*Terttu Nevalainen & Helena Raumolin-Brunberg**

Sociolinguistics and Language History: The Helsinki Corpus of Early English Correspondence[1]

Abstract

The paper introduces our new project on diachronic sociolinguistics, focusing on the problems of compiling a representative corpus for this purpose. We study long-term linguistic change in the Late Middle and Early Modern English periods (1420-1680) in a computer-readable corpus of personal letters, which is designed specifically for the purposes of sociohistorical research. When completed, the *Helsinki Corpus of Early English Correspondence* will comprise some 1.5 million running words representing all the literate social ranks of the time, both sexes, and different ages and occupations. In our case, the issues that a corpus compiler must deal with include the coverage of all the sociolinguistically relevant categories of data, authenticity of extant materials, and the quality of editing.

1. Introduction

Our new project investigates the extent to which modern sociolinguistic models and methods are applicable to diachronic linguistics. Despite all the recent work on historical corpora, the kind of material that we need is not available in sufficient quantities in electronic form. We have therefore had to start by compiling a sociolinguistically representative corpus for the periods that we are interested in. As its title suggests, the *Helsinki Corpus of Early English Correspondence* consists of letters. It covers the Late Middle and Early Modern English periods from 1420 to 1680.

---

1 The people working on the project are Terttu Nevalainen, director of the project, Helena Raumolin-Brunberg, full-time researcher (both are happy to answer any inquiries), and two part-time research assistants, Arja Nurmi and Minna Palander-Collin. Our postal address is: Department of English, P.O. Box 4 (Hallituskatu 11), FIN-00014 University of Helsinki. E-mail: tnevalainen@cc.helsinki.fi, raumolinbrun@cc.helsinki.fi.

* *Terttu Nevalainen & Helena Raumolin-Brunberg*
University of Helsinki
Department of English
P.O.Box 4 (Hallituskatu 11)
00100 Helsinki (FIN)

Our problems in constructing the framework for our study are in principle the same as modern sociolinguists are faced with today. Our whole undertaking is licensed by the principle of uniformitarianism, the claim that the kind of social factors that operate in present societies should also have applied in societies of the past. In other words, if social factors, roles and structures, commonly correlate with language change today, they should also have been relevant to linguistic change three or four hundred years ago (Labov 1994, 21-25).

It should be emphasized that the aim of our project is to explore the extent to which modern sociolinguistic models and methods apply to the past. We are, by no means, taking them for granted and advocating any naïve one-to-one correspondence between such constructs as social class in present-day urban sociolinguistics and the social ranks of pre-industrial societies. One of our problems is precisely to come up with analytic tools that are historically justified.

This paper concentrates on different aspects of representativeness of the corpus we are compiling. Section 2 introduces the current state of our work and compares it with the diachronic part of the *Helsinki Corpus of English Texts*. Section 3 discusses the 'hard' facts that we use to screen our materials and assess their validity. Section 4 completes the picture with the 'soft' facts that must be taken into account if we are to make the most of the data that we have at our disposal. Both these aspects will influence not only our corpus organization as a whole but also the kind of information to be supplied for each text in the corpus. This point is taken up in section 5, and some of our results so far are discussed in section 6.

2. Current state of the project

Financed by the Academy of Finland, our project started in September 1993 and is due to finish by the end of 1995. Our team has so far sampled data from about forty collections of correspondence. Some 1.4 million running words have been selected for further processing, and 1.1 million words have already been stored in computer-readable form. Out of these, some 650,000 words have been proofread once, and constitute the basic material of our pilot studies. Within the limits of our budget, we are aiming at a total of 1.5 million running words. This is estimated to furnish enough material for both real-time and apparent-time studies of language change.

These figures are much higher than those for the letters sampled for the Late Middle and Early Modern English sections of the diachronic part of the Helsinki Corpus of English Texts (less than 100,000 running words of correspondence altogether), or any other historical corpus containing letters from our period (see Kytö et al. 1993). All the Helsinki Corpus data are supplied with participants' descriptions, but the small number of letters included makes it less suitable for detailed sociohistorical studies than we had hoped.

We shall of course have to bear in mind that the Helsinki Corpus is a longitudinal general-purpose corpus, covering a time span of about one thousand years from the 8th to the 18th century. It was originally compiled for studies in textual variation in exactly the same way as the Brown and LOB corpora were. This means that its criteria for selection were oriented towards textual rather than social representativeness. The Early Modern English section, for instance, consists of fifteen different text types, which together make up 550,000 words (see Nevalainen & Raumolin-Brunberg 1993). As our present interests are more focused, the criteria of selection applied are also differently weighted. It goes without saying that the main criterion is the authenticity of our material.

3. 'Hard' facts, or problems of authenticity

The authenticity of our primary data involves a large number of problems. Since it is outside the scope of our type of corpus work to edit unedited manuscripts, the only sources that we have consist of letters which are available in an edited form.

In an ideal case we have in front of us a carefully edited collection based on letters that were actually delivered from one person to another. It is also important that these letters were written personally by people whose social backgrounds are fully indentifiable. Collections like this exist, but unfortunately their number is not very large. Good examples are the *Barrington Family Letters* (1628-1632) and the well-known letters by Dorothy Osborne to her future husband, Sir William Temple (1652-1657).

On the other hand, there are many collections where the majority of letters, but not all, are based on autograph sources. A typical instance is a collection compiled around one person who was the recipient of several authentic letters but, as far as his or her own letters are con-

cerned, the only material that remains is a collection of drafts or a letter-book of copies. This is for example the case with the diarist Samuel Pepys' family letters (1663-1680). Copies of letters sent by Pepys are found in a letter-book, written in a secretary's hand, which is every now and then interspersed by corrections in Pepys' own writing.

One step further, there are whole collections of letters edited from copies, like *The Letters of John Holles, 1587-1637*, which were edited by P.R. Seddon from four letter-books that had been copied by the eldest son of John Holles at different times. Going further still, the letters written by the Plumpton family from Yorkshire (1480-1549/50) were edited in 1839 from early seventeenth-century copies.

Furthermore, it was customary for persons in high administrative offices to use secretaries and amanuenses for writing letters. This was more or less the rule for royal letters, at least the non-private ones. At the other end of the social scale, due to widespread illiteracy, the lower and middle sections of society and especially women had to rely on secretarial help in their correspondence. A good piece of evidence of the extent of illiteracy at the turn of the seventeenth century is the Southampton captain Thomas Stockwell, who, although considered a gentleman, used a mark instead of a signature in his letters and business contracts.

The above examples raise the general issue of how to deal with drafts and copies and with non-autograph letters. We must not forget that the corpus we are compiling is not designed simply to represent the language of one period or another. Our aim is rather to create a socially representative corpus, and hence it is vital that the language we study is something that was produced by an identifiable person.

It was relatively easy to decide how to deal with drafts and copies that were written by the sender personally. They are treated like authentic letters; in any case they are autograph and could have been delivered to the recipient. The letter-book copies in a secretary's hand but corrected by the sender have also been given a high priority. As the editor H.T. Heath of the Pepys collection (1955) argues, the procedure was probably such that Pepys first dictated a letter to his secretary, who wrote it down in the letter-book. Then it was corrected by Pepys, and on this corrected draft the secretary wrote the letter, which Pepys signed.

As far as uncorrected secretarial letters are concerned, there is no way of knowing whether they were dictated or written in accordance with some general instructions of the sender (see Davis 1971, xxxviii-xxxix). We cannot consider them really representative of the sender's language. Our original thought was to exclude letters based on secretarial copies of this type. During the course of our work we have come to realize that there are subperiods and social ranks which are impossible to reach if we totally ignore these copies. As a compromise we have decided to include some of them as supplementary data.

The quality of editing may also create problems. The existing editions vary a great deal. Some date from the early nineteenth century, while others are very recent. Some have been produced for historians by historians without any philological training, whereas some others testify to outstanding historical and linguistic expertise. Recently edited collections usually have good accounts of editorial principles, while some of the older ones hardly provide any information at all.

Our main problem is the frequent modernization of spelling. Although there are research topics that could be pursued despite changed orthography, at least at this stage of our work we consider modernized spelling a sufficient drawback to exclude the edition from the corpus. We shall nevertheless have to allow for some minor changes that have been made in most collections. Typically, capitalization and punctuation have been modernized and abbreviations expanded.

4. 'Soft' facts, or problems of coverage

All these 'hard' facts apply at the microlevel, that is, they will have to be determined with respect to each and every letter included in the corpus. There are also certain macrolevel considerations that have guided our selection procedure as a whole. We shall here discuss two closely related issues: socioregional representativeness, on the one hand, and linguistic representativeness, on the other. It is true of both of them, of course, that we are dealing with an imperfect historical record. Although the number of texts that have been preserved from the Late Middle and especially Early Modern period is considerable, there are bound to be gaps. We are in fact lucky to have so many letter-writers who were keen archivists themselves, or had the misfortune of being sued by their adversaries, with the result that their personal correspondence was confiscated by courts as legal evidence.

The requirement of social representativeness means that we shall have to obtain data from both sexes, young and old people alike, and from as many social ranks as possible. The most difficult requirement is the last one. To gain some idea of the social stratification of the times we are studying, we might look at the system of rank and status in Stuart England. Table 1 is reconstructed on the basis of Laslett (1983, 38).

Table 1. Rank and status in Stuart England (based on Laslett 1983, 38).

	Grade		Title	Occupational Name
G E N T R Y	Nobl e men	1. Duke Archbishop		
		2. Marquess		
		3. Earl	Lord, Lady	none
		4. Viscount		
		5. Baron Bishop		
	Gent lemen	6. Baronet	Sir, Dame	Professions: Army Officer, Doctor of Medicine, Merchant, etc.
		7. Knight		
		8. Esquire	Mr, Mrs	
		9. Gentleman		
		Clergyman		
		10. Yeoman	Goodman, Goodwife (Goody)	
		11. Husbandman		Husbandman
		12. Craftsman Tradesman Artificer	none	(Name of Craft) Carpenter, etc.
		13. Labourer		Labourer
		14. Cottager Pauper		none

Concentrating on the major distinctions, it is relatively easy to obtain data from people representing nobility and gentry, and professional people such as clergymen. The problem is obtaining representatives of the majority of the English population who rank below the gentry. Throughout our period, but especially in the first half, the rate of full literacy (both reading and writing) is extremely low among the lower social ranks. Hence we should be prepared to relax our 'hard'-fact criteria if we are fortunate enough to find correspondence by people representing these ranks. Usually it is the background information that is missing in the case of our yeomen and petty merchants.

Full literacy is also amazingly low among women. According to the findings of David Cressy (1980, 119-121), the overall literacy of English women in the second half of our period was at the same level as that of lowest-ranking men, manual labourers. So letters attributed to women even in poorly edited sources are valuable to us. So far we have only excluded modernized editions containing women's letters.

Not surprisingly, full regional coverage is not a possible goal in our case. For one thing, the best representatives of rural dialects were illiterate. With the rise of the standard language, a dialect atlas of Early Modern English is not envisaged even by historical dialectologists. Our practical solution has been to concentrate, whenever possible, on three broad areas: London and its vicinity, Norfolk, and the 'North'. They are all linguistically motivated; London because of its role in the standardization process of the English language, Norfolk and the Northern counties as different regional varieties and as input to the process of standardization. All three are well-represented in historical records and offer a fair amount of diachronic continuity, sometimes even within one and the same family, such as the Pastons and the Bacons.

Linguistic representativeness ranks high on our agenda of 'soft' facts. As we are mostly interested in the diffusion of morphosyntactic changes in English, single-letter informants are not very useful in this respect. Our average letter is far too short to provide enough instances of morphosyntactic variables for statistical analysis. Whenever possible we have tried to secure a minimum of ten letters per informant, even if not all of them were dated, for instance. People represented by fewer letters can nevertheless be valuable, especially when they come from the lower ranks. The information they yield can be pooled from different sources and used in studies of social stratification.

5. Corpus organization

Finally, the question remains how much and in what form to incorporate extralinguistic information into the corpus. At the moment we feel that we can supply only some of the facts. Our participants' descriptions are time-consuming to compile and some of the information needed is far from easy to come by (see Johansson 1993). Some of it is also conjectural, and hence a matter of our interpretation rather than solid facts. We have therefore decided to work cyclically on three levels, hoping to learn from experience as we go on.

At the moment we are experimenting with three databases intended for different purposes. We are in the process of building a letter database, so far in hard copy format only, which will ideally contain the background information associated with each individual letter, its sender and recipient, and their social identities. We are similarly in the process of devising a sender database of all our informants, and a database of all our letter collections. The latter will contain, for each collection, the kind of hard and soft information that has been discussed above.

6. Prospects

The first results of our project are very encouraging. Linguistic change in apparent time, such as variation in relation to age, for instance, can be detected in the language of late 15th-century London wool-merchants (Raumolin-Brunberg & Nevalainen, forthcoming). We are also at the moment busy studying the forms of address in the data. It is common knowledge that a radical simplification of these forms took place in our period. What we do not know is the route through which the change was implemented; was it from the highest ranks to the lowest, vice versa, or perhaps through social equals in each estate? In fact, one of the key issues we are going to explore more fully is the question whether the notion of social stratification is of relevance to language change in our diachronic context of study. From what modern sociolinguistics tells us, it ought to be.

References

1. Letter collections

Barrington = Barrington family letters, 1628-1632. Ed. by A. Searle. Camden fourth series 28. London: Royal Historical Society, 1983.

Holles = Letters of John Holles, 1587-1637. Ed. by P. R. Seddon. Vol 1. Thoroton Society 31. Nottingham, 1975.

Osborne = The Letters of Dorothy Osborne to William Temple. Ed. by G. C. Moore Smith. Oxford: Clarendon Press, 1959 [1928].

Paston = Paston letters and papers of the fifteenth century. Vol. 1. Ed. by N. Davis. Oxford: Clarendon Press, 1971.

Pepys = The letters of Samuel Pepys and his family circle. Ed. by Helen Truesdell Heath. Oxford: Clarendon Press, 1955.

Plumpton = Plumpton correspondence. A series of letters, chiefly domestick, written in the reigns of Edward IV, Richard III, Henry VII and Henry VIII. Ed. by T. Stapleton. Camden Society 4, 1839.

2. Articles and monographs

Cressy, D. (1980): *Literacy and the social order: Reading and writing in Tudor and Stuart England*. Cambridge: Cambridge University Press.

Johansson, S. (1993): Some aspects of the recommendations of the Text Encoding Initiative, with special reference to the encoding of language corpora. In: M. Kytö, M. Rissanen & S. Wright, eds., 203-212.

Kytö, M., M. Rissanen & S. Wright, eds. (1993): *Corpora across the centuries, Proceedings of the First International Colloquium on English Diachronic Corpora*. Amsterdam: Rodopi.

Labov, W. (1994): *Principles of Linguistic Change* Vol. 1: Internal factors. Oxford: Blackwell.

Laslett, P. (1983): *The world we have lost (further explored)*. London: Methuen.

Nevalainen, T. & H. Raumolin-Brunberg (1993): Early Modern British English. In: M. Rissanen, M. Kytö & M. Palander-Collin, eds., *Early English in the computer age, Explorations through the Helsinki Corpus* (Topics in English Linguistics 11), 53-73. Berlin: Mouton de Gruyter.

Raumolin-Brunberg, H. & T. Nevalainen (forthcoming). Like father (un)like son: a sociolinguistic approach to the language of the Celys. In: J. Fisiak, ed., *Studies in Middle English*. Berlin: Mouton de Gruyter.