on the choice between the pronominal variants. I used the Windows version of GoldVarb 2001, originally developed by Rand and Sankoff (1990).

The VARBRUL results of anaphoric and cataphoric contexts are presented independent of each other. Following Paolillo's suggestion (2001) for handling sparse data and avoiding knockout values in the cells, the results are divided into centuries, allowing us to analyze linguistic variation of low-frequency items. The main interest lies in the writers' selection of the innovative *they* over *he* in anaphora and *those* over *they*/*he* in cataphora, the constraints for the factors being presented in terms of their probability of appearing with the innovative variant. The external, social factors whose interplay is investigated include gender and register as discussed earlier.[19]

In anaphora, the language-internal factor is the form of the antecedent element, more particularly, the choice between the four paradigmatic alternatives in Figure 1: (a) the compounds in *-man*; (b) the compound forms of *-one* and *-body*; (c) simple, independent indefinite pronouns; and (d) noun phrases of determiners and generic nouns. Figure 12 displays the chronological shifts in the frequencies of *he* and *they* with these forms. The figure shows the internal dynamism in anaphoric epicene pronouns brought about by the changes in the antecedents. In the 16th century, *-man* indefinites were referred to with singular variants, whereas the simple, independent forms received the plural *they*. The other types of reference were a clear minority. The 18th-century variation is more diverse, as the other antecedent types had increased their proportion, projected in the pronoun variants in Figure 12.

The VARBRUL results for three centuries of anaphora are in Table 4, which shows the three factors, and the maximum and minimum values for the factor weights, which vary between 0 and 1.[20] The variation ranges are calculated by subtracting the minimum value from the maximum. The range of variation establishes the relative impact of the independent factors in the slow shift toward *they*. The pattern emerging from the statistical analysis is clear, corroborating the results in the cross-tabulations above. Table 4 confirms that the slow spread of the *they* variant is promoted by female writers. Throughout the three centuries, the variation range is considerably higher in gender than in register variation.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

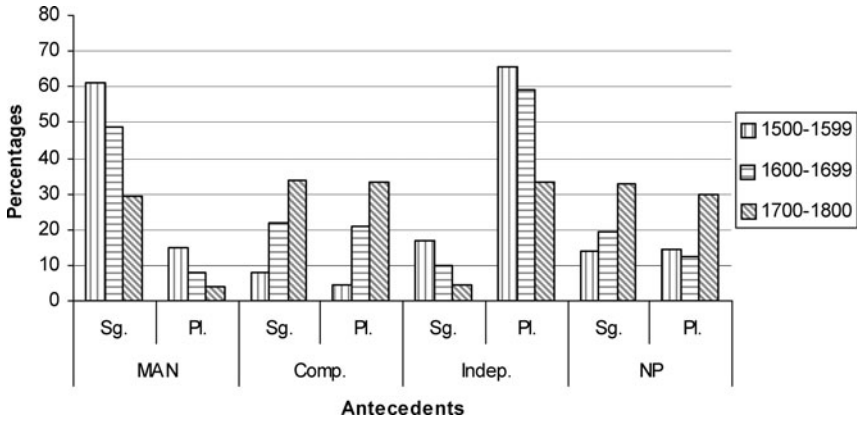


FIGURE 12. The anaphoric variants singular *he* and plural *they* with the different antecedent types.

As to the language-internal factors, Table 4 shows that the independent antecedent types clearly promote the uses of *they*, an effect which is consistent in the time frame. The indefinite pronouns in *-man*, on the other hand, inhibit it. The variation ranges of the language-internal factor are indeed greatest throughout the three centuries examined. These results indicate the role of changing antecedent types, the indefinite pronouns, to be highly relevant.

In cataphora, the basis for language-internal factors is the relative pronoun type, the relative pronouns being *who* and *that*.[21] Previous scholarship on relative structures has shown that *who* as a relative of personal antecedents was introduced later than *that* (Rydén, 1966). My research question concerns the relative impact of the relativizers along with the social factors of gender and register on the selection of the innovative *those*. Figure 13 shows that even though *who* is later than *that*, *he who* is more frequent in indefinite reference than *he that* in the 16th century. When used with the demonstrative pronouns, *those that* is more frequent than *those who* at the early stages of its emergence. *Those who* only picks up after the increase in *who* in the 17th century. *Who* gained ground with *those* in cataphora only once it had been established as the relative for personal antecedents. *He who* and *he that* then started to decline in frequency.[22]

The three periods show the spread of *those* first at its new and vigorous stage (the 16th century at 25% level), at midrange (47%), and finally in the 18th century when nearing completion (74%). The constraint probabilities shown in the three tables indicate the strength of *those* appearing in the data. The factors whose relative weightings have been inspected are gender, register, and the relative pronoun type.

The VARBRUL results for the three centuries are shown in Table 5, which shows that gender variation is a more significant language-external factor than register variation in the three phases of change. Men favor the innovative form at the early stages of change, and the differences in the midrange are minimal.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

TABLE 4. *The probability of anaphoric* they *in the multivariate analysis of the three factors in three centuries (1500–1800).*

Period	Factor	Factor weight	%	*N*
1500–1599	Gender (range = .388)			
	Women	.869	92	13
	Men	.481	53	185
	Register (range = .133)			
	Distant	.536	57	149
	Close	.403	50	49
	Antecedent (range = .585)			
	Independent	.780	82	130
	NP	.481	56	29
	-one/-body	.343	40	9
	-man	.195	23	30
	Input:	.578		
	Total *N*:			198
1600–1699	Gender (range = .303)			
	Women	.737	83	206
	Men	.434	56	78
	Register (range = .049)			
	Close	.518	63	183
	Distant	.469	60	101
	Antecedent (range = .704)			
	Independent	.839	90	168
	-one/-body	.358	60	59
	NP	.317	51	35
	-man	.135	20	22
	Input:	.683		
	Total *N*:			284
1700–1800	Gender (range = .402)			
	Women	.761	84	67
	Men	.359	53	84
	Register (range = .000)			
	Distant	.500	54	27
	Close	.500	66	124
	Antecedent (range = .762)			
	Independent	.874	92	51
	NP	.478	61	45
	-one/-body	.386	62	49
	-man	.112	19	6
	Input:	.695		
	Total *N*:			151

The table shows that female leadership of this particular change is assumed only when the diffusion of *those* is nearing completion in the 18th century. The only notable impacts (weightings over .550) in social variables occur in the last period with women clearly promoting the use of *those*.

Separating the impact of the relative pronoun variants from gender and register, Table 5 shows that the relatives played a significant role in the adoption of *those* over *they* in cataphora. Keeping in mind Figure 13, which indicates that the

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

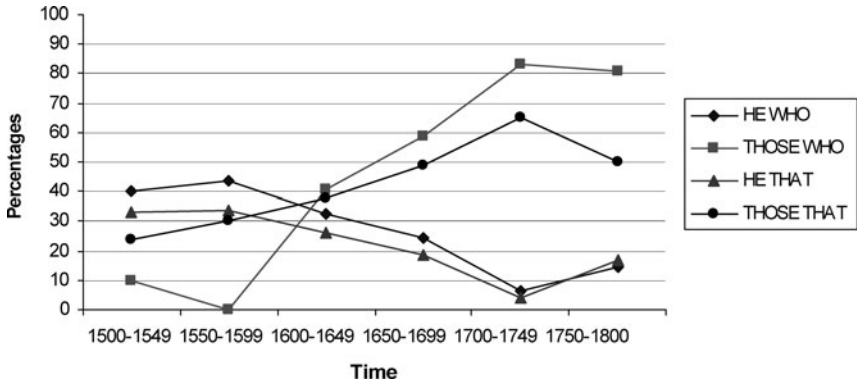


FIGURE 13. The cataphoric variants *he* and *those* with the two relative pronoun types.

TABLE 5. *The probability of cataphoric* those *in the multivariate analysis of the three factors in three centuries (1500–1800)*

1500–1599	Gender (range = .209)			
	Men	.520	37	45
	en	.311	16	2
	Register (range = .144)			
	Distant	.539	40	39
	Close	.395	22	8
	Relative pronoun (range = .473)			
	that	.586	41	46
	who	.113	5	1
	Input:	.319		
	Total *N*:			47
1600–1700	Gender (range = .085)			
	Men	.521	61	123
	Women	.436	53	34
	Register (range = .075)			
	Close	.525	61	108
	Distant	.450	56	49
	Relative pronoun (range = .154)			
	who	.611	69	51
	that	.457	55	106
	Input:	.598		
	Total *N*:			157
1700–1800	Gender (range = .218)			
	Women	.643	88	48
	Men	.425	85	90
	Register (range = .009)			
	Distant	.507	87	29
	Close	.498	86	109
	Relative pronoun (range = .481)			
	who	.629	93	114
	that	.148	64	24
	Input:	.903		
	Total *N*:			138

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

disappearing *he who* was the dominant variant in the 15th century, Table 5 shows that the range of variation with the relatives exceeds that of the other factors in the 16th century. The use of *that* promotes the innovative *those* at the early stages. *Who* is used in cataphora in the 16th century, but predominantly with the disappearing *he* and *they* variants. The range of variation in the early stages of the diffusion of *those* with *that* is greater than in the 17th century when the change is at its midrange phase (.473 versus .154). When *those* is nearing completion, the variation favoring the relative is once again markedly similar to the situation in the 16th century, but this time the relative is *who*.

CONCLUSIONS

This article has explored the sociolinguistic patterns of a development that concerns typological tendencies in English. I have focused on three pronoun variants that may be used in a human indefinite function but which can be realized in two contexts: (a) human indefinite anaphora and (b) cataphoric personal reference. Using quantitative corpus evidence, it was argued that the less marked pronoun variants (in terms of gender, number, and linguistic function) assumed an expanded functional role, which is related to a larger change in English pronoun use (cf. Laitinen, forthcoming). My approach in this paper has been inspired by two theoretical proposals. First, the decision to include the pronominal realizations of both anaphora and cataphora was prompted by Croft (2003:271), who proposes that a diachronic study of grammaticalization could benefit from investigating not only individual linguistic elements, but also larger constructions. The study presented here naturally deals with abstract semantic constructions such as realizations of indefiniteness. Second, the framework for the sociolinguistic enquiry comes from the suggestion made by Labov (2001) that more abstract linguistic phenomena do not necessarily exhibit sociolinguistic patterns of socially evaluated variables. The present results are, to my knowledge, the first systematic corpus-based account of anaphoric and cataphoric pronoun realizations in the indefinite sense in Early and Late Modern English in which the pronominal variants are correlated with social variables. More importantly, I hope that the approach adopted here could be repeated with other morphosemantic functions.

My findings can be summed up as follows. The corpus data from CEEC and its *Extension* strongly suggest that the diachronic evolution in both contexts was highly similar, leading to the diffusion of one variant at the expense of the others. Despite the similarities, the developments were also highly divergent. In anaphora, the corpus data indicate that *they* in an anaphoric relationship to indefinite pronouns has been in use for centuries. There was indeed a change toward the greater tendency to *they*, but the shift has been extremely slow, and barely observable even in large corpora. In cataphora, on the other hand, the 1500–1800 period saw the near completion of the change toward the demonstrative *those* over the personal-pronoun variants *he* and *they*.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

The three language-external, social variables were gender, sender-addressee relations, and region. The cross-tabulations demonstrate that the writers' gender in anaphora, in particular, turned out to be interesting. The data indicate that women clearly preferred the plural *they* variant to a degree that represents a clear case of a stable sociolinguistic variable exhibiting long-term variation persisting over several centuries. Women use *they* in proportions over 80% throughout the three centuries, whereas men are more equally divided between the two variants. The results for register variation also initially suggest that *they* became a sociolinguistic stereotype in the Labovian sense in the course of the 18th century (Labov, 1972:314–315).

In cataphora, the development toward *those* seems to proceed independently of the social variables, as the corpus data show no clear-cut correlations with any one independent variable. The comparisons with the results from the multigenre *Helsinki Corpus* indicate that the direction in change was from below. My suggestion is that the process of change in cataphora was partly driven by language-internal factors; the innovative form, *those*, never rose into consciousness, and the change was allowed to proceed, unlike the development of *they* in anaphora, which attracted attention in normative grammars.

In broad terms, this article has taken a sociolinguistic approach to grammaticalization. This process has traditionally been seen as language-internally driven, but, as J. Milroy (2005) argues, language change takes place from person to person. The axis of internal *versus* external factors has sometimes been seen as a dichotomy in which one excludes the other instead of complementing each other and offering two interrelated views to language change. The article, therefore, introduced a linguistic component into the analysis. A multivariate analysis (VARBRUL) was carried out to estimate the relative impact of the social variables and the linguistic elements in the immediate cotext of our pronoun variants. This analysis suggested that social variation in terms of gender exceeds register variation, corroborating the implicational scale in Nevalainen and Raumolin-Brunberg (2003:199). The multivariate statistical testing also indicates that in both contexts there was a considerable linguistic component involved in the change. That is to say that the linguistic subsystems in these constructions interacted heavily with the pronoun variants along with the social factors.

NOTES

1. The main data come from the *Corpus of Early English Correspondence* and its *Extension*.

2. The nominative forms *he* and *they* are used as generic terms in this article, including the more adjectival possessive pronouns *his* and *their*, the nonreflexive oblique cases *him* and *them*, and the reflexive pronouns ending with *-self* and *-selves*.

3. The pronouns *she, it*, and disjunctive forms *he* or *she* may also function as epicenes, but it should be noted that no such cases have been found in the texts for this study; this, of course, does not mean that they do not occur. Present-day English studies (Baranowski, 2002; Gerner, 2000) have shown that *he* and *they* are the major alternatives; to make the manual determination of anaphoric relations feasible, the research question only includes pronoun number as the category to be scrutinized.

4. The terms, anaphora and cataphora, are used as common names to cover the two linguistic contexts in which the pronoun variants occur. It should of course be noted that there is a wide range of relations in person deixis not covered in this article. Readers are referred to Huddleston and Pullum (2002:1449–1564) for a comprehensive discussion of these two linguistic phenomena.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

5. I have discussed the indefiniteness requirements in Laitinen (2002; 2006). The criteria are not elaborated further in this article.

6. Rydén's data from Sir Thomas Elyot's prose show that generic *he* or *they* followed by a relative clause (especially with *which* and *that*) was very frequent in the early 16th century (990 occurrences), whereas *those* plus a relative clause was found only in 29 cases, and not all of them had an indefinite, generic sense (1966:6–8, 60–62, 157–160).

7. My study is carried out in the variationist framework, which is based on the assumption that no change is abrupt but rather always involves variation (Weinreich et al., 1968:188). In broad terms, the article contributes to the basic fact-finding in historical sociolinguistic work on grammaticalization. The empirical part provides one answer to Janda's (2001) challenge to historical sociolinguistic research, suggesting that transmission of grammaticalization in language change affects speech communities at large, not between generations.

8. CEEC—*Corpus of Early English Correspondence* (1998). Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin at the Department of English, University of Helsinki. CEECE—*Corpus of Early English Correspondence Extension* (in progress). Compiled by Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Terttu Nevalainen, Arja Nurmi, Minna Palander-Collin, Helena Raumolin-Brunberg, Anni Sairio (née Vuorinen) and Tuuli Tahko at the Department of English, University of Helsinki.

9. This pronoun variation essentially concerns the aftermath of the restructuring in the indefinite pronoun system, which caused writers to emphasize indefiniteness in larger proportions of the plural anaphora.

10. In (10), the two nominal lexemes illustrate the restructuring that the indefinite pronouns underwent at the paradigmatic level, and Spencer Cowper selects both *-man* and *-one* in his indefinites.

11. All the relevant occurrences of both anaphora and cataphora have been analyzed and meet the indefiniteness requirements detailed in Laitinen (2002; 2006). In this respect, the cataphoric cases are easier to define than anaphoric. See also Curzan's (2003) detailed discussion of the problems facing a historical corpus linguist in identifying referents in anaphoric relationships (Appendix 2).

12. The 16th century has been excluded, because they were only beginning to emerge at the time, as has been shown by Raumolin-Brunberg (1994), and the frequencies remain low.

13. Authorized Version—The Holy Bible, containing the Old and New Testaments, translated out of the original tongues: and with the former translations diligently compared and revised by his majesty's special command. Authorized King James Version. Oxford: Printed at the University Press, London: Oxford University Press. The Authorized Version was accessed electronically at http://www.hti.umich.edu/k/kjv/index.html.

14. Helsinki Corpus—*Diachronic part of the Helsinki Corpus of English texts*. Available at http://helmer.aksis.uib.no/icame/hc/index.htm.

15. My division of the texts into genres follows that used in Raumolin-Brunberg (2005). The results from the Helsinki Corpus in Table 3 are limited and serve the purpose of complementing the main data of the CEEC. It should also be noted that most of the texts in the early modern section E1 in the Helsinki Corpus are later than the *Tyndale Bible* of 1534.

16. I have not considered social directionality of diffusion, because Figure 5 does not indicate that gender differentiation played a significant role in the spread.

17. The rise of epicene *they* to social awareness coincides with the rise of prescriptivism in the late 18th century, as exemplified in the quote from Murray's grammar (1795) in the section entitled "Discussion of the role of gender."

18. The question of outstanding individuals in the data, accounting for large shares of variants in certain periods, provides an example of the "bad data" problem in historical sociolinguistics (Labov, 1994:11). Considering the three independent variables in this study, gender, region, and register, in terms of situation of use, Arabella Stuart is the only informant who accounts for a large proportion of the pronoun variants to the extent that it skews the overall tendencies observed in the data.

19. I decided to exclude region from the multivariate analysis because of unbalanced data distribution between the pronoun variants in the three regions.

20. The factors have been selected based on previous historical sociolinguistic research (Nevalainen & Raumolin-Brunberg, 2003:185–201); the objective is to account for the observed variability in pronouns and include one language-internal, one sociodemographic, and one register-related factor in the multivariate analysis (see Paolillo, 2001:42–43 for choosing the factors to be investigated in variationist sociolinguistics).

21. The most infrequent variant, *which/the which*, has been excluded to avoid knockout cells in the analysis.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

22. *They* is used mostly with *that* and only occasionally with the other relative pronoun types and has been excluded from the multivariate analysis to avoid knockout cells.

REFERENCES

Bailey, Guy, Maynor, Natalie, & Cukor-Avila, Patricia. (1989). Variation in subject-verb concord in Early Modern English. *Language Variation and Change* 1:258–300.

Baranowski, Maciej. (2002). Current usage of the epicene pronoun in written English. *Journal of Sociolinguistics* 6(3):378–397.

Baron, Dennis E. (1986). *Grammar and gender*. New Haven: Yale University Press.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan, & Finegan, Edward. (1999). *Longman grammar of spoken and written English*. London: Longman.

Bodine, Ann. (1990). Androcentrism in prescriptive grammar: Singular "they," sex-indefinite "he," and "he or she." In Deborah Cameron (ed.), *The feminist critique of language: A reader*. London: Routledge. 166–186.

Chambers, J. K. (2003). *Sociolinguistic theory: Linguistic variation and its social significance*. 2nd ed. Oxford: Blackwell.

Comrie, Bernard. (1986). Markedness, grammar, people and the world. In Fred R. Eckman, Edith A. Moravcsik, & Jessica R. Wirth (eds.), *Markedness*. New York: Plenum Press. 85–196.

Croft, William. (2003). *Typology and universals*. 2nd ed. Cambridge: Cambridge University Press.

Curzan, Anne. (2003). *Gender shifts in the history of English*. Cambridge: Cambridge University Press.

Dekeyser, Xavier. (1975). *Number and case relations in 19th century British English. A comparative study of grammar and usage*. Antwerp, Belgium: Nederlandsche Boekhandel.

_____. (1984). Relativizers in Early Modern English. In Jacek Fisiak (ed.), *Historical syntax*. Berlin: Mouton de Gruyter. 61–87.

Denison, David. (2005). Syntactic surprises in some English letters: The underlying progress of the language. *Plenary paper presented at Language History from Below: Linguistic Variation in the Germanic Languages 1700–2000*, University of Bristol, April 7, 2005.

Finegan, Edward, & Biber, Douglas. (1994). Register and social dialect variation: An integrated approach. In Douglas Biber & Edward Finegan (eds.), *Sociolinguistic perspectives in register*. Oxford: Oxford University Press. 315–347.

Gerner, Jürgen. (2000). Singular and plural anaphors of indefinite personal pronouns in spoken British English. In John M. Kirk (ed.), *Corpora galore. Analyses and techniques in describing English. Papers from the nineteenth International Conference on English Language Research on Computerised Corpora (ICAME 1998)*. Amsterdam: Rodopi. 93–114.

Goldberg, Adele E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7(5):219–224.

Halliday, M. A. K. (2001). *An introduction to functional grammar*. 3rd ed. (Revised by Christian M. I. M. Matthiessen). London: Arnold.

Hawkins, John A. (2004). *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hernández-Campoy, Juan Manuel, & Conde-Silvestre, Juan Camilo. (2005). Sociolinguistic and geolinguistic approaches to the historical diffusion of linguistic innovations: Incipient standardisation in Late Middle English. *International Journal of English Studies* 5(1):101–134.

Huddleston, Rodney, & Pullum, Geoffrey K. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Janda, Richard D. (2001). Beyond "pathways" and "unidirectionality": On the discontinuity of language transmission and the counterability of grammaticalization. *Language Sciences* 23:265–340.

Jespersen, Otto. (1949). *A Modern English grammar on historical principles. Part III: Syntax*. Vol. 2. London: George Allen & Unwin; Copenhagen: Ejnar Munksgaard.

Kirby, J. (1746) *A new English grammar*. Menston, UK: Scolar Press.

Kytö, Merja. (1996). *Manual to the diachronic part of the Helsinki Corpus of English Texts. Coding conventions and the lists of source texts*. 3rd ed. Helsinki: University of Helsinki.

Labov, William. (1972). *Sociolinguistic patterns*. Philadelphia: University of Philadelphia Press.

_____. (1994). *Principles of linguistic change. Internal factors*. Oxford: Blackwell.

_____. (2001). *Principles of linguistic change. Social factors*. Oxford: Blackwell.

Laitinen, Mikko. (2002). Singular HE and plural THEY in indefinite anaphora in written present-day English. *International Journal of Corpus Linguistics* 7(2):137–164.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

———. (2006). Expressing human indefiniteness in English: Typology and markedness of pronouns. In Terttu Nevalainen, Juhani Klemola, & Mikko Laitinen (eds.), *Types of variation: Diachronic, dialectal and typological interfaces.* Amsterdam: John Benjamins. 203–239.
———. (forthcoming). Epicene HE and THEY and the development of English indefinite expressions during the period 1500–1800. In Leena Kahlas-Tarkka, R. W. McConchie, & Matti Rissanen (eds.), *The change of meaning and the meaning of change*. Helsinki: Société Néophilologique.
Milroy, James. (2005). Variability, language change, and the history of English. *International Journal of English Studies* 5(1):1–11.
———. & Milroy, Lesley. (1993). Mechanisms of change in urban dialects: The role of class, social network and gender. *International Journal of Applied Linguistics* 3(1):57–77.
Milroy, Lesley. (2004). Off the shelf or under the counter? On the social dynamic of sound changes. *Plenary paper presented at Studies in the History of the English Language (SHEL) 3*, University of Michigan, May, 7 2004.
Murray, Lindley. (1795). *English grammar*. In Gerald Nelson (ed.), *Landmarks in English grammar. The eighteenth century.* (CD-Rom.) London: University College London.
Nevalainen, Terttu. (1987). Change from above: A morphosyntactic comparison of two Early Modern English editions of the *Book of common prayer*. In Leena Kahlas-Tarkka (ed.), *Neophilologica Fennica*. Helsinki: Société Néophilologique. 295–315.
———. (2000a). Gender differences in the evolution of standard English: Evidence from the *Corpus of Early English correspondence. Journal of English Linguistics* 28(1):38–59.
———. (2000b). Processes of supralocalization and the rise of standard English in the Early Modern period. In Ricardo Bermúdez-Otero, David Denison, Richard M Hogg, & C. B. McCully (eds.), *Generative theory and corpus studies: A dialogue from 10 ICEHL.* Berlin: Mouton de Gruyter. 329–372.
———. (2004). Three perspectives on grammaticalization: Lexico-grammar, corpora and historical sociolinguistics. In Hans Lindquist & Christian Mair (eds.), *Corpus approaches to grammaticalization in English.* Philadelphia: John Benjamins. 1–31.
———. (2006). Vernacular universals? The case of plural *was* in Early Modern English. In Terttu Nevalainen, Juhani Klemola, & Mikko Laitinen (eds.), *Types of variation: Diachronic, dialectal and typological interfaces.* Amsterdam: John Benjamins. 351–369.
———. & Raumolin-Brunberg, Helena. (1996). The corpus of Early English correspondence. In Terttu Nevalainen & Helena Raumolin-Brunberg (eds.), *Sociolinguistics and language history. Studies based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi. 39–54.
———. (2003). *Historical sociolinguistics: Language change in Tudor and Stuart England*. London: Longman.
Palander-Collin, Minna. (1999). *Grammaticalization and social embedding. I THINK and METHINKS in Middle and Early Modern English*. Helsinki: Société Néophilologique.
Paolillo, John C. (2001). *Analyzing linguistic variation: Statistical models and methods*. Stanford, CA: CSLI.
Puttenham, George. (1589). *The arte of English poesie*. Reprinted 1968. Menston, UK: Scolar Press.
Rand, David, & Sankoff, David. (1990). *GoldVarb version 2: A variable rule application for the Macintosh*. Montréal: Centre de recherches mathématiques, Université de Montréal.
Raumolin-Brunberg, Helena. (1994). The development of the compound pronouns in *-body* and *-one* in Early Modern English. In Dieter Kastovsky (ed.), *Studies in Early Modern English*. Berlin: Mouton de Gruyter. 301–324.
———. (2000). WHICH and THE WHICH in Late Middle English: Free variants? In Irma Taavitsainen, Terttu Nevalainen, Päivi Pahta & Matti Rissanen (eds.), *Placing Middle English in context*. Berlin: Mouton de Gruyter. 209–226.
———. (2002). Stable variation and historical linguistics. In Helena Raumolin-Brunberg, Minna Nevala, Arja Nurmi, & Matti Rissanen (eds.), *Variation past and present. VARIENG studies for Terttu Nevalainen.* Helsinki: Société Néophilologique. 101–116.
———. (2005). The diffusion of subject YOU: A case study in historical sociolinguistics. *Language Variation and Change* 17(1):55–73.
———, & Kahlas-Tarkka, Leena. (1997). Indefinite pronouns with singular human reference. In Matti Rissanen, Merja Kytö, & Kirsi Heikkonen (eds.), *Grammaticalization at work. Studies of long-term developments in English*. Berlin: Mouton de Gruyter. 17–86.
Raumolin-Brunberg, Helena, & Nevalainen, Terttu. (2007). Historical sociolinguistics: The *Corpus of Early English correspondence*. In J. C. Beal, Karen Corrigan, & Hermann Moisl (eds.), *Creating and digitizing language corpora.* Vol. 2: *Diachronic databases*. Houndmills, UK: Palgrave Macmillan.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press

Rydén, Mats. (1966). *Relative constructions in early sixteenth century English. With special reference to Sir Thomas Elyot*. Uppsala, Sweden: Almqvist & Wiksells.
Schilling-Estes, Natalie. (2004). Investigating stylistic variation. In J. K. Chambers, Peter Trudgill, & Natalie Schilling-Estes (eds.), *The handbook of language variation and change*. Oxford: Blackwell. 375–401.
Sundby, Bertil, Bjørge, Anne Kari, & Haugland, Kari E. (1991). *A dictionary of English normative grammar* 1700–1800. Amsterdam: John Benjamins.
Tagliamonte, Sali. (1998). *Was/were* variation across the generations: View from the city of York. *Language Variation and Change* 10:153–191.
Weinreich, Uriel, Labov, William, & Herzog, Marvin I. (1968). Empirical foundations for a theory of language change. In Winifred P. Lehmann & Yakov Malkiel (eds.), *Directions for historical linguistics*. Austin: University of Texas Press. 95–189.

https://doi.org/10.1017/S0954394508000045 Published online by Cambridge University Press