i sall synge in haboundance of gastly softne: /h/ insertion in Middle English — Methodology, data mining and some first interpretations[1]

Daniel Schreier, Milan Marković & Saša Petković
University of Zurich

This paper discusses some of the methodological challenges for a first corpus-based analysis of so-called /h/ insertion in English, a feature that has been widely observed yet not analysed empirically so far. We survey the existing literature and present what is known about historical variation and change, before describing our data-driven approach in the *Helsinki Corpus* and the *Corpus of Early English Correspondence* and presenting some first results on internal conditioning, restriction to word type, and overall frequency. We show that /h/ was inserted on English as well as French loanwords, nouns, adjectives, verbs, adverbs, and numerals, and that there was a positive match of identical lexical items in the two corpora (*able*, *am*, *it*, and *itself*), making this a historically robust feature.

Keywords: /h/ insertion; corpus analysis; historical variation and change; Helsinki Corpus; Corpus of Early English Correspondence

---

[1] This paper was delivered as a plenary talk at the 2017 Conference of the *Spanish Society for Medieval English Language and Literature* (SELIM), held at the University of Málaga on 21–23 September 2017. We wish to acknowledge the feedback and contributions from the audience present. A number of colleagues shared data and experience, commented on earlier drafts of the manuscript and pointed our interest to relevant issues, and our thanks go to: Rhona Alcorn, Jorge Luis Bueno-Alonso, Javier Calle-Martín, Juan Camilo Conde-Silvestre, Ray Hickey, Jesús Romero-Barranco, Olga Timofeeva, and Christine Wallis. Any remaining mistakes are our own.

Daniel Schreier, Milan Marković & Saša Petković, *Selim* 23 (2018): 115–136.
ISSN 1132-631X

1. Introduction

It is perhaps not exaggerated to say that /h/ insertion (in words that receive initial stress and begin with a vowel, *island*, *abundance*, *it*, etc.) is one of the least understood variables in the history of the English language. It is reported in overviews of English historical phonology (Lass 1992, Milroy 1992) and can be traced back to Early Middle English. Milroy (1992) has shown that the *Norfolk Gilds* (late 14th century) or the *Paston Letters* (late 15th century) exhibit variable use of <h> spellings, namely both absence (in *alpenie* 'halfpenny') and un-etymological insertion (in *hoke lewes* 'oak leaves'). It is frequently used as a stylistic device in Charles Dickens' renderings of working-class London English ("gas microscopes with hextra power", in the *Pickwick Papers*), and historical corpora (such as the *Linguistic Atlas of Early Middle English*, LAEME) provide evidence also. Notwithstanding, inserted /h/ has very near to extinction in Present-day British English (surviving in cases of occasional hypercorrection, such as in the letter <h>, pronounced /heitʃ/) but has been maintained in several post-colonial English varieties around the world (such as Tristan da Cunha English; see Schreier forthc.). All in all, /h/ insertion is well-known and mentioned in the literature, yet always on a descriptive level, and quantitative analyses are lacking. So far, it has not yet been studied from a standpoint of variationist sociolinguistics or (with the exception of Milroy 1992) historical sociolinguistics.

This paper goes some way of redressing this imbalance and reports on a first attempt to reconstruct historical variation in /h/ insertion in two corpora, namely the *Helsinki Corpus* (HC) and the *Corpus of Early English Correspondence Sampler* (CEECS). A prime concern of our focus is methodology and we will detail our data mining procedures in detail here, in the hope that /h/ insertion is studied in other historical corpora as well. We will also provide some information on the lexical dimension of /h/ variation and compare the two corpora with the aim of uncovering parallel patterns of /h/ insertion in earlier forms of English. We begin with a general discussion of /h/ insertion in English and present what is known about this variable, based on existing sources. Following the general discussion, we detail the methodology developed for our research project, describe the two corpora used (HC, CEECS), present our findings, and end with a conclusion and an assessment of future research on the variable.

2. /h/ insertion in English: What we know so far

One problem with assessing the historical validity of /h/ insertion is that the comments are sporadic, non-representative, and selective. The high degree of awareness of this variable somewhat contrasts with the sources for which it is reported, and over the last two centuries, commentators (such as Jespersen 1949) focus on very few sources where it is mentioned. To give an example: one of the most widely used 19th-century sources is the direct speech of working-class London English as reported in Charles Dickens' novels. Indeed, /h/ variation is a rather prominent feature here:

Adam Bede:
- Mr Casson, a butler, talks to a traveler. "He'll be comin' of **hage** this 'ay-'arvest"

Nicholas Nickleby:
- "This is the **hend**, is it? continued Miss Squeers, who being excited aspirated her h's strongly!" (p. 518)

Great Expectations:
- Pip writes to Joe: "mI deEr JO i opE U r krWitE wEll i opE i shAl soN B **haBelL** 4 2 teeDge U JO" (i, p. 98)

The Pickwick Papers:
- "If they wos a pair o' patent double million magnifyin' gas microscopes of **hextra** power, p'raps I might be able to see through a flight o' stairs and a deal door" (Chapter 34)

Previous studies on variation and change of English dialect features in dialect contact scenarios overseas (such as the /v–w/ merger, Trudgill et al. 2003) have readily used Dickens as a resource and taken dialect reports of this kind as *prima facie* evidence of working-class London English (assuming that it was brought to overseas areas). This is not unproblematic, however. From a standpoint of English historical linguistics, there is no warranty that the direct speech reported in a source is an adequate reflection of how Londoners used to speak at the time (this, of course, goes for all historical sources, Chaucer, Shakespeare, etc.), let alone that London English served as a donor variety, which is a considerable conundrum. Playwrights have always taken liberties

when creating their characters; reported dialogues at odds with adequate (or authentic) renderings of speech and may in fact be nothing more than literary artefacts, not to be taken at face value. Mugglestone reminds historical linguists that

> Dickens's lower-class characters are, correspondingly, often depicted as being seemingly incapable of pronouncing *hand* other than as *and, hungry* other than as *ungry*, or, conversely, *under* other than as *hunder*. Such systematic patterns are, in real terms, fictions just as much as the characters themselves. (Mugglestone 1995: 137)

The same observation, albeit stronger, is found in Jespersen:

> Many novelists would have us believe, that people who drop their aspirates place false aspirates before every vowel that should have no [h]; such systematic perversion is not, however, in human nature. But they sometimes inadvertently put a [h] between two vowels (rarely after a consonant), especially when the word is to receive extra emphasis, and of course, without any regard to whether the word 'ought to' have [h] or not. The observer [...] is struck with the instances of disagreement, deducing from them the impression of a systematic perversion ('Am an' heggs'). (Jespersen 1949: 378)

Both Mugglestone and Jespersen call for a cautious interpretation of historical materials (of which Dickens' novels are just one source, of course, even though a rather widely quoted one). However, this is not to refute their validity. Jespersen himself discusses /h/ variation in the history of English (1949: 378–390) and quotes other sources, such as James Elphinston's (1787) *Principles of English Grammar*: "E 1787 (vol 2.254 ff.) complains of exactly the same errors in this respect as are met with nowadays: *ils, ouzes, eariing the owls in the hevening, orse, art, arm,* etc."; John Walker's (1791) *Critical Pronouncing Dictionary*: "W 1791 speaks of the 'fault of the Londoners: not sounding *h* where it ought to be sounded, and inversely'."; and Thomas Batchelor's (1809) *Orthoëpical Analysis*: "B 1809 p. 29 says: 'the aspirate *h* [...] is often used improperly, and is as frequently omitted where it should be used. *Give my orse some hoats* has been given as an example of these opposite errors from the Cockney dialect'."

Additional evidence comes from regional dialectology. The *Survey of English Dialects* (SED, Luick 1964) reports fieldwork data of so-called NORM (Non-mobile Older Rural Male) speakers throughout England. Fieldworkers

were instructed to elicit data via questionnaires and to transcribe the responses. With regard to /h/ insertion, Luick concludes that

> Anm. 1. Dass *h* nicht gänzlich geschwunden ist (so Wright, Dial. Gram. 254), ergibt sich aus den Aufzeichnungen bei Ellis (vgl. Güning 3ff) und neueren Einzeluntersuchungen. Danach würde heute in Kendal im nördlichen Cumberland und in Suffolk *h* zumeist richtig gesprochen (Hirst 12, 111, Brilioth 93, Kökeritz 106) und ware in Penrith und Süd-Durham bei manchen Sprechern oder in gewissen Orten regelmässig vorhanden (Reaney 133, Orton 6, 141). *Für in der Emphase falsch angefügtes h bieten Beispiele: Hackness (Cowling 101), Adlington (Hargreaves 73), Pewsey (Kjederqvist 113), West-Somerset (Kruisinga 93), Stokesley (Klein 76), Oldham (Schilling 107)*. (Luick 1964: 1093, emphasis added)
> 'Obs. 1: The fact that *h* has not entirely disappeared (cf. Wright, Dial. Gram. 254), can be seen in the notes of Ellis (cf. Güning 3ff) and some more recent case studies. Accordingly, in Kendal in northern Cumberland and in Suffolk *h* was pronounced correctly most of the time (Hirst 12, 111, Brilioth 93, Kökeritz 106) and it was regularly found in some speakers and certain localities in Penrith and South-Durham (Reaney 133, Orton 6, 141). Examples for h that are added incorrectly in emphatic positions: Hackness (Cowling 101), Adlington (Hargreaves 73), Pewsey (Kjederqvist 113), West-Somerset (Kruisinga 93), Stokesley (Klein 76), Oldham (Schilling 107).'

This suggests that /h/ insertion was well and alive in the mid-20th century (at least in older speakers of traditional dialects) and that it was rather widespread regionally. This is confirmed in late Middle English, where Milroy's (1992) analysis of the *Linguistic Atlas of Late Medieval English* (LALME) provides evidence in texts from the East Midlands, East Anglia and the South (in *c.* 1190–1320, the feature was attested in a region from Lincolnshire and Norfolk to the southern counties but "the instability seems to be greatest in the East Midlands", Milroy 1992: 140). Milroy also finds that /h/ insertion was attested in the *Norfolk Gilds* (late 14th century) and the *Paston Letters* (late 15th century; *alpenie* 'halfpenny' and *hoke lewes* 'oak leaves').

Alcorn (personal communication, August 2016) reports even earlier evidence of /h/ insertion in LAEME, where a search surfaced

> multiple examples in multiple LAEME texts from multiple counties, including Gloucestershire, Kent, Lincolnshire, Norfolk, Suffolk, and Worcestershire.
> Adj[ective] examples include: *holde* 'old', *hiuel* 'evil, *hunkinde* 'unkind'.
> Adv[erb] examples include: *heuere* 'ever', *hinne* 'in', *hout* 'out'.

> Conj[unction] examples include: *hif* 'if', *her* 'ere', *has* 'as'.
> Prep[osition] examples include: *huntil* 'until', *hafter* 'after', *hat* 'at'.

Alcorn informs us that /h/ insertion is widespread, both externally in terms of region, and internally in all parts of speech (which is noteworthy, as most of the later examples come from nouns; see below).

Like other sources, Milroy emphasises that there is a relationship between /h/ dropping and /h/ insertion, though he is more sociolinguistically sensitive to the variable nature of the two processes than scholars such as Jespersen are. Milroy (1992) finds that /h/ variation, e.g. *ate* for 'hate' and *om* for 'home', *halle* for 'alle' and *his* for 'is', has a complex sociolinguistic history:

> Many Early M[iddle] E[nglish] sources exhibit variable use of the letter *h* in syllable-initial positions before vowels (that is, in such words as *hate*, *hopper*). Sometimes it is omitted where it is historically expected to be present, and sometimes it is added where it is not expected. (Milroy 1992: 140)

Milroy offers a historical-sociolinguistic explanation for /h/ variation, namely that there was linguistic insecurity as to when to pronounce /h/. One of the most openly stigmatised sociolinguistic variables in the history of English is /h/ dropping, and Milroy suggests that instability in /h/ usage leads to hyper-correction and insertion of /h/ where it is not etymological. Speakers, in other words, over-compensate in order to avoid using stigmatised innovations —which would explain why one and the same speaker would have both dropping and insertion at the same time (which, of course, makes it very noticeable). A similar explanation was offered by Christine Wallis (unpublished ms.), who studied the *Monasteriales Indicia*, a short description of sign language in a monastery during times of silence (commonly used items to do with liturgy, eating and drinking, clothing, and people in the abbey) and was dated to Canterbury, *c.* mid-11th century. In this text, Wallis noted both deletion (*is* 'his', *abban* 'to have') and insertion (*hunlocan* 'unlock', *his* 'is', *halban* 'alb') and concluded that "the scribe is inconsistent in his use of <h> [...] 'overcompensating' in speech for h-dropping. Therefore, the text would reflect the writer's speech. Alternatively, it could be a written feature reflecting only h-lessness in speech."

Of course, one problem with such an approach is that one would expect the variable to be above the sociolinguistic radar. If stigmatised features are 'corrected', then they would have to be noticeable and subject to open

comments and discourse, particularly in the usage guides and orthoepic treatises of the 18th century. However, there are "few comments on '[h]-dropping' (and [h] insertion) before 1800. The very few comments also refer to Cockney, or London English" (Milroy 1992: 138). This is also reflected in Jespersen's discussion (see above).

Linguistic insecurity is questioned by others as well:

> Until the beginning of the sixteenth century, there was no evidence of association between h-dropping and social or educational status, but the attitudes began to shift in the seventeenth century, and by the eighteenth century [h]-lessness was stigmatized in both native and borrowed words. (Minkova 2013: 107; cf. Mugglestone 2006)

> Still, compared with the situation in the following century, any adverse sociophonetic consequences of [h]-dropping and adding are, if anything, relatively unstressed by late-eighteenth-century observers. (Jones 2006: 257)

To substantiate these claims, we checked a selection of grammar books and usage guides published in the second half of the 18th century (see Table 1), and it is striking that not a single one of them listed /h/ insertion.

This, in our view, is evidence that the feature was not in the public eye and underneath the sociolinguistic radar (discussion in Schreier forthc.), making Milroy's explanation difficult to uphold. The three sources quoted by Jespersen (1949) and reproduced in other publications seem to be the only ones that make explicit reference to the feature.

To sum up this very short overview of what we know about the history of /h/ insertion in British English: the evidence we have at present is anecdotal and sporadic at best. /h/ deletion and insertion are commonly discussed hand in hand and considered as a manifestation of /h/ variation in general. The earliest reports we could locate come from the mid-11th century (*Monasteriales Indicia*), then inserted /h/ is reported throughout the Middle English and Early Modern English periods (*Norfolk Gilds*, late 14th century; *Paston Letters*, late 15th century; and also in historical dialect corpora: LAEME, LALME).

Table 1. 18th-/19th-century usage guides checked for comments on /h/ insertion

Year	Author	Title
1726	N. Bailey	*An Introduction to the English Tongue: Being a Spelling Book*
1750	S. Hammond	*A New Introduction to Learning; or, A Sure Guide to the English Pronunciation and Orthography*
1762	G. Sharp	*A Short Treatise on the English Tongue: Being an Attempt to Render the Reading and Pronunciation of the Same More Easy to Foreigners*
1764	A. Johnston	*Pronouncing and Spelling Dictionary*
1773	W. Kenrick	*A New Dictionary of the English Language*
1781	T. Sheridan	*A Rhetorical Grammar of the English Language. Calculated Solely for the Purposes of Teaching Propriety of Pronunciation, and Justness of Delivery, in that Tongue, by the Organs of Speech*
1784	R. Nares	*Elements of Orthoepy*
1786	W. Scott	*A New Spelling, Pronouncing, and Explanatory Dictionary of the English Language*
1791	J. Walker	*A Critical Pronouncing Dictionary and Expositor of the English Language*
1792	W. Fogg	*Elementa Anglicana; or, The Principles of English Grammar Displayed and Exemplified*
1793	W. Perry	*The Royal Standard English Dictionary*
1795	W. Smith	*An Attempt to Render the Pronunciation of the English Language More Easy to Foreigners*
1799	J. Adams	*The Pronunciation of the English Language Vindicated from Imputed Anomaly & Caprice*

As for insights from regional dialectology and historical sociolinguistics, Ellis (1889: 1) reports the feature in "uneducated people, speaking an inherited language, in all parts of Great Britain where English is the ordinary medium of communication between peasant and peasant", and it is found in a broad range of diverse areas, including Yorkshire, Lancashire, Wiltshire, and Somerset. Following Milroy, there was a regional concentration in the Southeast and the Midlands, and by the late 19th century it had developed a strong association with working-class London English (Cockney). This no doubt fuelled its prominent usage in the literature (as in Dickens' novels), where it was portrayed (perhaps even stereotyped) as an East London feature. Later on, it was all but lost from British English; whereas dialect atlases such as the SED still report it as common (at least in the speech of older non-mobile men in

the countryside), it became obsolescent from the 1950s onwards and is now very near extinction (or already extinct).

As quantitative evidence is lacking, the present study is a first attempt to develop a methodology that allows us to retrace historical variation with regard to /h/ variation. We would like to present some first historical data to explore that variable nature of /h/ and also get a better understanding of its frequency and internal conditioning in Middle and Early Modern English. With this aim, we selected two corpora, the HC and the CEECS, and searched for cases of /h/ insertion. In a next step, we detail the methodology we employed before we go on to present some first findings.

3. /h/ insertion in Middle English: Methodology and some first findings

While certainly not the largest of the historical corpora available today, the *Helsinki Corpus* and the *Corpus of Early English Correspondence* (CEEC) are particularly well-suited for diachronic investigation of /h/ insertion, since they were both designed and compiled along principles of historical sociolinguistics. These two collections of written language represent a proverbial 'window into the past', as they provide a very good opportunity to extend the analysis of /h/ insertion into the domain of English historical linguistics and to assess patterns of variation and change.

The HC (Rissanen et al. 1991) is comprised of two major components: a diachronic and a dialectal one. The diachronic component of the HC was released in 1991 and includes a selection of texts that span the period from around 730 to 1710 (thus almost one thousand years). These texts are organised into three large sections, which correspond to the Old, Middle and Early Modern English periods. In total, the HC includes *c.* 450 texts and 1,572,800 words. The Old English section contains 413,300 words, the Middle English section 608,600 words, and the Early Modern English section 551,000 words. The representativeness of the language included in the corpus is described by Kytö, who notes:

> [T]he selectional criteria adopted for including a text [...] reflect the principles of socio-historical variation analysis [...] Periodization has been of primary importance [...] but attention has also been paid to geographical dialect, type and register of writing (text type, relationship to spoken language, setting on

> formal-informal axis) and sociolinguistic variation (different author-related parameters such as gender, age, social rank). (Kytö 1996)

However, even though the corpus covers nearly an entire millennium in the history of the English language, only its Middle and Early Modern English components, which together span the period from 1150 to 1710, were included for the present investigation. Its Old English component, on the other hand, which covers the period from 730 to 1150, was not taken into consideration.

In a similar vein, the original purpose of the CEEC (Nevalainen et al. 1998) was to examine the ways in which a methodology employed for modern sociolinguistic research can be applied to historical data. This was made possible by including "an extensive database containing background information about letter writers" (Nevalainen et al. 1998) into the corpus metadata. The original version of the CEEC was released in 1998 and covers the time period between 1410 and 1681. It is comprised of around 5,961 letters and amounts to more than 2.5 million words. The CEECS, its publicly-available version, was released in the same year (Nevalainen et al. 1998), and it is this version of the corpus that we used for the purposes of the present paper. The sampler itself covers the period between 1418 and 1680 and contains 1,147 letters from 194 individual writers, amounting to around 450,000 words in total. It is worth mentioning that this sampler represents a "fairly accurate small-scale copy of the full CEEC" as it provides very "similar results for many linguistic phenomena" (Nevalainen et al. 1998).

As mentioned previously, so far very few attempts have been made to account for /h/ insertion in quantitative terms. As a result, we set out to devise a universal, reliable and robust set of criteria, which would allow us to accurately track the usage of this sociolinguistic variable over long periods of time (and that would be general enough so that it could be applied to other corpora as well). Our methodological procedure, which is essentially an extension of the methodology first employed to investigate /h/ insertion in Tristan da Cunha English (see Schreier forthc.), is described in detail in the passages below, in the hope that it will help advance further quantitative analyses of /h/ variation in the domain of English historical linguistics.

During the initial stage of the analysis, it was necessary to identify and extract all the instances of /h/ insertion from the two corpora in order to get some first understanding of its use. This presented us with a first major obstacle. Specifically, while this procedure would have been a relatively

straightforward task when dealing with a smaller, specialised corpus, where the feature investigated is usually marked via some type of notation in order to facilitate its searchability and recall, the same does not apply when dealing with large, general-purpose corpora, such as the HC and the CEEC, which strive for universality and where the level of non-standardness in terms of notation is reduced to a bare minimum. To make matters more complicated, upon running some of our preliminary searches, we also encountered a number of ambiguous word forms, a variety of different spelling realisations for the same form as well as forms carrying inflectional suffixes which have long disappeared from usage in the English language. This is illustrated in Table 2, which shows the complete list of word forms attested in the HC data for the lexical units ASK and ABIDE. This, as a consequence, rendered looking for any underlying patterns, which could facilitate the retrieval of /h/ tokens, a virtually impossible task.

Table 2. List of word forms for ASK and ABIDE attested in the *Helsinki Corpus of English Texts*

Lexeme	**Word form**
ASK	ask, asks, asking, asked, ask'd, askande, askede, askeden, askeing, askeinge, asken, askes, askest, asket, asketh, askeyd, askeþ, askid, askis, askist, askit, askith, askt, askte, askus, askyd, askyde, askyn, askyng, askynge, askynges, askys, askyth, axing, akseþ, aksy, axande, axe, axed, axede, axeden, axeing, axeinge, axes, axelyne, axest, axet, axen, axeth, axeyd, axeþ, axid, axis, axist, axit, axith, axt, axte, axus, axyd, axyde, axyn, axyng, axynge, axynges, axys, axyth, hasked, haske
ABIDE	abide, abiding, abidað, abidæn, abidan, abidand, abidas, abiddan, abide, abideð, abiden, abideth, abideþ, abiding, abidiþ, abidon, abidyng, abidynge, abyd, abydand, abyde, abyden, abydest, abydeth, abydith, abydyn, abydyng, abydynge, abydyth, abydythe, habyde, habydynge

Hence, in order to substantially limit the scope of our investigation and bring it to a somewhat more manageable level, we first extracted every single word form which starts with the letter *h*. This procedure was carried out using corpus query language; however, the same result can also be achieved with the use of regular expression syntax. Unsurprisingly, running a search of such general scope resulted in an extremely high number of hits in both the HC and the CEECS (N=109,920 and N=27,335, respectively). The concordanced results were then exported as a comma separated file (.csv) and imported into an Excel spreadsheet for further analysis.

During the next stage, each of the hits was manually inspected for the presence of a potential /h/ prefix. Only those cases where /h/ insertion was positively attested were retained in the tabulations, while the rest of the data was discarded. As a result, we managed to identify a total of 1,661 instances of /h/ insertion in the HC and 520 instances of /h/ insertion in the CEECS.

Nevertheless, upon closer examination of the results, we noticed two distinct types of tokens with an /h/ prefix in the HC data. The first type was comprised of word forms such as *habilite* (ability), *hable* (able), *haboundance* (abundance), *haske* (ask), *hevere* (every), *hit* (it), etc. These instances of /h/ insertion 'proper' were retained in the analysis. On the other hand, we also noticed a number of tokens occurring in the Middle English component of the HC (1150–1350, specifically), which were transformed into *wh-* words over time. This type consisted of word forms such as *huanne* (when), *huere* (where), *huet* (what), *huich* (which), *hwile* (while), etc. We treated these instances as members of the 'Don't Count' class and removed them entirely from the scope of our investigation. With the exclusion of these word forms, the total number of /h/ tokens in the HC data was reduced from the original 1,661 to a somewhat lower value of 1,309.

In the next step, we categorised each of the /h/ tokens obtained from the two corpora according to their corresponding lexical values. Upon the completion of this process, we managed to identify nineteen unique lexical items in each of the two corpora. These lexemes were then used to find the corresponding word forms where /h/ insertion could but did not occur. This step was crucial in order to ensure that our method remains in compliance with Labov's (1982:30) 'principle of accountability'.

Specifically, the method which was employed to extract the /h/ variation data was the following. First, the entire inflectional paradigm was reconstructed for each of the lexemes by adding all the possible suffixes, depending on its part of speech membership (e.g. ABILITY: *ability*, *abilities*; ALLOW: *allow*, *allows*, *allowing*, *allowed*; ABLE: *able*, *abler*, *ablest*). Next, /h/ prefixed equivalents were created for each of the word forms obtained from the above procedure. Finally, all the prefixed and non-prefixed word forms were combined to create a list in order to extract both realisations of the dependent variable.

However, while this procedure works very well with the data from the Modern English period (see Schreier forthc.), it alone is not sufficient when earlier forms of English are taken into account, due to the variability in form and spelling mentioned earlier. Hence, in order for our procedure to be

considered methodologically robust, the word list needed to be expanded with all the archaic realisations of the lexemes as well. These were identified with the help of two major Middle English resources: LAEME (Laing 2013) and the *Middle English Dictionary* (MED, Lewis 1959). In the last step, all the archaic word forms attested in our subsequent searches were included in the final list. An illustrative sample of such a list, which was used to extract the /h/ variation data from the HC, is presented in Table 3.

Table 3. Sample of the list used for extraction of Ø and /h/ tokens in the *Helsinki Corpus of English Texts*

Lexeme	**Ø**	**/h/**
ABILITY (N)	ability	hability
	abilities	habilities
	abilite	habilite
	abilitie	habilitie
	abilitys	habilitys
ALLOW (V)	allow	hallow
	allows	hallows
	allowing	hallowing
	allowed	hallowed
	allow'd	hallow'd
	allowd	hallowd
	allowe	hallowe
	allowes	hallowes
	alloweþ	halloweþ
ABLE (Adj)	able	hable
	abler	habler
	ablest	hablest
	abel	habel
	abil	habil
	abill	habill
	abyl	habyl
	abyll	habyll
ABUNDANTLY (Adv)	abundantly	habundantly
	abundauntly	habundauntly
	aboundantlie	haboundantlie
	aboundantly	haboundantly

3.1 The *Helsinki Corpus of English Texts*

As shown in Table 4, the initial results of the analysis on the HC show quantitative evidence of /h/ insertion in the Middle English varieties. Even though only at 3.8 per cent, the overall insertion rate in texts ranging from 1150 to 1710 stems from a relatively high token count of 1,309 (34,507 in total). Considering that these tokens are extracted from the records of written language when only the educated and the nobility were literate, a substantial number of /h/ prefixed instances such as this warrants further investigation.

Table 4. *Helsinki Corpus of English Texts*: Overall /h/ insertion rate

Ø (N)	**/h/ (N)**	**Total (N)**	**/h/ (%)**
33,198	1,309	34,507	3.8

Categorised into periods ranging from seventy to one hundred years in length, the analysis of /h/ insertion rates reveals a drastic decrease and eventual loss of this feature over time. Texts from the earliest period between 1150 and 1250 show a very high /h/ insertion rate amounting to 41.1 per cent (527 out of 1282), which very noticeably contrasts with the results from the other periods. Even though the periods from 1250 to 1710 are above the overall insertion rate, they exhibit very low insertion rates in comparison. Specifically, the periods from 1250 to 1350 and 1420 to 1500 display /h/ insertion rates of 6.6 per cent and 7.4 per cent, respectively. The remaining periods show a substantial decrease in insertion rates, dropping to 1.7 per cent between 1350 and 1420, then to 0.9 per cent between 1500 and 1570, and finally to 0.3 per cent between 1570 and 1640. There are no occurrences of /h/ insertion in the text records from the period between 1640 and 1710. This is an indication that the feature was lost, or at the very least, that there was a considerable decline of /h/ insertion in written language, as seen in Figure 1.

On the lexical level, the HC results show a varied distribution of /h/ insertion rates. These findings are presented in Table 5. Due to their low token counts, the lexemes *ability, abound,* and *abundantly* will not be discussed further. The adjective *abundant* is one of the exceptions since it presents a very interesting case, as does the reflexive pronoun *itself* in connection to the pronoun *it*. Further investigation shows that eight out of sixteen remaining lexemes display low insertion rates. In the case of *abide*, only one example of /h/ insertion out of 171 total tokens was discovered, which points to a 0.6 per cent insertion rate. Similarly, the lexeme *in* appears only twice with the /h/

prefix out of 18,357 total tokens. The lexemes *old* (N=563), *up* (N=519), and *upon* (N=755) have high total token counts but the number of their prefixed tokens is low, so that their insertion rates range from 0.3 per cent to 0.6 per cent. These appear to be actual instances of /h/ insertion; nevertheless, they should be examined further on a larger data sample before drawing any conclusions here. The lexeme *ask*, on the other hand, is only minimally more reliable with four /h/ prefixed tokens out of 536, producing a 0.7 per cent insertion rate, which is identical to the lexeme *every* with six out of 902 total

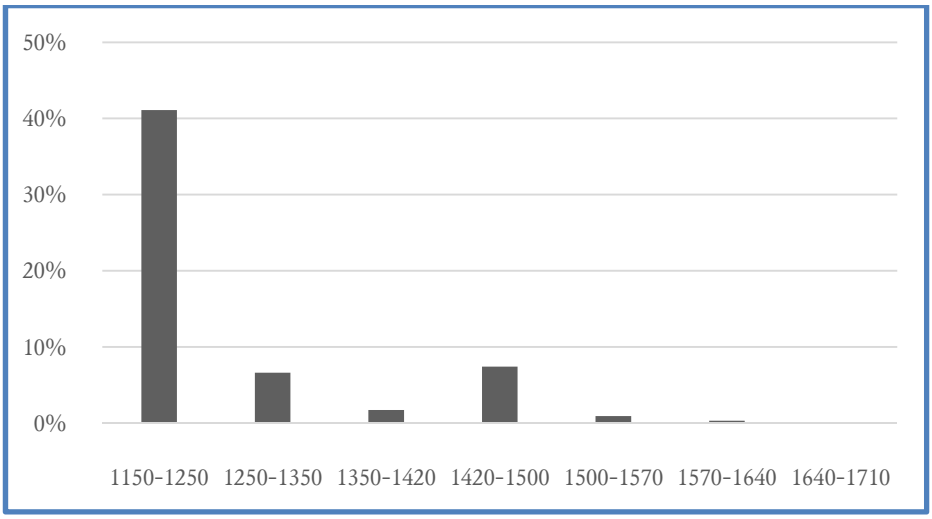

Figure 1. *Helsinki Corpus of English Texts*: /h/ insertion rates across different time periods

tokens. Likewise, the possessive pronoun *our,* displays a conclusively low insertion rate of 0.7 per cent with eleven /h/ inserted tokens out of 1,623 in total.

Moving on to lexemes with higher insertion rates, the verb *allow* with thirty-five total tokens appears only twice in the examined data sample (insertion rate 5.7%). Similarly, with a total token count of fifty-two, the preposition/conjunction *until* also appears only twice in the corpus yielding a 3.8 per cent insertion rate. The lexeme *able* appears prefixed by /h/ four times out of the total 153 tokens with an insertion rate of 2.6 per cent. Nevertheless, these lexemes require further research on a larger data sample in order to

verify the reliability of these results. The first person singular verb *am* on the other hand provides a much clearer picture of its insertion rate (twenty-one out of 920, or 2.3 per cent). Furthermore, the remaining lexemes are all examples of substantially higher /h/ insertion rates, which warrants closer examination. The adjective *abundant,* although low in total token count, appears almost categorically with the /h/ prefix in the data sample. With

Table 5. *Helsinki Corpus of English Texts*: /h/ insertion rates per lexeme

Lexeme	**/h/ (N)**	**Ø (N)**	**Total (N)**	**/h/ (%)**
ABIDE	1	170	171	0.6
ABILITY	4	7	11	36.4
ABLE	4	149	153	2.6
ABOUND	1	14	15	6.7
ABUNDANCE	12	22	34	35.3
ABUNDANT	12	1	13	92.3
ABUNDANTLY	1	13	14	7.1
ALLOW	2	33	35	5.7
AM	21	899	920	2.3
ASK	4	532	536	0.7
EVERY	6	896	902	0.7
IN	2	18,355	18,357	0.01
IT	1,214	8,601	9,815	12.4
ITSELF	4	15	19	21.1
OLD	3	560	563	0.5
OUR	11	1,612	1,623	0.7
UNTIL	2	50	52	3.8
UP	3	516	519	0.6
UPON	2	753	755	0.3
Total	**1,309**	**33,198**	**34,507**	**3.8**

twelve insertion examples out of the total thirteen tokens, its insertion rate amounts to 92.3 per cent. Correspondingly, perhaps in line with the previous result, the noun *abundance,* with twelve examples of /h/ insertion out of thirty-four total tokens, produces an insertion rate of 35.3 per cent.

The most interesting finding in the data sample, however, is the personal pronoun *it*. With 1,214 occurrences of /h/ insertion out of the total 9,815 tokens, this lexeme has an insertion rate of 12.4 per cent. Its insertion rate is not as high as that of the previous two lexemes, nevertheless, with such a large number of /h/ inserted occurrences (92.7% of all the /h/ tokens in the HC), the pronoun *it* is a perfect candidate for further study of the underlying linguistic constraints which condition /h/ insertion. Similarly, although the reflexive pronoun *itself* has a low total token count (N=19), with four occurrences of /h/ insertion it produces the insertion rate of 21.1 per cent and merits additional research on a larger data sample, if only in connection with the pronoun *it*.

3.2 The *Corpus of Early English Correspondence Sampler*

Although much smaller in size, the CEECS shows an overall insertion rate of 2.7 per cent, which is very close to the HC rate of 3.8 per cent (see Table 6). Out of 19036 total tokens, 520 have the inserted /h/ and, considering that the CEEC sampler version is less than half the size of the selected data sample from the HC, these results promise to be reliable, comparable and generalisable.

Table 6. *Corpus of Early English Correspondence Sampler*: Overall /h/ insertion rate

Ø (N)	**/h/ (N)**	**Total (N)**	**/h/ (%)**
18,516	520	19,036	2.7

Investigating the insertion rates in the CEECS on the lexemic level reveals the interesting fact that both corpora produce the total of nineteen unique lexemes prefixed with /h/ in written records (see Table 7). Of those nineteen only four appear in both data samples: *able, am, it*, *itself*. Again, out of the nineteen lexemes three (*abandon*, *ache*, *ale*) will not be discussed due to their low total token numbers. Out of the remaining sixteen, seven lexemes (*answer*, *as*, *at*, *one*, *order*, *over*, *us*) have high total token counts but appear only once or twice with the /h/ prefix in the corpus so they will also be disregarded as

possible *hapax legomena*. The reflexive pronoun *itself* will only be mentioned in connection to the lexeme *it* due to its very low number of occurrences (N=10). First of all, the lexeme *all* produces the insertion rate of 0.2 per cent appearing four times with the /h/ prefix out of the total 2,314 tokens, as does the lexeme *any*, occurring three times out of 1,322. The adjective/pronoun *other* appears four times out of 832 tokens and produces an insertion rate of 0.5 per cent.

Table 7. *Corpus of Early English Correspondence Sampler*: /h/ insertion rates per lexeme

Lexeme	**/h/ (N)**	**Ø (N)**	**Total**	**/h/ (%)**
ABANDON	1	2	3	33.3
ABLE	25	155	180	13.9
ACHE	1	1	2	50.0
ALE	2	5	7	28.6
ALL	4	2,310	2,314	0.2
AM	14	1,062	1,076	1.3
ANSWER	1	336	337	0.3
ANY	3	1,319	1,322	0.2
ARMS	2	26	28	7.1
AS	1	4,652	4,653	0.02
AT	2	2,326	2,328	0.1
IT	450	3,797	4,247	10.6
ITSELF	1	9	10	10.0
ONE	2	707	709	0.3
ORDER	2	169	171	1.2
OTHER	4	828	832	0.5
OVER	1	238	239	0.4
UNCLE	2	22	24	8.3
US	2	552	554	0.4
Total	**520**	**18,516**	**19,036**	**2.7**

Two of the lexemes, *arms* occurring with the /h/ prefix two out of twenty-eight times, and *uncle*, two out of twenty-four times, produce higher insertion rates (7.1% and 8.3%, respectively). However, these lexemes require further research on a larger data sample in order establish how robust and frequent this insertion process occurs. Regarding the lexemes that appear in both data samples, the verb form *am* shows an insertion rate of 1.3 per cent (occurring fourteen out of 1,076 times), one per cent less frequently than in the HC (2.3%). The adjective *able*, on the other hand, produces very high insertion rate of 13.9 per cent (appearing twenty-five out of 180 times), which is considerably higher than the 2.3 per cent rate from the HC results.

Although the two corpora show some similarities as well as differences in both lexeme choice and insertion rates, the most interesting finding in both data samples is the case of the pronoun *it* (see Table 8). In the CEECS, this lexeme appears prefixed by /h/ 450 times out of the total 4,247 tokens, yielding an insertion rate of 10.6 per cent. The pronoun accounts 86.5 per cent of all /h/ prefixed occurrences in the sampler corpus. In the HC, the pronoun *it* has a 12.4 per cent insertion rate (occurring 1,214 out of 9,815 times). Even though the occurrences, token numbers and the size of the examined data samples are 50-60 per cent smaller for the CEECS compared to the HC, the results are similar, and the extracted word form variants are nearly identical. The personal pronoun *it* is definitely a promising candidate for further and more granular research as it produces consistent results across two very different data samples. Moreover, the reflexive pronoun *itself* may also provide some insight if examined in more detail on a larger corpus.

Table 8. *it*: /h/ insertion rates across the corpora

Corpus	**/h/ (N)**	**Ø (N)**	**Total**	**/h/ (%)**
CEECS	450	3,797	4,247	10.6
HC	1,214	8,601	9,815	12.4

4. Conclusion: What we know and need to know

Our analysis of inserted /h/ in two historical corpora has provided some first evidence of historical variation and change. We found that /h/ insertion operated on a limited set of lexical items only. The comparison of the HC and the CEEC suggests that there are some parallels. First of all, the overall

insertion rate is low yet similar in the two corpora, second, /h/ insertion is reported in identical lexical items (which in our view attests to the robustness of the process), and third, there are similarities regarding the diachronic manifestation, particularly when it comes to obsolescence. For one, based on the criteria selected here, we show that /h/ was inserted on English as well as French loanwords, nouns, adjectives, verbs, adverbs, and numerals, and that there was a positive match of identical lexical items in the two corpora (*able*, *am*, *it*, *itself*). However, the overall token count was low (in fact even lower had we disregarded *it*, which deserves special mention) so we could not carry out a detailed variationist analysis and can (for the moment) only provide a superficial understanding of the internal and external factors that correlate with variation. As for timing, our analysis indicates that there was ongoing obsolescence of the feature, as the earlier periods (1250–1350 and 1420–1500) display higher /h/ insertion rates, whereas there is a substantial decrease during the periods between 1350 and 1640. This supports existing claims (e.g. Milroy 1992) and suggests that /h/ insertion became less frequent from the Early Middle English period onwards.

There are a few desiderata for future research in historical sociolinguistics. For one, one would need to apply this methodological framework for the analysis of other corpora. Given that our findings indicate a decline in the 14th century, LAEME would be an ideal starting point. Not only would this allow us to push the timeframe back by about 150 years, but such a study would also bring to light considerably higher numbers of tokens. This ideally would enable us to conduct a variationist analysis so that we can take a further step in order to unwrap the envelope of variation. Such knowledge would help us to contextualise synchronic findings, namely data from present-day varieties of English where inserted /h/ has survived (Tristan da Cunha in the South Atlantic, Palmerston Island in the Pacific, various Caribbean settings) in order to locate founder processes and, though this may be a long shot, contact-induced innovation mechanisms that operate in the formation of Englishes around the globe. Finally, scrutinising lexical variation throughout the history of English may provide us with more information on the possible origins of /h/ insertion in English (i.e. whether this originated as a contact-induced phenomenon with French or whether it was present already in Old English, which would point to earlier origins or perhaps even feature legacy from the Germanic dialects that served as inputs to Old English). We hope that this paper provides a further stepping stone for these research questions.

References

Batchelor, T. 1809: *An Orthoëpical Analysis of the English Language*.

CEEC: *Corpus of Early English Correspondence*. 1998: T. Nevalainen, H. Raumolin-Brunberg, J. Keränen, M. Nevala, A. Nurmi & M. Palander-Collin comps. Helsinki, Department of Modern Languages, University of Helsinki.

Ellis, A. 1889: *On Early English Pronunciation.* Part V. London, Truebner and Co.

HC: *Helsinki Corpus of English Texts*. 1991: M. Rissanen, M. Kytö, L. Kahlas-Tarkka, M. Kilpiö, S. Nevanlinna, I. Taavitsainen, T. Nevalainen & H. Raumolin-Brunberg comps. Helsinki, Department of Modern Languages, University of Helsinki.

Jespersen, O. 1949: *A Modern English Grammar on Historical Principles*. Vol. I: *Sounds and Spellings*. London, Allen and Unwin.

Jones, C. 2006: *English Pronunciation in the Eighteenth and Nineteenth Centuries*. Basingstoke, Palgrave Macmillan.

Kytö, M. ed. 1996: *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*. 3rd ed. Helsinki, Department of English, University of Helsinki.

Labov, W. 1982: Building on Empirical Foundations. In W. P. Lehmann & Y. Malkiel eds. *Perspectives on Historical Linguistics*. Amsterdam & Philadelphia, John Benjamins: 17–92.

LAEME: *A Linguistic Atlas of Early Middle English, 1150–1325*. 2013–: M. Laing comp. Version 3.2. Edinburgh, The University of Edinburgh. http://www.lel.ed.ac.uk/ihd/laeme2/laeme2.html.

Lass, R. 1992: Phonology and Morphology. In N. Blake ed. *The Cambridge History of the English Language*. Vol. II: *1066–1476*. Cambridge, Cambridge University Press: 23–155.

Luick, K. 1964: *Historische Grammatik der englischen Sprache*. Oxford, Blackwell.

MED: *Middle English Dictionary*. 1959: R. E. Lewis, ed. Ann Arbor, University of Michigan Press.

Milroy, J. 1992: *Linguistic Variation and Change: On the Historical Sociolinguistics of English* (Language in Society 19). Oxford & Cambridge, Blackwell.

Minkova, D. 2013: *A Historical Phonology of English*. Edinburgh, Edinburgh University Press.

Mugglestone, L. C. 1995: *"Talking Proper": The Rise of Accent as Social Symbol*. Oxford, Clarendon Press.

Mugglestone, L. C. 2006: English in the Nineteenth Century. In L. C. Mugglestone ed. *The Oxford History of English*. Oxford, Oxford University Press: 274–304.

Schreier, D. Forthcoming: Tracking Language Change via Dialect Transplantation: 1,200 Years of /h/ Insertion in English.

Trudgill, P., D. Schreier, D. Long & J. P. Williams 2003: On the Reversibility of Mergers: /w/, /v/ and Evidence from Lesser-Known Englishes. *Folia Linguistica Historica* 37 (Issue *Historica* Vol. 24.1–2): 23–46.

Walker, J. 1791: *A Critical Pronouncing Dictionary* (R. C. Alston, *English Linguistics 1500-1800* 117). Menston, The Scolar Press.

Wallis, C. <*h*> *in Old English*. (Unpublished ms.)

Author's address
English Department
University of Zurich
Plattenstrasse 47
CH-8032 Zürich
Switzerland
e-mail: schreier@es.uzh.ch, milan.markovic2@uzh.ch, sasa.petrovic@uzh.ch

received: 15 March 2018
revised version accepted: 5 April 2018