How to Handle Small Samples: Bootstrap and Bayesian Methods in the Analysis of Linguistic Change

Alexander Hinneburg
Institute for Informatics, Martin-Luther University, Halle/Saal, Germany

Heikki Mannila
HIIT Basic Research Unit, University of Helsinki and Helsinki University of Technology, Finland

Samuli Kaislaniemi, Terttu Nevalainen and Helena Raumolin-Brunberg
Department of English, University of Helsinki, Finland

Correspondence:
Prof. Heikki Mannila,
Helsinki Institute of
Information Technology,
P.O. Box 68, 000140
University of Helsinki,
Finland.
E-mail:
mannila@cs.helsinki.fi

Abstract

Estimating the relative frequencies of linguistic features is a fundamental task in linguistic computation. As the amount of text or speech that is available from a given user of the language typically varies greatly, and the sample sizes tend to be small, the most straightforward methods do not always give the most informative answers. Bootstrap and Bayesian methods provide techniques for handling the uncertainty in small samples. We describe these techniques for estimating frequencies from small samples, and show how they can be applied to the study of linguistic change. As a test case, we use the introduction of the pronoun *you* as subject in the data provided by the *Corpus of Early English Correspondence* (c. 1410–1681).

1 Introduction

There are several alternative ways of saying the same thing, arising from both register differences and speaker variables, such as gender, age, socioeconomic class, and regional background. This variationist or sociolinguistic approach to language contests the traditional view of language as tending towards a balanced state where each function is realized by one form only. The analysis of linguistic variation has proved essential for our understanding of the processes of language change. In studies of differences in linguistic usage between populations or points in time, a crucial task is to quantify the frequencies of linguistic features. In its simplest form, the question can be posed as follows: suppose there are two alternative forms A and B of a word, grammatical construct, etc.; what is the frequency of the use of each? Of course, several problems have to be tackled even before this question can be asked. Are the forms A and B really well-defined? What are the populations we are comparing? Is the sample representative? And so forth.

In this study, we consider the seemingly simple task of estimating the frequency of alternative forms on the basis of available data. In diachronic research

doi:10.1093/llc/fqm006 Advance Access Published on 12 April 2007

in particular, sample sizes tend to be small, and the amount of linguistic data is beyond the control of the researcher. It is, therefore, important to get information about the uncertainty in the result: we need methods that provide information about how reliable the estimate of the frequency of the linguistic variable is.

The simplest method to use is pooling, i.e. combining all the observations, and computing the frequency of each form over a given time period. Almost as simple is computing the average of individual averages. These methods provide a single estimate, and give no direct indication of how reliable or unreliable that estimate is. Bootstrapping (Efron and Tibshirani, 1993) is a general technique that can be applied in this type of situation. In bootstrapping, one repeatedly creates new data sets from the original data by sampling with replacement. For each such data set one computes the frequency by pooling or averages of averages, and uses the distribution of the frequency in the sampled data sets as a way of obtaining confidence intervals for the frequency in the original data. Bootstrapping is easy to implement, and has been used in a wide variety of tasks. Linguistic applications are described in, e.g. Ogura and Wang (1996), Collins *et al.* (2004), and Spencer *et al.* (2003).

Pooling and averages of averages are maximum likelihood methods, i.e. they try to obtain one estimate for the value of the unknown frequency u of the new form. Bootstrapping and other similar methods provide ways of measuring the uncertainty in this estimate. Bayesian methods, on the other hand, consider the unknown parameter u as a random variable, and estimate the probability distribution for u, not just a single (most likely) value as in maximum likelihood methods. Bayesian statistics is very well-developed (Bernardo and Smith, 1994), and it has been used in many domains. For applications in linguistics, see for example, Eisner (2002) and Clement and Sharp (2003).

In this article, we will consider the bootstrap method and the Bayesian approach for analyzing the frequency of use of an incoming form. Having described both methods and discussed their properties, we will demonstrate their application to the data provided in Nevalainen and Raumolin-Brunberg (2003), drawn from the 2.7-million-word *Corpus of Early English Correspondence* (CEEC). The software for the Bayesian method is available at http://www.cs.helsinki.fi/u/mannila/smallsamples/.

The rest of this article is organized as follows: Section 2 describes the background for the study and the material sampled from the CEEC; Section 3 considers the introduction of the subject *you*, a linguistic phenomenon we use as our case study; in Section 4, we present the basic principles of the methods, describe their implementation, and study some simple examples; Section 5 examines the results for the data from the CEEC in more detail, and Section 6 is a short conclusion.

2 Background

Sociolinguistics investigates the multitude of functions that language performs in modern society. At any given time, there are different ways of saying the same thing, not only in terms of register but in terms of speaker variables. This approach to language variation calls into question the traditional view of a one-to-one correspondence between meaning and form. Correlations between linguistic and external variables have also proved relevant to our understanding of the processes of language change.

Understanding language change is one of the major goals of sociolinguistic inquiry. One of the principal tenets of sociolinguistics has long been that the present could be used to explain the past (Labov, 1972, 1994). It is argued that, as living languages continually vary and change in patterned ways, the principles of language change could be arrived at by studying the present. From a language historian's point of view, it is, however, equally evident that present-day sociolinguistics cannot replace the study of linguistic variation in the past in its own right.

The key to understanding the social embedding of language change, past as well as present, is to examine texts that come as close as possible to day-to-day interaction between people. The data used in this study comes from the *Corpus of Early English Correspondence* (CEEC)[1], about 2.7 million running words, consisting of 6,039 letters written by 778 people between the early fifteenth and late

Table 1 Distribution of the writers in the *Corpus of Early English Correspondence*, 1410–1681, according to social status, domicile, and gender. Total number of writers: 778

Social status (%)	Domicile (%)	Gender (%)
Royalty (3)	Court (8)	Female (26)
Nobility (15)	London (14)	Male (74)
Gentry (39)	East Anglia (17)	
Clergy (14)	North (12)	
Professionals (11)	Other (49)	
Merchants (8)		
Others (non-gentry) (10)		

seventeenth centuries (Table 1). The vast majority of these writers represent the higher social ranks, the gentry, and professionals, but the aim of the corpus compilers has been to provide as balanced a social coverage as possible with the resources available. Besides chronological and regional variation, the external variables studied using this material include gender, domicile, social status, mobility (social and geographic), level of education, and register variation defined in terms of the relations between the writer and the addressee (Nevalainen and Raumolin-Brunberg, 2003, 43–9, 190).

3 Test Case: The Introduction of Subject *you*

As the linguistic example for this article we have chosen the introduction of the traditional object pronoun *you* in the subject function, which took place during the period c. 1450–1600. It is one of the fourteen changes discussed in detail in Nevalainen and Raumolin-Brunberg (2003).

The whole second-person pronoun system underwent two major changes in Late Middle and Early Modern English (1350–1700). Firstly, the singular pronoun *thou/thee* was replaced by the plural pronoun *ye/you*. Secondly, the object form *you* came to be used in the subject function, gradually ousting the historical subject form *ye* from the language. After a gradual start in the fifteenth century, this change diffused rapidly and was completed in the sixteenth century. Examples (1)–(4) illustrate the use of these second-person pronouns. Example (1) contains only *ye* in the subject position, while (2) is one of the earliest occurrences of *you* in the CEEC. Examples (3) and (4) illustrate variable use during the rapid diffusion of *you*.

(1) Plese it you to vnderstond that Will Cely told me that *ye* had no knowledge from me fir payment of the xx li. *ye* of your curtesy delyuerd vnto William Lemster my seruaunte/to my gret marvel. (William Dalton, 1487; CELY, 228)[2]

(2) I wnderstonde that *yow* haue ben sore seke ande now well rewiwid, . . . (Thomas Kesten, 1479; CELY, 67)

(3) I perceve by your seid lettre, for asmyche as *ye* be ridyng forster in the New fforest undyr the Duke of Suffolke, *ye* say that *you* may lawfully take your eaise in ony lodge within the seid forest. (William Fitzalan, Earl of Arundel, 1530s; WILLOUGHBY, 24)

(4) *you* knowe for a certenty and a thinge without doubt, that *you* be bownden to obey your souerain lorde your Kyng. And therfore are *ye* bounden to leaue of the doute of your vnsure conscience in refusinge the othe, . . . (Sir Thomas More, 1534; MORE, 505)

The first occurrences of subject *you* were found in ambiguous linguistic contexts, for example, after the verb in questions and imperatives. As for what caused the change, several suggestions have been made. Phonological confusion is a likely cause, since both pronouns apparently had a similar weak or fast-speech form pronounced [jə] or [ju]. A further factor increasing confusion may have been the personalization of impersonal verbs, e.g. *if you please*, *if ye please* (Lutz, 1998). After the early period, the linguistic constraints seem to disappear, and mixed cases such as Examples (3) and (4) become frequent. The diffusion of the new use of *you* was quite a rapid process, completed by the second half of the sixteenth century.

4 Different Ways of Estimating Frequencies

Let us assume that we have texts from a set of *N* individuals from a certain time, and that for each text we know how many times form A (such as a new form like *you*) and form B (a corresponding old form like *ye*) occur. We assume that there is an underlying frequency of form A in the population, and we want to estimate this. Such estimates are denoted by $\hat{u}$.

4.1 Simple methods

We next describe the principles of two simple approaches. Suppose we have data from N individuals. For each individual $i = 1, \ldots, N$, let a_i be the number of times form A occurs in the texts by i, and let b_i be the number of times form B occurs in the texts by i.

4.1.1 *Pooling*

The simplest approach is to pool all counts from the individuals in a given time period. That is,

$$\widehat{u} = \frac{\sum_{i=1}^{N} a_i}{\sum_{i=1}^{N} (a_i + b_i)} \tag{1}$$

Pooling suffers, of course, from the problem that if one individual has used the form A or B or both very often, i.e. if the value of a_i or b_i is large, then that individual dominates the result.

4.1.2 *Averaging the averages*

Another simple method is to calculate the frequency of use of form A for each individual, and average those percentages over the given time interval:

$$\widehat{u} = \frac{1}{N}\left(\sum_{i=1}^{N} \frac{a_i}{a_i + b_i}\right) \tag{2}$$

Averaging suffers from the opposite problem from pooling: as individuals have equal weight, a person who has used form A once and form B also once has the same impact on the frequency estimate as an individual who has used form A 100 times and form B once.

As an example, consider the data shown in Table 2 on all the male informants in the period 1540–59. There are large differences in the number of observations per individual, and since the change from *ye* to *you* is just taking place, the individual ratios of usage of the new form, $a_i/(a_i + b_i)$, vary considerably. Pooling gives us an estimate of 0.33, while averaging the averages gives 0.57.

Table 2 Counts of subject *you* (a_i) and *ye* (b_i) for the male informants in the CEEC sample during the period 1540–1559

Name	*you*	*ye*	Name	*you*	*ye*
Robert Andrew	21	1	Robert Lake	2	0
James Bassett	8	0	Henry Marmion	1	0
Charles Brandon	0	2	John Master	2	26
Christopher Breten	3	116	Peter Master	9	0
Thomas Brudenell	2	3	William Paget	109	52
William Butts	0	4	Thomas Phillipson	15	0
Ambrose Cave	7	0	Ralph Pinder	1	1
Anthony Cave	29	417	Richard Preston	3	78
John Coope	5	8	Henry Radcliffe	1	0
Thomas Cromwell	0	31	Richard Sandell	16	3
John Doddington	10	2	William Sandell	4	0
Thomas Egglesfield	1	3	Ambrose Saunders	68	42
William Fitzwilliam	0	5	Blase Saunders	2	2
Henry Garbrand	0	33	Laurence Saunders	19	0
Stephen Gardiner	11	86	Robert Saunders	18	0
John Gery	0	6	Henry Savill	2	34
George Graunt	2	1	Thomas Saxby	52	1
James Haddon	1	2	John Scrope	0	1
Henry VIII	12	3	Thomas Smith	8	7
Thomas Holland	1	0	Henry Southwick	57	0
Bartholomew Hosse	11	2	Edward Stanley	3	7
John Johnson	18	534	John Tupholme	17	0
Otwell Johnson	245	6	William Tupholme	1	0
Richard Johnson	5	155	Bartholomew Warner	41	0
Thomas Jolie	1	0	(Woodriff) Woodruffe	5	2
Andrew Judde	1	0	Thomas Wyatt	3	20

4.1.3 *Bounds on the occurrence frequencies*

The problems with pooling and averaging can be avoided to a certain extent by suitable selection of data. For averaging one might require that in order for the data for person i to be included in the analysis, the number of occurrences of the linguistic variable (i.e. $a_i + b_i$) should be higher than a given quantity. For pooling, one could do the opposite, i.e. remove persons with very high counts of a_i or b_i or both from the analysis. Individual quotas have also been applied to limit the influence of people with very high numbers of occurrences. Such methods are slightly *ad hoc*, but they have nevertheless been used with good success. Nevalainen and Raumolin-Brunberg (2003, 214–17) compare averaging with pooling, with and without individual quotas, and introduce three rather similar curves for two Early Modern English changes. The major differences detected occur when one informant clearly dominates, like the merchant John Johnson writing in 1540–59, who preferred the conservative form *ye*. Johnson's material accounts for about 20% of all the occurrences of the variable, and it is clear that in this case pooling gives a lower percentage for the new form *you* than the other two methods.

4.2 Bootstrap methods

The idea in bootstrap methods (Efron and Gong, 1983; Efron and Tibshirani, 1993) is to resample the data several times, i.e. select a number of random sets of the observations, and then to compute the frequency of occurrence of the new form in each resample. This is done for, say 1,000 samples, and the variation in the frequency in the resamples gives us a confidence interval for the frequency in the original data.

The resample is formed with replacement. In other words, given the data (a_i, b_i) for $i = 1, \ldots, N$, we generate N random integers $i_1, i_2, \ldots, i_N$, from $1, \ldots, N$, and the resample D' consists of the observations $(a_{i_1}, b_{i_1}), (a_{i_2}, b_{i_2}), \ldots, (a_{i_N}, b_{i_N})$. Note that the resample can contain many copies of an original observation, and that some of the original observations are not included in the resample.

Given the resample D', we compute the frequency u of the new form by, e.g. pooling or averaging the averages. This is repeated for many resamples D', which gives us an empirical distribution of u. An estimate for u is obtained from the median of u in this distribution, and confidence bounds for the frequency u can be found by taking the (say) 2.5 and 97.5 percentile of this distribution.

As an example, using the data in Table 2 and creating 1,000 samples, we obtain the empirical distributions of the pooled frequencies and averages of averages shown in Fig. 1. The median and confidence interval for the pooled estimate are 0.34 (0.17–0.64) and those for the average of averages are 0.57 (0.46–0.68). We see that the pooled estimates of the probability have a much larger variability than those of the averages of averages. This is due to the presence of individuals such as John Johnson with a large number of observations: the presence or absence of these persons in the bootstrap sample has a strong effect on the pooled frequency of the sample.

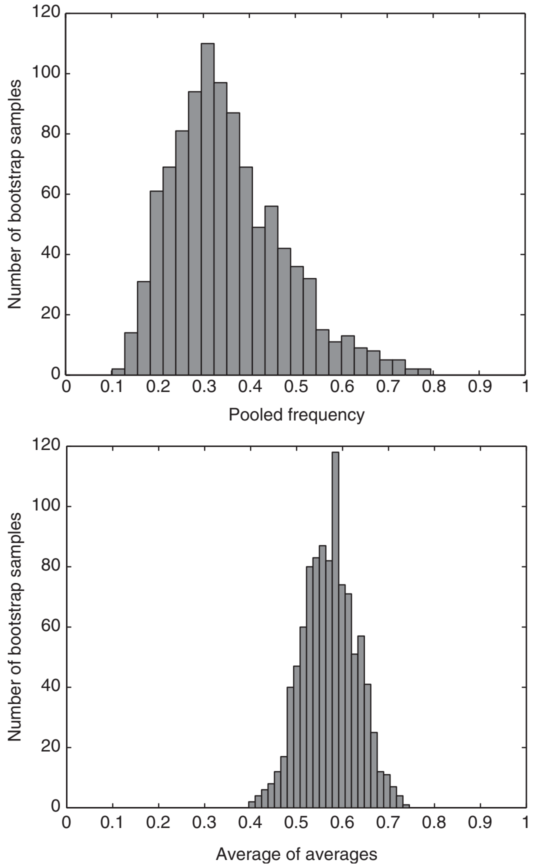

Fig. 1 Histograms of the pooled frequency and average of averages in 1,000 bootstrap resamples from the data in Table 2

The bootstrap method has a wide range of applications; see, e.g. Zoubir and Boashash (1998). There is an elaborate theory of the use of the bootstrap, and it can be shown that the estimates and confidence intervals obtained by using it have many desirable properties.

4.3 Bayesian approach

Maximum likelihood approaches aim at finding the most likely value for a parameter. In many situations it is not reasonable to assume that there would be one true value of the parameters. Bayesian statistics (Gelman *et al.*, 2004; Bernardo and Smith, 1994; Gamerman, 1997; Gilks *et al.*, 1996) is a widely used method in which the basic idea is that the parameters are also considered to be random variables, i.e. quantities with variation. For our example case of the frequency of usage of form A for the individuals in question, in the Bayesian approach we assume that this frequency is the quantity u, and then we try to estimate the distribution of the u, instead of a single estimate $\hat{u}$.

Let us assume that we have a prior probability $P(u)$ for each possible value of u. For example, we can assume that each possible value of u between 0 and 1 is equally probable. Let us denote by u_i the probability that individual i will use the new form; we assume that there is a probability $P(u_i \mid u)$ that u_i will have some value, given the value of u. Finally, let $P(a_i,b_i \mid u_i)$ be the probability that individual i uses the new form a_i times and the old one b_i times, given the value of u_i. This likelihood is the well-known binomial probability:

$$P(a_i, b_i|u_i) = \binom{a_i + b_i}{a_i} u_i^{a_i}(1 - u_i)^{b_i}. \qquad (3)$$

Note that this probability does not depend on u, but only on the individual probability u_i.

Let us denote the data by D; thus D consists of the values a_i and b_i for N individuals. Given the values of the parameters u and u_i for all i, the likelihood of the data D is

$$P(D|u, u_1, \ldots, u_n) = \prod_{i=1}^{N} P(a_i, b_i|u_i) \qquad (4)$$

That is, given the parameters, the probability of observing the data is the product of the probabilities of the data from each individual.

Using Bayes' rule, we can invert the probability in Equation (4):

$$\begin{aligned} &P(u, u_1, \ldots, u_N|D) \\ &= Z^{-1}P(u)\prod_{i=1}^{N} P(u_i|u)\prod_{i=1}^{N} P(a_i, b_i|u_i) \\ &= Z^{-1}P(u)\prod_{i=1}^{N} P(u_i|u)P(a_i, b_i|u_i) \qquad (5) \end{aligned}$$

Here Z is a normalizing constant. That is, the probability of having a certain combination of values for the parameters u and u_i for all i, given the data D, is proportional to the prior probability of the value of u [the term $P(u)$] times the probability of the individual parameters having the values times the probability of the data.

In general, we are not very interested in the individual probabilities u_i; rather, we want to find the distribution of the general probability u. This can be done by summing up all possible values of the parameters u_i, as follows:

$$\begin{aligned} &P(u|D) = Z^{-1}P(u) \\ &\qquad \times \sum_{u_i}\sum_{u_2}\cdots\sum_{u_N}\prod_{i=1}^{n} P(u_i|u)P(a_i, b_i|u_i). \quad (6) \end{aligned}$$

This can also be expressed as

$$\begin{aligned} &P(u|D) = Z^{-1}P(u) \\ &\qquad \times \prod_{i=1}^{n}\sum_{u_i} P(u_i|u)P(a_i, b_i|u_i), \qquad (7) \end{aligned}$$

a form that is much easier to compute. That is, the posterior probability of a value of the general probability u is proportional to the prior probability $P(u)$ of u times the product of the probabilities of observing the data of the individuals for some values of the individual probabilities u_i.

The Bayesian estimate $\hat{u}$ for u is obtained by computing the expected value of u under this distribution:

$$\hat{u} = \sum_{u} u\, P(u|D), \qquad (8)$$

To complete the description of the model, we need to specify the prior probability $P(u)$ and the probability $P(u_i \mid u)$ of an individual probability, given the value of u. For the prior probability $P(u)$ we simply use the uniform distribution: every value of u is equally probable. For the probability $P(u_i \mid u)$ we use

$$P(u_i|u) \propto \exp(-C(u_i - u)^2), \qquad (9)$$

where C is a constant. The role of C is to determine how strongly the individual parameters u_i depend on the general probability u. If C is large, then deviations of the parameters u_i from u is unlikely, while for small values of C such deviations can happen more easily. The choice of C does not effect the results very strongly; we use $C = 10$.

The Bayesian model has a form that can be implemented in a simple way. We precompute for each data point (a_i, b_i) and possible value of the parameter u_i the probability $P(a_i, b_i \mid u_i)$. Similarly, for each pair of possible values of u and u_i we compute $P(u_i \mid u)$. An accuracy of 0.01, for example, yields 10,000 different pairs of u and u_i, and computing $P(u_i \mid u)$ for all the pairs is quite feasible. The Equation (5) now makes it possible to compute for each possible value of u an estimate of the posterior probability $P(u \mid D)$. More complex Bayesian models can be handled by using MCMC techniques (Gamerman, 1997; Gelman *et al.*, 2004; Gilks *et al.*, 1996); sophisticated software exists for estimating the posterior probability for many types of models (Lunn *et al.*, 2000).

From the posterior distribution $P(u \mid D)$ we can further compute the interval (r,s) of values such as to give a probability of at least 95% that the value of u is within that interval; such an interval is called a 95% *posterior interval* in Bayesian literature.

Next we give some comments on the different approaches. Table 3 summarizes the results of the different methods for the data in Table 2. The pooling method estimates that the frequency of *you* is about one-third. However, the presence of persons like Anthony Cave and John Johnson has a strong influence on this estimate, as the number of cases for them is very high. The average of averages gives a higher estimate, as the number of persons preferring *you* is high, ~57%. The Bayesian estimate is in between these two estimates, as more weight is given to individuals with larger numbers of observations.

Table 3 Pooling, averaging the averages, and Bayesian estimates for the probability of subject *you* for the data in Table 2

	Estimate	**2.5%**	**97.5%**
Pooling	0.34	0.16	0.62
Averaging	0.57	0.46	0.67
Bayesian ($C = 10$)	0.54	0.45	0.62

The 2.5% and 97.5% bounds are the percentiles of the corresponding estimate for the pooling and averaging methods, and the percentiles of the posterior distribution for the Bayesian method (i.e. the posterior interval).

Some of the differences between the Bayesian approach and the bootstrap method for ascertaining confidence or posterior intervals are most clearly observed by looking at artificial data. As an extreme example, suppose we have one individual who used the new form once and did not use the old form at all, i.e. we have one observation of the form (1,0). All the bootstrap resamples are identical, and the pooled estimates and average of averages for each resample will be exactly 1, implying that the confidence interval is also (1,1). In general, for any sample, no matter how small, in which the one of the forms does not occur at all, yields in the bootstrap approach a confidence interval of width 0.

For this tiny data set with one observation (1,0), the Bayesian approach uses the prior information $P(u)$ about the probability of different values of u, and yields the estimate 0.66, with posterior interval (0.06–0.98), showing the wide range of variability in the estimate. Assume next that we have data from five individuals, and that each of them has used the new form five times and the old form zero times. Then all the bootstrap resamples will have five occurrences of the pair (5,0), and thus the bootstrap method again indicates that the usage frequency of the new form is equal to 1. On the other hand, the Bayesian estimate is 0.93 (0.72–0.99) for the parameter u. A sample of twenty observations (5,0) gives an estimate of 0.97. Increased evidence for the unequivocal use of the new form increases the Bayesian estimate, as the weight of the data is balanced against the uniform prior assumption of all values of u being equally likely.

We also implemented another version of the Bayesian approach for hierarchical binomial data, described in Gelman *et al.* (2004, 118 ff.). In this approach, the general probability u is parametrized by using a beta distribution with two parameters. The prior is an uninformative prior on a transformed scale of the two parameters. The results obtained using the model described in Gelman *et al.* (2004) are in fairly good agreement with the results obtained by the model given earlier (data not shown).

5 Results for the Change from *ye* to *you*

5.1 General remarks

As aforementioned, we have chosen the introduction of *you* to illustrate the Bayesian approach to the investigation of the diffusion of linguistic changes.

Based on the CEEC and the *Helsinki Corpus of English Texts*, Nevalainen and Raumolin-Brunberg (2003) and Raumolin-Brunberg (2005) studied the introduction of *you* by using the simple statistical methods referred to in Section 3.1. According to their findings, gender played a role after the incoming form had reached the frequency of 20%, women leading the change. No geographical origin for the change was detected, but significant regional differences emerged as the change progressed, the capital area being more advanced than the rest of the country. Informal spoken language seems to have been the origin of this change, which allows us to characterize it as a change from below in terms of social awareness (Labov 1994, 78). As far as the diffusion of the change in terms of social stratification is concerned, Nevalainen and Raumolin-Brunberg (2003) suggest that this change originated in the middle ranks, merchants and professional people, and spread from there to the upper ranks, but not to the lower. The endorsement of *you* by the upper ranks may in fact have been the factor explaining the rapid supralocal diffusion of this change.

5.2 Comparison of the different approaches

We computed the pooling, averaging the averages, and Bayesian estimates for all twenty-year periods from 1410 to 1600, for the subgroups of females and males, and for different regions, as well as for different social classes. We omitted all subgroups with less than four individuals. This resulted in 130 different subgroups. We analyzed each of these using the bootstrap method for pooling and averages of averages, and the Bayesian method with $C = 10$.

The correlations between the different estimates are very high, 0.97–0.99, the highest being the correlation between the average of averages and the Bayesian estimate. This is also visible in the pairwise scatterplots of the three estimates for these 130 groups, shown in Fig. 2. The correlations are strongly influenced by the agreement of the methods for the cases where the estimates are close to 0 or close to 1. As can be seen from, e.g. the right panel of Fig. 2, there is large variation for estimates that are in the middle.

The widths of the confidence intervals are shown in Fig. 3. In general, the differences are not large; the Bayesian and averages of averages methods are in better agreement than the Bayesian and pooling or pooling and averages of averages.The mean width of the posterior interval for the Bayesian method is 0.23, slightly larger than the mean width of the confidence interval for pooling (0.19) and for averages of averages (0.17). This is mainly due to the subgroups with no occurrences of *ye* or *you*; as mentioned, for such data sets the bootstrap approaches yield estimates of 1 and a confidence interval of width 0. For the eighty-five subgroups for which both forms occur, the mean sizes of the intervals are 0.24 for the Bayesian method, 0.27 for the pooling approach and 0.24 for averages of averages. The width of the Bayesian posterior intervals is somewhat influenced by the choice of the parameter C; setting $C = 5$ extends the mean width of the posterior interval to 0.27 (all subgroups) or 0.28 (subgroups in which both forms occur).

5.3 Linguistic observations

Figure 4, showing the Bayesian estimate on gender differences in the use of *you* and *ye*, confirms the female leadership of this change, previously suggested by simple methods of calculation. As regards the different methods, Table 4 indicates in general only minor differences between the approaches in the analysis of women's usage. In the case of men, the three methods yield quite divergent results.

As an example, consider the periods 1500–19 and 1540–59. For 1500–19, pooling gives an estimate of 0.19 and the Bayesian approach only 0.04. The reason is that a majority (seventeen out of twenty-three) of the letter writers in the data set have not used the feature at all; the few individuals with a large number of occurrences of *you* do not offset the

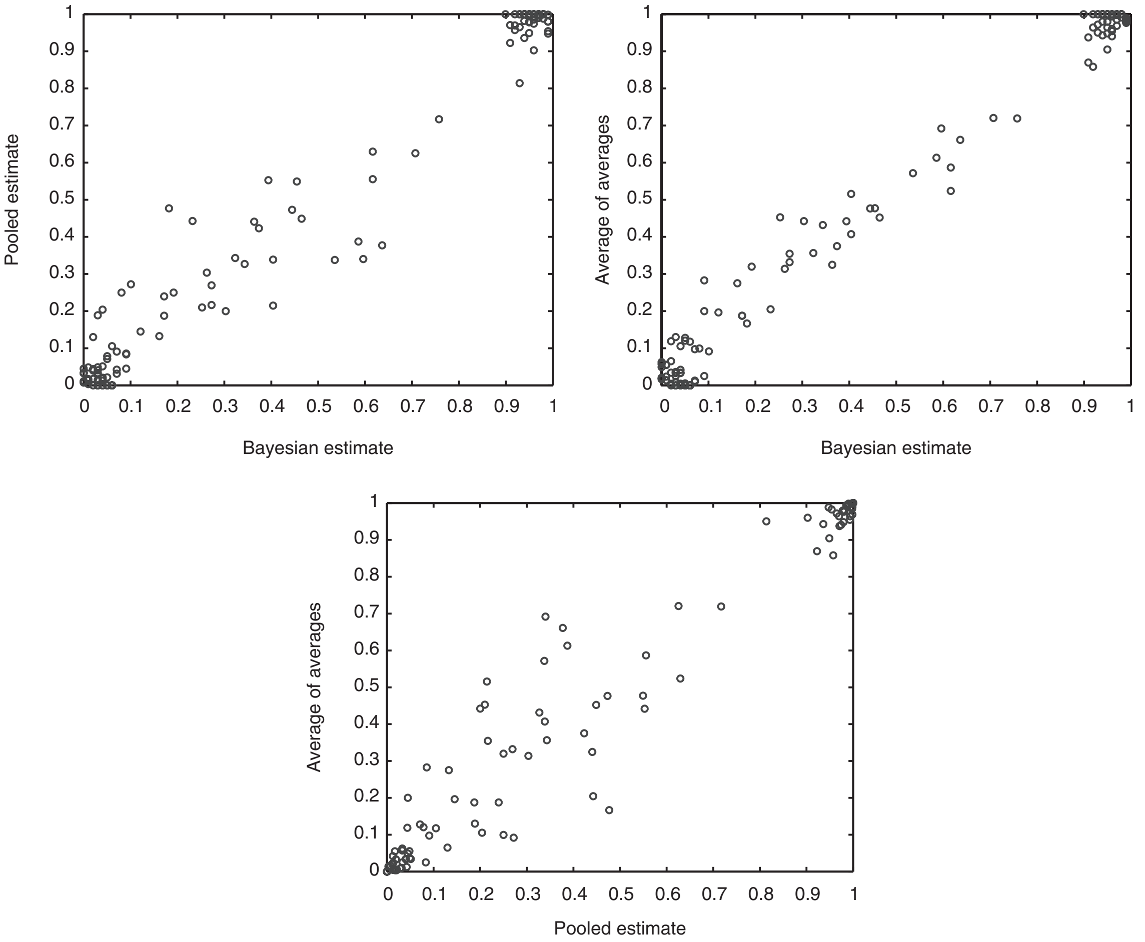

Fig. 2 Scatterplots of the Bayesian, pooled, and averages of averages estimates for 130 subgroups of the CEEC data

effect of the majority. During the period that John Johnson dominates, 1540–59 (Table 2), both averaging and the Bayesian approach take into account the differences in individual inputs, which pooling cannot handle.

Figure 5, showing the Bayesian estimates for the change in different regions, also confirms the results of previous research. For the study of regional variation in the CEEC, the usage in four areas was compared (Nevalainen and Raumolin-Brunberg 2003, 157–84). These were (1) the North of England, comprising counties north of Lincolnshire, (2) East Anglia, covering Norfolk and Suffolk, (3) London, covering the city, its suburbs and Southwark, and (4) the Court, less of a geographical unit, which consists of the royal family, courtiers and higher administrative officers, many of whom lived in Westminster. The City of London and the Court formed the capital area, which had a large number of in-migrants and weak-tie social networks, in other words, conditions which are likely to promote language change. For this article, the regional division was somewhat simplified, in that we compared the capital region (London, the Court and the Home Counties) with the North and East Anglia. We also contrasted the Capital region with the rest of the country, consisting of the North, East Anglia, and any data available from other regions in a given period.

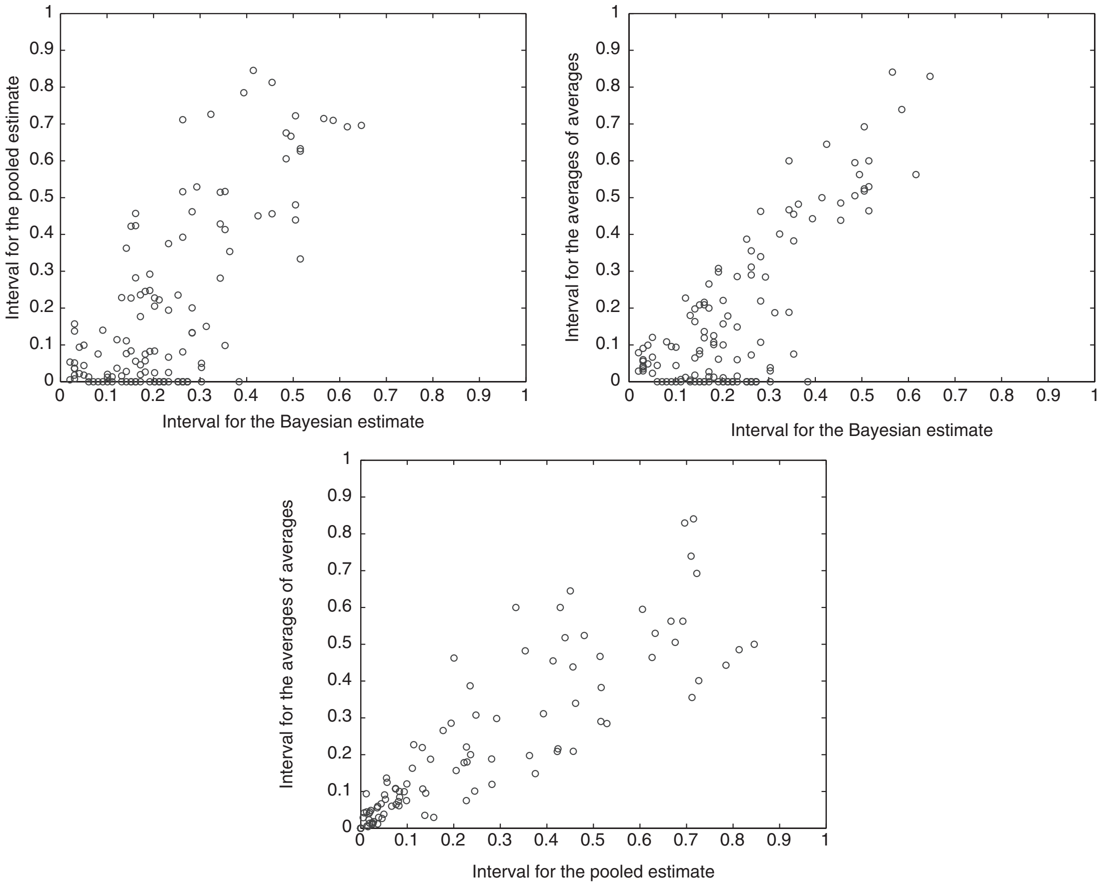

Fig. 3 Sizes of confidence and posterior intervals for Bayesian, pooled, and averages of averages estimates for 130 subgroups of the CEEC data

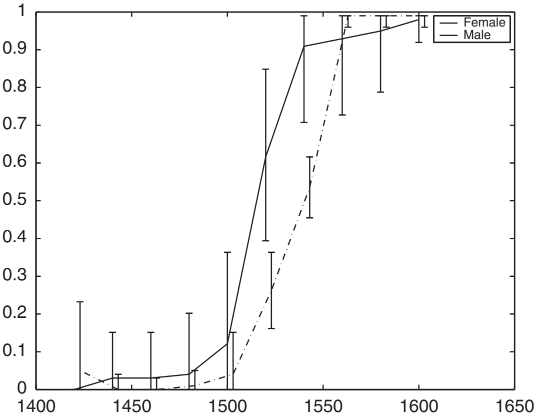

Fig. 4 Bayesian estimates for the probability of subject *you* for females and for males

Figure 5 shows that, after 1500, the capital region was in the lead when compared with the North and East Anglia until 1560, when the diffusion of subject *you* was practically completed. The comparison of the capital region against the rest of the country yields slightly different results, as there is a cross-over in 1540–59, the other areas taking the lead. This graph lends support to the concept of the snowball effect (e.g. Ogura and Wang, 1996), according to which the innovative form often spreads faster in an area which adopts it later. It is noteworthy that Table 5 indicates clear differences in the estimates for 1540–59 in all the regions. In all cases, pooling gives much lower estimates than the other two methods.

Table 4 The results of Bayesian, pooling, and averages of averages for female and male informants for the probability of subject *you*

Period	Females				Males			
	N	Bayes	Pool	Ave	*N*	Bayes	Pool	Ave
1410–39	<4	–	–	–	5	0.05	0.02	0.01
1440–59	9	0.03	0.00	0.00	39	0.00	0.01	0.02
1460–79	12	0.03	0.03	0.03	70	0.00	0.03	0.06
1480–99	10	0.04	0.02	0.03	48	0.01	0.05	0.06
1500–19	7	0.12	0.14	0.20	23	0.04	0.19	0.10
1520–39	14	0.62	0.63	0.53	54	0.26	0.31	0.32
1540–59	9	0.91	0.92	0.87	52	0.54	0.34	0.57
1560–79	6	0.93	1.00	1.00	59	0.99	0.99	0.98
1580–99	11	0.95	0.98	0.95	65	0.99	0.95	0.99
1600–19	21	0.98	1.00	1.00	51	0.99	1.00	0.98

N: the number of individuals in the subgroup; Bayes: the Bayesian estimate for the frequency; Pool: the pooled estimate; Ave: the average of averages.

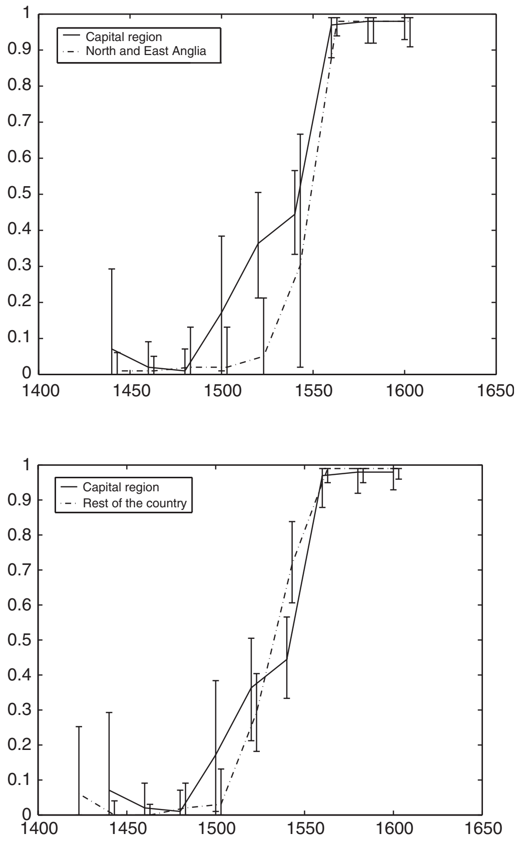

Fig. 5 Bayesian estimates for the probability of subject *you* in different regions. ‘Capital region’ includes London, the Court, and the Home Counties; ‘Rest of the country’ includes everything that is not ‘Capital region’

Our material makes it clear that the new form *you* was very rapidly diffusing among the population in 1540–59. This type of rapid diffusion typically involves extensive variation between individuals, including both those who strongly favor the old form and those who prefer the new one, as well as those who do not show categorical use of either. In cases like this, striving for one single estimate for all users may not be the best method to adopt.

6 Discussion and Conclusions

We have presented techniques for estimating the frequency of alternative forms from small samples. The basic maximum likelihood approaches are pooling or averaging the averages; their confidence intervals can be obtained by bootstrapping. Bootstrapping is based on the idea of repeatedly resampling the data, computing the frequency in the sample, and inferring the confidence intervals from the distribution of frequencies in the samples.

In the CEEC material used in our case study, pooling seems to be less useful than averaging the averages: pooling is susceptible to strong influence by individuals with lots of data, and at least in the CEEC data, results for averaging of averages seem to behave in a more consistent manner. Bootstrapping methods are very easy to implement and show robust behavior. One of their drawbacks is that if the data is very sparse, e.g. one of the forms is not used at all, bootstrapping gives a confidence interval of width 0, no matter how small the data is.

Table 5 The estimates for the probability of subject *you* for different regions

Period	Capital region (London, Court, Home Counties)				The North and East Anglia				All areas other than Capital region			
	N	Bayes	Pool	Ave	*N*	Bayes	Pool	Ave	*N*	Bayes	Pool	Ave
1410–39	<4	–	–	–	<4	–	–	–	6	0.07	0.00	0.00
1440–59	5	0.07	0.01	0.08	24	0.01	0.00	0.01	43	0.00	0.01	0.01
1460–79	24	0.02	0.04	0.09	33	0.01	0.02	0.03	58	0.00	0.03	0.04
1480–99	24	0.01	0.03	0.03	15	0.02	0.03	0.01	34	0.02	0.06	0.07
1500–19	11	0.17	0.23	0.23	13	0.02	0.04	0.03	19	0.03	0.15	0.07
1520–39	18	0.36	0.33	0.37	15	0.05	0.11	0.12	50	0.29	0.37	0.36
1540–59	28	0.44	0.27	0.50	4	0.30	0.18	0.44	33	0.72	0.62	0.71
1560–79	18	0.97	0.99	1.00	33	0.98	0.99	0.99	47	0.99	0.98	0.97
1580–99	26	0.98	0.93	0.98	31	0.98	1.00	1.00	50	0.99	0.99	0.98
1600–19	27	0.98	1.00	0.97	19	0.98	1.00	1.00	45	0.99	1.00	1.00

N: the number of individuals in the subgroup; Bayes: the Bayesian estimate for the frequency; Pool: the pooled estimate; Ave: the average of averages.

We also described a Bayesian model. The model has a global parameter for the frequency, and an individual parameter for each individual. Given the probabilities of the individual parameter given the global parameter, and the likelihood of particular data for an individual given the individual parameter, the Bayesian method gives a posterior distribution of the frequency parameter. This posterior distribution can then be used to obtain posterior intervals for the global parameter. Bootstrapping and Bayesian methods are especially useful when the sample sizes are small, and hence more traditional ways of determining confidence intervals cannot be used. Bootstrapping and Bayesian methods can, of course, also be used for large samples.

The Bayesian method is fairly easy to implement and gives robust results. The model looks at individuals, allows for variation in individual frequencies, and the amount of data that is available from an individual does affect the result. In this respect, one could say that the method is in between pooling (which does not take individuals into account) and averaging averages (in which the amount of data from a person does not have an effect).

The results we obtained for this case study confirm our previous findings in that they show how rapidly the second-person pronoun *you* replaced *ye* in the subject function. The process was completed in less than a hundred years, and the better part of it took place in the first half of the sixteenth century. The data provide us with clear statistical confirmation that women were instrumental in spreading the incoming form in the most rapid phase of the process in the sixteenth century.

As also shown by our earlier studies, the capital region (London, the Court and the Home Counties) clearly favored the incoming form in the first half of the sixteenth century when compared with geographically more peripheral regions such as East Anglia and the North. However, the Bayesian estimates also suggest that, compared with the data available from all other regions, the capital region was overtaken and superseded by the rest of the country when the change reached its mid-course in the 1550s. These new findings agree with the basic principle suggested by Ogura and Wang (1996) that localities that are later in adopting an ongoing change will, when it advances, not only catch up with those that started earlier, but surpass them.

In these cases, bootstrap and Bayesian methods have allowed us to assert findings based on small samples with more confidence. They have also helped us make some new discoveries. Equally importantly, they have alerted us to cases where the sample size is too small to allow statistical inferencing.

One problem in the model we have described is the dependence of the results on the parameter *C* used to weigh the importance of deviations of individual parameters from the global parameter. Varying *C* does not typically have a strong impact on the estimate of the frequency, but it does influence the width of the posterior interval. On the other hand, the Bayesian method gives reasonable results even for extremely small sample sizes, as it takes into account the prior information about the global parameter.

All the methods considered in this article are based on the assumption of one estimate of the frequency of the form for each subgroup. This might not be a valid assumption, however, and it would be interesting to consider methods that would search for possible components in the sample. Techniques such as mixture modeling could be useful in this regard.

In our sociolinguistic case study individual informants form the basic unit. It goes without saying that the methods introduced can be applied to other types of linguistic material as well; e.g. individual texts can be chosen as the sampling unit.

Acknowledgements

The research carried out for this article by S.K., T.N. and H.R.-B. was supported by the Academy of Finland Centre of Excellence funding for the VARIENG Research Unit. The work of H.M. was supported by the Academy of Finland.

References

Brooks, S. P. (1988). Markov chain Monte Carlo and its applications. *Statistician*, **47**: 69–100.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. New York, NY: John Wiley Inc.

Clement, R. and Sharp, D. (2003). Ngram and Bayesian classification of documents for topic and authorship. *Literary and Linguistic Computing*, **18**: 423–47.

Collins, J., Kaufer, D., Vlachos, P., Butler, B. and Ishizaki, S. (2004). Detecting collaborations in text comparing the authors' rhetorical language choices in the Federalist Papers. *Computers and the Humanities*, **38**: 15–36.

Corpus of Early English Correspondence (CEEC). (1998). Compiled by Nevalainen, T., Raumolin-Brunberg, H., Keränen, J., Nevala, M., Nurmi, A., Palander-Collin, M. University of Helsinki.

Efron, B. and Tibshirani, B. (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.

Efron, B and Gong, B. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, **37**: 36–48.

Eisner, J. (2002). Discovering syntactic deep structure via Bayesian statistics. *Cognitive Science*, **26**(3): 255–68.

Gamerman, D. (1997). *Markov Chain Monte Carlo. Stochastic Simulation for Bayesian Inference*. Texts in Statistical Science. London: Chapman and Hall.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd edn. New York: Chapman and Hall.

Gilks, W., Richardson, S., and Spiegelhalter, D. (eds) (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Helsinki Corpus of English Texts (HC). (1991). Compiled by the Helsinki Corpus project team at the Department of English, University of Helsinki.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1994). *Principles of Linguistic Change. Volume 1: Internal Factors*. Oxford, UK and Cambridge, USA: Blackwell.

Lunn, D. J., Thomas, A., Best, N. G. and Spiegelhalter, D. J. (2000). Win-BUGS – a Bayesian modelling framework: concepts, structure and extensibility. *Statistics and Computing*, **5**(6): 321–33.

Lutz, A. (1998). The Interplay of External and Internal Factors in Morphological Restructuring: The Case of *you*. In Fisiak, J. and Krygier, M. (eds), *Advances in English Historical Linguistics (1996)*. Berlin and New York: Mouton de Gruyter, pp. 189–210.

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

Mosteller, F. and Wallace, D. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. MA: Addison-Wesley.

Ogura, M. and Wang, W. S.-Y. (1996). Snowball Effect in Lexical Diffusion: The Development of -s in the Third Person Singular Present Indicative in English. In Britton, D. (ed.), *English Historical Linguistics 1994: Papers from the 8th International Conference on English Historical Linguistics*. Amsterdam: Benjamins, pp. 119–41.

Raumolin-Brunberg, H. (2005). The diffusion of YOU: a case study in historical sociolinguistics. *Language Variation and Change*, **17**(1): 55–73.

Spencer, M., Bordalejo, B., Robinson, P. and Howe, C. J. (2003). How reliable is a stemma? An analysis of Chaucer's Miller's Tale. *Literary and Linguistic Computing*, **18**(4): 407–22.

Zoubir, A. M. and Boashash, B. (1998). The bootstrap and its application in signal processing. *IEEE Signal Processing Magazine*, **15**: 56–76.

Notes

[1] The *Corpus of Early English Correspondence* (1998 version) was compiled by Jukka Keränen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, and the directors of the Sociolinguistics and Language History project, Terttu Nevalainen and Helena Raumolin-Brunberg. A sampler version of the corpus (CEECS) is available from the Oxford Text Archive (OTA; http://ota.ahds.ac.uk) and the International Computer Archive of Modern and Medieval English (ICAME; http://helmer.aksis.uib.no/icame.html). A tagged and parsed version of the corpus (PCEEC), available from the Oxford Text Archive, was released in 2006.

[2] The quotations from the CEEC have a text identifier consisting of the writer's name, the year of writing, the short title of the collection the letter comes from, and a page number.