The challenge of historical data: from sources and corpora to answering research questions

Wal, M.J. van der

ROZHLEDY

Marijke van der Wal

The challenge of historical data: from sources and corpora to answering research questions

ABSTRACT: At the start of each research enterprise, historical sociolinguists have to deal with the key issue of the required historical data. In the present world of big data optimism, the idea may arise that the time-consuming compilation of specialised corpora is no longer needed. Can taking a shortcut still lead us to convincing results? In this article I discuss the crucial relationship between specific research questions and appropriate historical data. This methodological issue will be illustrated by concentrating on three historical-sociolinguistic research programmes, conducted at Leiden University: *Letters as loot: Towards a non-standard view on the history of Dutch* (2008–2013), *Going Dutch: The construction of Dutch in policy, practice and discourse* (2013–2018) and *Pardon my French? Dutch-French language contact in the Netherlands, 1500–1900* (2018–2023). What do we learn from these large-scale projects which address different research questions and focus on different periods in the history of Dutch? The use of specific sources, handwritten material such as ego-documents, a multi-genre approach and details of corpus compilation will be discussed. The various approaches and results are considered against the background of methodological developments and current debates in historical sociolinguistics. I argue and conclude that the careful compilation of specialised corpora remains essential as a solid foundation for historical sociolinguistic research.

Key words: digital corpora, private letters, ego-documents, literacy, language history from below, language variation, language change, codification, language norms, French influence

1. The data issue

The present-day mature discipline of historical sociolinguistics covers a rich diversity of approaches, time periods and different languages (see i.a. Hernández-Campoy & Conde-Silvestre 2012; Nevalainen & Raumolin-Brunberg 2012; Säily et al. 2017). Notwithstanding this diversity, historical sociolinguists all share a general issue at the start of a research enterprise: the key matter of the required historical data. As historical sociolinguists we have to decide what our data are and determine from which sources data can be collected. Both large historical corpora and more specialised corpora are available for English and the compilation of digital corpora for other languages has increased in recent decades (see i.a. Elspaß 2012: 162–163; Whitt 2018: 3–8).[1] Thus researchers may have a choice between either using existing digital corpora or compiling new specialised

[1] Examples of large corpora for English are the *Helsinki Corpus of English Texts*, *CEEC/ PCEEC ((Parsed) Corpus of Early English Correspondence)* and *ARCHER (A Representative Corpus of Historical English Registers)*; an example of a more specialised corpus is the *CED (Corpus of English Dialogues 1560–1760)*.

ones. Moreover, in a world of big data optimism, the idea may arise that the time-consuming compilation of specialised corpora is no longer needed. Instead, using randomly collected big data from which patterns would automatically appear is sometimes suggested as a fast option, but does such a short-cut lead to convincing results?

It is beyond the scope of this article to discuss the problems and challenges of big datasets.[2] Nor will I go into the practice of using existing historical corpora, but will rather concentrate on cases that required the compilation of new specialised corpora. Dealing with the data-related issues, I will focus on the crucial relationship between specific research questions and appropriate historical data.[3] This methodological issue will be illustrated and elaborated by concentrating on three research programmes conducted at Leiden University as part of the historical sociolinguistic line of research I initiated for the history of Dutch. Two of these projects, *Letters as loot: Towards a non-standard view on the history of Dutch* (2008–2013) and *Going Dutch: The construction of Dutch in policy, practice and discourse* (2013–2018) have been successfully completed; the third project *Pardon my French? Dutch-French language contact in the Netherlands, 1500–1900* (2018–2023) is still in progress.[4] These research programmes will be briefly introduced in Section 2 before I discuss the approaches, research questions and corpora of the three large-scale projects in Sections 3 to 8. In Section 3 the language history from below approach and an exceptional source of ego-documents will be described. Section 4 goes into the details of the compilation of the *Letters as loot* corpus and the problems that had to be overcome. Section 5 presents some illustrative phenomena and reflects on the *Letters as loot* results. The relationship between the *Going Dutch* research questions and a newly created multi-genre corpus is the topic of Sections 6 and 7. Section 8 explains the design of the third corpus which was compiled in order to establish and describe the influence of French on the Dutch language over a period of four centuries. Reflections on the challenge of historical data and final conclusions are presented in Section 9.

2. Introduction of three research programmes

The first project, *Letters as loot*, concentrated on confiscated 17th- and 18th-century private letters, sent by people from various walks of life. These letters, which I explored as programme leader and supervisor together with members of my team (post-doc Gijsbert Rutten and PhD students Judith Nobels and Tanja Simons), miraculously survived the dangers of Anglo-Dutch wars. The rediscovery of this archival source of letters allowed us to examine the language of middle and lower class people, both men and women, and consequently to obtain a non-standard view of the history of Dutch. Non-standard as

[2] See i.a. Hiltunen, McVeigh & Säily (2017) for a detailed publication on the problems and challenges of big data.

[3] I also dealt with this topic in the invited lecture *The challenge of 'bad data': from sources and corpora to answering research questions*, presented in the bad data lectures series at the *Historical Sociolinguistics Young Researchers Forum (HSYRF)* on 6 March 2020 at the VUB / Free University of Brussels.

[4] The three programmes were funded by The Netherlands Organisation for Scientific Research (NWO).

opposed to a standard view, that is a history of language as a history of standardisation, a language history of literary authors, printed texts and upper ranks of society. What is lacking in the standard view, which has dominated histories of Dutch and other European languages for a long time, is diversity, the language of middle and lower ranks, non-literary texts, handwritten documents. This meant that blank spots needed to be filled in the Dutch language history (van der Wal 2006).

The second project, *Going Dutch*, initiated by Gijsbert Rutten (co-supervisor), in which I was involved as supervisor of PhD students Andreas Krogull and Bob Schoemaker, focused on the construction of Dutch in the period around 1800. It is in a historical context of nation building, educational laws and a national school system, when a national orthography (1804) and grammar (1805), intended for the domains of education and administration, were written and issued on behalf of the government (see Rutten 2019). This period and context determine the relevant research questions concerning the effectiveness of the language policy and the effects of the official codification on actual language use.

The third project, Gijsbert Rutten's *Pardon my French*, which includes PhD student Brenda Assendelft (supervisors Gijsbert Rutten and Marijke van der Wal, co-supervisor Rik Vosters) and post-docs Andreas Krogull and Jill Puttaert, concentrates on four centuries of Dutch-French language contact. *Language choice* and *language change* are the main focal points of this research on French influence.

3. Language history from below and an exceptional source

The *Letters as loot* research programme is characterised by both the choice of what is known as the *language history from below* approach and its focus on an extraordinary source of handwritten ego-documents. The standard view of language was heavily criticized in the first decade of the 21st century, a turning point which led to the view that historical linguistics should be concerned with all sorts of written data from the past, not just literary or religious texts, the language of print and the language use of the elites (Watts & Trudgill 2002; Elspaß 2007). The *Letters as loot* programme was embedded in the similar awareness of the linguistic diversity of the past versus formerly assumed uniformity as a result of standardisation. *Letters as loot*, therefore, became an example of the *language history from below* approach, the term used for an alternative language history, characterised by a switch from the focus on the language use of experienced writers from the upper classes to the language use of the lower ranks of society, both men and women.[5] For an alternative history of Dutch *ego-documents* (handwritten private letters, diaries and travelogues) would be perfect material, but ego-documents from women in general and from both men and women from middle and lower ranks were available only in very small numbers, scattered over various municipal and provincial archives in the Netherlands (van der Wal 2006; van der Wal & Rutten 2013). Outside the

[5] Stephan Elspaß coined the term *language history from below / Sprachgeschichte von unten* at the HiSoN conference at the University of Bristol in 2005. What this approach implied was shown in great detail in his research of 19th-century German migrant letters (Elspaß 2005).

Netherlands, however, the British National Archives (Kew/London) contained an extraordinary source of documents, referred to as the *Prize Papers*. This source comprises a linguistic treasure trove of ca. 40,000 confiscated Dutch letters, including ca. 15,000 private letters, sent from the Netherlands to such areas as Asia, Africa, the Caribbean region and vice versa.

The *Prize Papers* survived due to warfare and privateering, the longstanding legitimised activity of conquering enemy ships. In the past, this activity was engaged in by all seafaring European countries and regulated by strict rules. When a ship was taken, all papers on board, including mail bags and the crew's private papers, were confiscated in order to demonstrate that the seized ship indeed belonged to the enemy and was a *legal prize*. In England, after the legal procedure, the confiscated papers remained in the High Court of Admiralty's Archives. As a result of the frequent warfare between England and the Netherlands from the second half of the 17th to the early 19th century, approximately 40,000 Dutch letters, both private and commercial, and a wide range of other material are currently stored in over a thousand boxes in the National Archives (van der Wal 2006; van der Wal, Rutten & Simons 2012; Rutten & van der Wal 2014: 1–2). The confiscated private letters appeared to be a unique and substantial source for the language history from below approach in *Letters as loot*, as they were sent by men and women from various social ranks, in particular lovers, spouses, parents, children, other relatives and friends who needed to communicate with their beloved ones at a geographical distance.

The research questions of *Letters as loot* were aimed at tracing variation in 17th- and 18th-century Dutch and revealing the course of language changes. The selection of material from the huge archive and the actual corpus compilation for linguistic research depended on both these research questions and the limitations of the source material. Two cross-sections were made with a deliberately chosen interval of about one hundred years, which would give us the opportunity to examine what had changed over roughly a century. As letter material from the First Anglo-Dutch war is not abundant, the first cross-section was made from the prelude to the Second Anglo-Dutch War to the end of the Third, based on the large number of letters available from the 1664–1674 period. The second cross-section selected was 1776–1784, from the start of the Fourth Anglo-Dutch War to the end of the American War of Independence (see Table 1; the figures in bold indicate the periods of the selected letters).

Table 1: Chronology of war periods with cross-sections made

1st Anglo-Dutch War	1652–1654
2nd Anglo-Dutch War	**1664/1665–1667**
3rd Anglo-Dutch War	**1672–1674**
War of the Austrian Succession	1739–1748
Seven Years' War	1756–1763
4th Anglo-Dutch War & American War of Independence	**1776–1784**
Napoleonic period	1793–1813

Once that decision had been taken, the time-consuming groundwork of selecting letters from the collection in the National Archives and taking digital photographs began,

followed by the second phase of transcribing the letters. Initial transcriptions of the letters were made by student assistants and volunteer transcribers of *Wikiscripta Neerlandica*, a project I initiated as a kind of early crowd-sourcing enterprise. This was followed by correction stages and the collection of the metadata related to the letter, the sender and the addressee. The result was a tri-partite collection of data which consisted of photographs, transcriptions and metadata. From this collection, a balanced corpus was compiled, for which the number of words per individual writer was limited (and not all transcribed letters were used). The ultimate corpus consisted of 933 letters, 716 different writers and around 425,000 words.

4. Choices of variables and the literacy problem

In order to trace the 17th- and 18th-century variations, the Labovian variables *gender, social class*, *age* and *region* were chosen. Of these variables, gender is relatively unproblematic, since, in almost all cases, the gender (male or female) of the sender can be deduced from his or her name in the letter. Age (<30, 30–50, >50) is more difficult to determine and often archival research was needed. The regional origin of the senders can be traced with some effort (for the regions distinguished, see Rutten & van der Wal 2014: 10–13). Social class, however, was more problematic, as social structures differ diachronically. Relying on research by social historians, we adopted four social ranks out of six for the 17th and 18th centuries (see the ranks 2, 3, 4, and 5 in Table 2).[6] The highest rank of noble and non-noble ruling classes is hardly present in the confiscated material, nor is the lowest rank of have-nots and beggars.

Table 2: Social stratification in the 17th and 18th centuries and *Letters as loot*

Social ranks	***Letters as loot* ranks**
1 Nobility and the non-noble ruling classes such as the "regenten" of cities	----
2 Bourgeoisie, e.g. wealthy merchants, ship owners, academics, commissioned officers	Upper Class (UC)
3 Prosperous middle class, e.g. large storekeepers, non-commissioned officers, well-to-do farmers	Upper-Middle Class (UMC)
4 Petty bourgeoisie, e.g. petty shopkeepers, small craftsmen, minor officials	Lower-Middle Class (LMC)
5 Mass of wage workers, e.g. sailors, servants, soldiers	Lower Class (LC)
6 Have-nots, e.g. tramps, beggars, disabled	----

In practice, to assign specific letter writers to social ranks, we used a variety of criteria, the most important being the male writer's profession or, in the case of women, the occupation of their father or their husband.

[6] This choice reflects Nevalainen's opinion that the best historical sociolinguists can do is to draw on models that have met with a certain level of consensus among social historians and historical sociologists (Nevalainen 1999: 503). See for reconstructing social stratification also Nevalainen & Raumolin-Brunberg (2017: 133–137).

I would like to stress that the precise labels for the senders of the letters (their age, social class, region) often had to be determined through elaborate genealogical and archival research in registers of baptism, marriage and deaths. Apart from social class, a further problem related to the historical context still had to be resolved: the problem of literacy and illiteracy. Although the literacy rate was relatively high in the Netherlands compared to other European countries, in the Early Modern period there were still quite a few people who were either illiterate or partly literate.[7] The semi-literates were able to read, but could not write. For the *Letters as loot* research, this meant that the crucial question of the autograph or non-autograph status of the letters had to be resolved. If the letters were not self-written by the senders, clearly, the sender information (gender, social class, age) could not be used for our research. The Leiden Identification Procedure (LIP) was developed to determine whether a letter was an autograph or a non-autograph, and in some cases letters had to be labelled as "uncertain" (for details of this procedure, see Nobels & van der Wal 2012; Nobels 2013: 53–76). These results had consequences for the compilation of the *Letters as loot* corpus.

Table 3: The *Letters as loot* (*LAL*) Corpus

Period	Letters (Autographs)	Writers	Words
1660s–1670s	549 (260)	424 (202)	228,000 (118,000)
1770s–1780s	384 (384)	292 (292)	196,500 (196,500)
Total	933	716	424,500

In the 18th century the literacy rate had increased. For our late-18th century corpus, this meant that we were able to compile a substantial corpus of only autographs, that is, 384 letters, 292 different writers, 196,500 words (see Table 3). For the 17th century, we also had a substantial corpus, but only part of it comprised autographs: 260 autographs, 202 writers, 118,000 words (see the numbers between brackets in Table 3). These autographs were appropriate material for exploring gender, social class, age and regional variation, but the question remained of how to deal with the illiteracy or semi-literacy reflected in the non-autographs and letters of uncertain status. The non-autographs and uncertain letters could still be useful for the region variable, as we may safely assume that people who were not able to write a letter themselves sought the help of a delegated letter writer in their own neighbourhood.[8] With a few exceptions, therefore, the region of the sender and the actual, often unknown writer of the non-autograph or letter with uncertain status are identical.[9] In sum, a digital corpus of 933 private letters was compiled, all transcribed from the original manuscripts and ready for the analysis of their linguistic phenomena.

[7] Higher literacy rates were found in the northern Netherlands and the Scandinavian regions, for instance, than in the southern Netherlands and Romance countries, and the literacy rate for men was generally higher than that for women (cf. Graff 1987: 173–248; van der Wal 2006; Houston 2013: 144–146, 169).

[8] For detailed research on 17th-century delegated writing of Dutch letters see van der Wal (2021).

[9] This conclusion could be safely drawn for non-autographs or letters of unclear status sent from the Netherlands, but not for such letters sent from abroad. In the latter case, the delegated writer, e.g. a literate aboard the

5. Reflecting on some illustrative phenomena and *Letters as loot* results

A particular source or particular data may have consequences for the selection of the linguistic phenomena to be analysed. In the *Letters as loot* research we selected both linguistic phenomena which are characteristic for letters, such as epistolary formulae, and phonological, morphological, morphosyntactic and syntactic phenomena which had been the topic of linguistic debate. The phenomena of epistolary formulae, H-dropping, apocope, diminutives, clause chaining, relative clauses, forms of address, negation, genitive and alternative constructions were frequent in our corpus and thus allowed quantitative research for the various variables. For full details of our results, I refer to the *Letters as loot* monograph (Rutten & van der Wal 2014). Here, I will discuss a few examples to illustrate our achievements and to answer the questions of whether the deliberately compiled corpus allowed us to answer our research questions and whether our analyses offered new insights into variation and language change.

The first example is an instance of variation, namely variation of the epistolary formula *Ik laat u weten dat* 'I let you know that…', a text-structural formula which (after an introductory part of the letter) initiates discourse or indicates a change of topic.[10] Figure 1 clearly shows that lower ranks in the 17th century use more formulae than higher ranks, and women use more formulae than men.

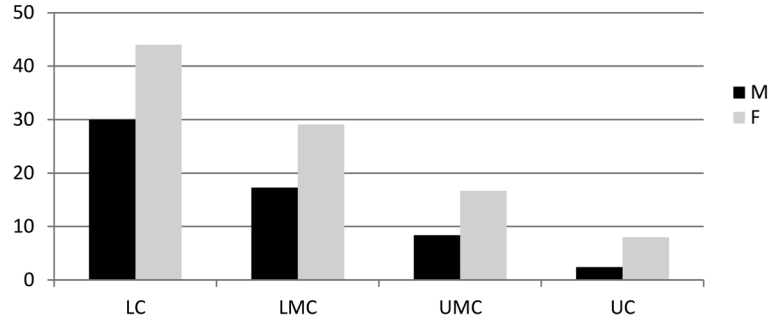

Figure 1: Frequency per 10,000 words, by social class and by gender of *ik laat u weten dat* 'I let you know that' in the seventeenth century; N=138 (Rutten & van der Wal 2014: 156)

Comparing the 18th-century with the 17th-century data, a decrease in the epistolary formula becomes clear, but the social class variation is still present in the 18th century. Lower ranks use more formulae than upper ranks; the highest rank even shows no formula at all in this case (see Figure 2).

ship, may not have originated from the same region as the sender of the letter. In that case, the letters received the regional code *Unknown* (Nobels 2013: 31–33).

[10] This formula is also found in both English and German letters of the Early and Late Modern period (Austin 1973:16; Elspaß 2005:165, 168–170), and in Finnish letters from the nineteenth century (Laitinen & Nordlund 2012: 69).

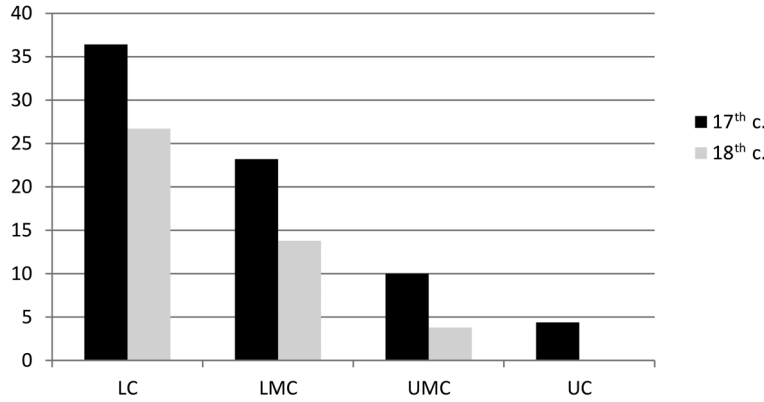

Figure 2: Frequency per 10,000 words, by social class of *ik laat u weten dat* 'I let you know that' in the seventeenth and eighteenth centuries; 17th c. N=138, 18th c. N=94 (Rutten & van der Wal 2014: 158)

Similar findings of variation and change were identified for other formulae and these had to be explained. The explanation for this social class and gender variation is their *different writing experience*. Fixed epistolary formulae were a help for less experienced writers in formulating a letter. These less-experienced writers, who did not use writing on a daily basis or in their profession, were women and lower class people who were therefore most in need of formulae.

Another example of change is what appeared to be a language change from above in the case of the epistolary forms of address. In the letters we find two epistolary forms of address: *ul* (*uwe liefde* 'your love') and the rising form *ue* (*uwe edelheid* 'your honour'). In the 17th-century graph (Figure 3), the new form *ue* increases in the Upper and Upper Middle Classes at the expense of *ul* and just a minor start can be noticed in the Lower Middle Class.

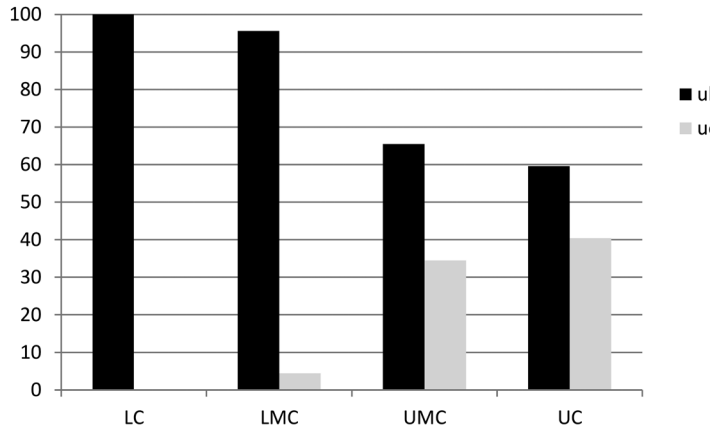

Figure 3: The proportion of *ul* and *ue* within the category of epistolary forms of address, by social class, in the seventeenth century N=1,606 (Rutten & van der Wal 2014: 228)

A hundred years later, the new *ue* is common in all social classes and the original *ul* is a minor variant in the Lower and Lower-Middle classes (see Figure 4).

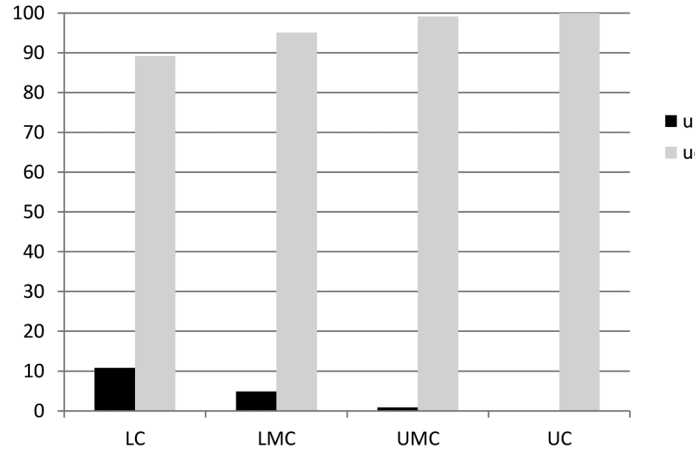

Figure 4: The proportion of *ul* and *ue* within the category of epistolary forms of address, by social class, in the eighteenth century N=4,110 (Rutten & van der Wal 2014: 233)

We may conclude that the rise of *ue* was a convincing change from above, a change which began in the higher social classes. In this case, the course of a specific language change was revealed. This change was also gender determined, as it started in the language of men from the higher ranks (see Rutten & van der Wal 2014: 227–236).

The stage of time-consuming groundwork, which consisted of taking digital photographs, transcribing letters, conducting genealogical and archival research, collecting metadata and creating a database resulted in the balanced *Letters as loot* corpus, ready for linguistic research. After having illustrated our linguistic research, the question of whether the whole enterprise was worth the effort has to be answered. The answer is positive: with the time-consuming preparations and our historical sociolinguistic analyses, we achieved our aims of a language history from below approach. We traced gender variation, social class variation, limited age variation, and regional variation, and we were able to describe the course of change diachronically, socially and regionally (see Rutten & van der Wal 2014; Nobels 2013; Simons 2013). The *Letters as loot* programme also produced the POS-tagged and lemmatised *Letters as loot* corpus, made available online for linguists, historians and the general public with text access, photos and advanced search facilities (see <https://brievenalsbuit.ivdnt.org>) and thus creating new research opportunities for others.[11]

[11] As a late spin-off, my additional *LAL-2* collection of another 1,386 transcribed letters was published online on 26 February 2021 at <https://brievenalsbuit2.ivdnt.org>.

6. Compiling a multi-genre corpus

The sociohistorical field of research is in a state of constant flux and the results of the language history from below approach allow a further step to be taken. Comparing variational usage with the tradition of norms and prescriptions is currently high on the agenda in various language areas. An illustrative example of this debate is the volume on *Norms and Usage in Language History, 1600–1900* (Rutten, Vosters & Vandenbussche 2014), which covers Dutch, English, French and German language histories. Actually, it is an old issue of the influence of language norms and language policy which has often been discussed, but hardly researched. That is also what the *Going Dutch* research programme was created for: establishing the influence of language policy and determining what impact the linguistic regulation had on language usage in the Netherlands. I will not discuss the language policy itself here, but will concentrate on the effects of language norms on language usage.[12] Did the norms intended for administration and education have influence beyond these domains? Self-evidently, first of all the question arises of how to examine the influence of language norms, in other words, what historical data are needed and what digital corpus design would be required for these different research questions and other chronological period.

In order to examine the normative influence, the 1770–1790 and 1820–1840 periods were chosen, that is, the generation *before* and *after* the official codification (1804/1805). Large-scale usage data were needed for these two time periods, that is, the *Going Dutch Corpus (GDC)* had to be compiled. On the assumption that genre mattered, a *multi-genre approach* was chosen and data from three genres were collected: in line with the language history from below approach the main sources were handwritten ego-documents, that is private letters (1) and diaries and travelogues (2), supplemented by newspapers as a third genre from the category of printed and published texts. There are some similarities with the *LAL* corpus: the choice of two periods, the manual transcriptions from original archive sources, and the size of 420,000 words. The *GDC* did not need to be compiled from scratch, as photos from the Delpher newspaper site (see <https://www.delpher.nl>), which consists of newspapers from the 17th to the 20th centuries, could be used and both photos and transcriptions of 18th-century private letters could be partly taken from the *LAL* corpus. All other ego-documents had to be collected from Dutch archives, which entailed the selection of documents and taking digital photos. During the following stage, student assistants made transcriptions of all the material collected and these transcriptions were enriched with a basic set of metadata and tags. Correction stages followed, and again, the corpus compilation, conducted mainly by PhD student Andreas Krogull, was a time-consuming enterprise.

Apart from time period and genre, the structure of the *GDC* depended on yet another variable: namely *region*. Seven regions of the Northern Low Countries were taken into

[12] For the language policy part, I refer to Schoemaker (2018) and Rutten (2019). The latter publication, Rutten's monograph *Language Planning as Nation Building* offers a detailed description of the *construction of Dutch* in policy, practice and discourse in the 1750–1850 period.

account in order to determine the effects of the national language policy (see Krogull 2018: 62–65 for the precise regions). Summarising, *time, genre* and *region* are the main variables, directly related to the *Going Dutch* research questions. The reader may wonder why the Labovian variables *gender*, *age*, *social class*, which played a major role in the *Letters as loot* programme, are absent. First of all, if we wanted to use these variables, they could be applied only for ego-documents, not for newspapers. As the ego-documents were written by men and women, gender variation could be examined, and it was, but age and social class are not variables in *Going Dutch* due to the limitations of the available material. It did not appear feasible to collect enough ego-documents written by people from different ages and different social classes for all regions and the two periods. Furthermore, diaries and travelogues were written mainly by upper-middle and upper-class writers. A similar selection, therefore, was made for the private letters, which resulted in a more or less homogeneous social class of writers. The ultimate structure of the *GDC*, which consists of three genres, each with two periods and seven regions, and two genders in the case of ego-documents, is shown in Figure 5.

Private letters (approx. 210,000 words)				**Diaries and travelogues** (approx. 140,000 words)				**Newspapers** (approx. 70,000 words)	
1770–1790		**1820–1840**		**1770–1790**		**1820–1840**		**1770–1790**	**1820–1840**
7 regions		**7 regions**		**7 regions**		**7 regions**		**7 regions**	**7 regions**
♂	♀	♂	♀	♂	♀	♂	♀		

Figure 5: The *Going Dutch Corpus (GDC)* (Krogull 2018: 48)

Looking at Figure 5, the different sizes of the three genres are striking: newspapers 70,000 words (5,000 per region); diaries and travelogues 140,000 words (10,000 per region); private letters 210,000 words (15,000 per region). These deliberately chosen differences are related to our knowledge of genre-dependent variation differences. It is known from previous research that printed newspapers show less variation than handwritten ego-documents. For newspapers, therefore, 5,000 words suffice to obtain a view of the regional variation in a particular time period. Private letters on the one hand and personal diaries and travelogues on the other also differ in variation. Private letters show more variation than diaries and travelogues. To cover this variation more words per region were needed for private letters than for diaries and travelogues.

7. The influence of norms and regulations

With the *GDC* ready, the actual linguistic analyses could start after suitable linguistic phenomena were selected. As we want to assess the influence of Matthijs Siegenbeek's orthography (1804) and Petrus Weiland's grammar (1805), this implies a selection of both orthographic variables (consonantal and vocalic) and morphosyntactic variables such as relative pronouns and genitive case. To be absolutely sure that we measured the influence of the 1804/1805 written language regulation, the previous 18th-century normative tradition also had to be taken into account and, therefore, a corpus of 18th-century normative

publications was compiled (Krogull 2018: 70–72). In summary, for each chosen phenomenon, language use in the late 18th-century period and in the early 19th-century period was analysed and compared with two kinds of norms: the 18th-century normative tradition and the Siegenbeek-Weiland regulation.

For detailed results I refer to Krogull (2018) and Rutten (2019: 243–267). Here we focus on the general question of normative influence and the role of the various variables. The influence of the Siegenbeek-Weiland regulation was indeed established: drastic changes in the direction of Siegenbeek's spelling prescriptions and a less obvious impact of Weiland's grammar were identified (see Krogull 2018: 295–308). What was also found was a *genre-specific* spreading of change, in particular of spelling norms. Newspapers adopted the spelling prescriptions almost completely. Ego-documents also adopted the prescriptions, but to a lesser extent, as the ego-documents also showed deviating spelling instances. And comparing private letters and diaries and travelogues, more deviation and variation were found in the private letters than in the diaries and travelogues. In terms of the variable *region*, 18th-century regional differences appeared to largely level out in the 19th century, after the written language regulation, which shows the successful spread of a uniform variety of Dutch across the Netherlands. *Gender* variation is present in 18th-century ego-documents. In the 19th-century period the prescribed variants increased considerably, but a minor gender difference was still found: more prescribed variants occurred in ego-documents written by men.

Both *Going Dutch* and *Letters as loot* were characterised by different research questions related to current historical sociolinguistic debates and reflecting the kaleidoscopic character of historical sociolinguistics. The two large-scale projects share similarities in corpus compilation and the attention paid to the historical context, and both projects led to an inclusive history of Dutch, that is including various types of variation. Nonetheless, still more viewpoints on the history of Dutch (or of any other language) are possible and one of them is strongly related to present-day interest in our multilingual everyday world (cf. Rutten et al. 2017). Another point of view, therefore, is the view of Dutch language history not as a monolingual, but as a multilingual past. Exploring multilingual practices of the past is a highly interesting and promising avenue in the third project, *Pardon my French*.

8. *Pardon my French*: the impact of a multilingual society

As already indicated in Section 2, *Pardon my French* concentrates on four centuries of Dutch-French language contact with language choice and language change as the main focal points. Here we focus on language change and the related research question of what happened to the Dutch language. To answer this question, a corpus-based, quantitative, diachronic analysis has to be conducted of the actual influence of French on Dutch. Also in this case, decisions had to be made on data material, corpus design, and variables.

When looking for sources of data, we have to consider the two main factors of French influence in the Northern Netherlands: migration and the Frenchification of the Dutch elite (cf. Frijhoff 2015). One aspect that is different from the two previous corpora is the

long diachronic period of four centuries (1500–1900), for which appropriate data had to be found. These data were found in the town of Leiden, a city with substantial migration of French-speaking individuals throughout history and where a 'Frenchified' elite lived. As Dutch-French language contact was thus guaranteed, Leiden was a suitable choice as a testing ground for the influence of French on Dutch and for obtaining a view of multilingualism in society. The assumption was made in advance that language choice and language contact differed in various domains of society. *Domain*, therefore, became a main variable in the corpus design of the *Language of Leiden* (*LOL*) Corpus.

Table 4: Overview of the *Language of Leiden (LOL)* Corpus (N.A.= not applicable) (Assendelft, Rutten & van der Wal to appear)

Domain	**Public opinion**	**Private**	**Academic**	**Religion**	**Literature**	**Charity**	**Economy**
Genre	Newspaper articles	Letters	Minutes	Minutes	Plays	Wills	Ordinances Requests
1500–1549	N.A.	-	N.A.	-	-	5,027	5,072
1550–1599	N.A.	4,449	5,046	5,305	5,116	5,229	5,118
1600–1649	N.A.	5,114	5,124	5,259	5,138	5,131	5,276
1650–1699	5,053	5,032	5,177	5,128	5,143	5,111	5,314
1700–1749	5,111	5,421	5,025	5,153	5,183	5,082	5,189
1750–1799	5,095	5,116	5,067	5,128	5,112	5,290	5,212
1800–1849	5,084	5,145	5,160	5,258	5,173	5,114	5,100
1850–1899	5,088	5,038	5,157	5,271	5,194	5,037	5,052
	25,431	**35,315**	**35,756**	**36,502**	**36,059**	**41,021**	**41,333**
Total word count: 251,417							

Table 4 shows the seven social domains that were distinguished: public opinion, the private and academic domains and the domains of religion, literature, charity and economy. The selected genre of documents was based on the availability of materials in the archives. Public opinion is represented by newspaper articles from Leiden-based newspapers. It should be noted that newspapers were not published before the second half of the 17th century; hence the N.A. (= not applicable) remark in the earlier cells. Private letters were used for the private domain, minutes of board meetings of Leiden University, founded in 1575, for the academic domain, minutes of church council meetings of a number of Protestant churches for the domain of religion and theatre plays for the domain of literature. The charity domain featured wills, in which individual Leiden citizens left bequests to charity organizations. The economic domain is represented by ordonnances from the city council concerning economic activity in the city, as well as requests from labourers and companies written to the city council. The 1500–1900 period was divided into eight periods of 50 years and for each cell about 5,000 words were collected. Lexical, morphological and morpho-syntactical phenomena (loan words, suffixes and relativisers) are chosen for the research on linguistic change.[13]

[13] See Rutten, Vosters & van der Wal (2015) for a first exploration of French loan suffixes and Assendelft, Rutten & van der Wal (to appear) for an analysis of French loan suffixes in the *LOL* corpus.

Pardon my French is a work in progress and detailed results cannot yet be discussed here, but I would like to add that Brenda Assendelft's *Language of Leiden* (*LOL*) corpus was compiled, with the help of student assistants, following the same method as in previous projects, namely the method of transcribing from original archive sources, both handwritten and printed. In sum, the current *Pardon my French* project also entailed a similar time-consuming procedure of compiling a specialised corpus, a corpus determined by the theoretical approach, and tailor-made for answering the relevant research questions. Moreover, the first preliminary results of the analysis of French loan suffixes in the *LOL* corpus point to the relevance of the distinguished domains, which differ from the high frequencies of such suffixes found in the domains of academy and charity to low proportions in those of literature and private life.

9. Discussion and conclusions

In the three research programmes, we have explored new grounds for the language history of Dutch. *Letters as loot* was the first extensive sociolinguistic analysis of confiscated Dutch private letters from the late 17th and 18th centuries. *Going Dutch* was the first study of the actual influence of language norms on Dutch language use in the 1750–1850 period and the *Pardon my French* corpus (*LOL*) is ready for the first corpus-based, quantitative, diachronic analysis of the actual influence of French on Dutch over a period of four centuries. We have seen how shifting perspectives and different historical contexts foregrounded different research questions and different linguistic phenomena. I have reflected on our approaches, embedded in historical sociolinguistic debates, our methods and the problems we had to overcome. Final conclusions will now follow, in particular for the issue of corpus compilation.

First of all, what I have demonstrated is that each project required its own specialised corpus. The diversity of research questions is reflected in these specialised corpora. Secondly, the importance of a convincing corpus structure, dependent on the research questions involved, was clearly shown. Thirdly, back to the sources is a leading theme, as the corpora under discussion were based on transcriptions of original sources, both in manuscript and print. This leads to the fourth conclusion that time-consuming groundwork by the research teams and with the support of student assistants and volunteers proved to be indispensable. Fifthly, we draw the general conclusion that corpus compilation is a crucial and carefully conducted enterprise, needed as a solid foundation for historical sociolinguistic research leading to reliable results.

Historical sociolinguistic research has been continuously uncovering variation, diversity and complexity, and its ultimate aim remains to understand and reconstruct the complex linguistic past. That is what took place and is still taking place in the successive Leiden research projects. In a continuing process to achieve the best results possible, the initiators and participants of these projects have shown how, within a sociohistorical approach and with specialised historical corpora, access is gained to the variational usage of the past, and fascinating issues of normative and foreign language influences are being solved.

SOURCES

Letters as Loot / Brieven als Buit Corpus. Leiden University. Compiled by Marijke van der Wal (Programme leader), Gijsbert Rutten, Judith Nobels and Tanja Simons, with the assistance of volunteers of the Leiden-based *Wikiscripta Neerlandica* transcription project, and lemmatised, tagged and provided with search facilities by the Institute for the Dutch Language (INT). First release 2013; second release 2015; third release 2021. Available online at: <https://brievenalsbuit.ivdnt.org>.

Letters as Loot-2 / Brieven als Buit-2. Leiden University. An additional collection of letters, compiled by Marijke van der Wal. First release February 2021. Available online at: <https://brievenalsbuit2.ivdnt.org>.

REFERENCES

Assendelft, B., Rutten, G. & van der Wal, M. (to appear): Tracing Frenchification: A sociolinguistic analysis of French loan suffixes in a historical corpus of Dutch. In: R. Franceschini, M. Hüning & P. Maitz (eds.), *Historische Mehrsprachigkeit: Europäische Perspektiven*. Berlin: De Gruyter.

Austin, F. (1973): Epistolary conventions in the Clift family correspondence. *English Studies* 54, 9–22.

Elspaß, S. (2005): *Sprachgeschichte von unten: Untersuchungen zum geschriebenen Alltagsdeutsch im 19. Jahrhundert*. Tübingen: Niemeyer.

Elspaß, S. (2007): A twofold view 'from below': New perspectives on language histories and language historiographies. In: S. Elspaß, N. Langer, J. Scharloth & W. Vandenbussche (eds.), *Germanic Language Histories 'from Below' (1700–2000)*. Berlin, New York: Walter de Gruyter, 3–9.

Elspaß, S. (2012): The use of private letters and diaries in sociolinguistic investigation. In: J. M. Hernández-Campoy & J. C. Conde-Silvestre (eds.), *Handbook of Historical Sociolinguistics*. Malden, Oxford: Wiley-Blackwell, 156–169.

Frijhoff, W. (2015): Multilingualism and the challenge of Frenchification in the Early Modern Dutch Republic. In: C. Peersman, G. Rutten & R. Vosters (eds.), *Past, Present and Future of a Language Border: Germanic-Romance Encounters in the Low Countries*. Berlin: De Gruyter, 115–140.

Graff, H. J. (1987): *The Legacies of Literacy: Continuities and Contradictions in Western Culture and Society*. Bloomington: Indiana University Press.

Hernández-Campoy, J. M. & Conde-Silvestre, J. C. (eds.) (2012): *The Handbook of Historical Sociolinguistics*. Malden, Oxford: Wiley-Blackwell.

Hiltunen, T., McVeigh, J. & Säily, T. (eds.) (2017): *Big and Rich Data in English Corpus Linguistics: Methods and Explorations* (= Studies in Variation, Contact and Change in English, 19). Helsinki: University of Helsinki.

Houston, R. A. (2013): *Literacy in Early Modern Europe: Culture and Education 1500–1800*. 2nd edition. London, New York: Routledge.

Krogull, A. (2018): *Policy versus Practice: Language Variation and Change in Eighteenth- and Nineteenth-Century Dutch*. University of Leiden dissertation. Utrecht: LOT.

Laitinen, L. & Nordlund, T. (2012): Performing identities and interaction through epistolary formulae. In: M. Dossena & G. Del Lungo Camiciotti (eds.), *Letter Writing in Late Modern Europe*. Amsterdam, Philadelphia: John Benjamins, 65–88.

Nevalainen, T. (1999): Making the best use of 'bad' data: Evidence for sociolinguistic variation in Early Modern English. *Neuphilologische Mitteilungen* 100 (4), 499–533.

Nevalainen, T. & Raumolin-Brunberg, H. (2012): Historical sociolinguistics: Origins, motivations, and paradigms. In: J. M. Hernández-Campoy & J. C. Conde-Silvestre (eds.), *The Handbook of Historical Sociolinguistics*. Malden, Oxford: Wiley-Blackwell, 22–40.

Nevalainen, T. & Raumolin-Brunberg, H. (2017): *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. 2nd edition. London, New York: Routledge.

Nobels, J. (2013): *(Extra)Ordinary Letters: A View from Below on Seventeenth-Century Dutch*. University of Leiden dissertation. Utrecht: LOT.

Nobels, J. & van der Wal, M. (2012): Linking words to writers: Building a reliable corpus for historical sociolinguistic research. In: N. Langer, S. Davies & W. Vandenbussche (eds), *Language and History, Linguistics and Historiography: Interdisciplinary Approaches*. Oxford, Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Wien: Peter Lang, 343–361.

Rutten, G. (2019): *Language Planning as Nation Building: Ideology, Policy and Implementation in the Netherlands, 1750–1850*. Amsterdam, Philadelphia: John Benjamins. Available online at: <https://doi.org/10.1075/ahs.9>.

Rutten, G., Salmons, J., Vandenbussche, W. & Vosters, R. (2017): Unraveling multilingualism in times past: The interplay of language contact, language use and language planning. *Sociolinguistica* 31, 1–12.

Rutten, G., Vosters, R. & Vandenbussche, W. (eds.) (2014): *Norms and Usage in Language History, 1600–1900: A Sociolinguistic and Comparative Perspective*. Amsterdam, Philadelphia: John Benjamins.

Rutten, G., Vosters, R. & van der Wal, M. (2015): Frenchification in discourse and practice: loan morphology in Dutch private letters of the eighteenth and nineteenth centuries. In: C. Peersman, G. Rutten & R. Vosters (eds.), *Past, Present and Future of a Language Border: Germanic-Romance Encounters in the Low Countries*. Berlin: De Gruyter, 143–169.

Rutten, G. & van der Wal, M. (2014): *Letters as Loot: A Sociolinguistic Approach to Seventeenth- and Eighteenth-Century Dutch*. Amsterdam, Philadelphia: John Benjamins. Available online at: <https://doi.org/10.1075/ahs.2>.

Säily, T., Nurmi, A., Palander-Collin, M. & Auer, A. (2017): *Exploring Future Paths for Historical Sociolinguistics*. Amsterdam, Philadelphia: John Benjamins.

Schoemaker, B. (2018): *Gewijd der Jeugd, voor taal en deugd: Het onderwijs in de Nederlandse taal op de lagere school, 1750–1850* [Dutch Language Education in Primary Schools, 1750–1850]. University of Leiden dissertation. Utrecht: LOT.

Simons, T. (2013): *Ongekend 18e-eeuws Nederlands: Taalvariatie in persoonlijke brieven* [Unknown 18th-century Dutch: Language Variation in Private Letters]. University of Leiden dissertation. Utrecht: LOT.

van der Wal, M. (2006): *Onvoltooid verleden tijd: Witte vlekken in de taalgeschiedenis* [Imperfect Past Time: Blank Spots in Language History]. Inaugural lecture, 17 November 2006, Amsterdam, Koninklijke Nederlandse Academie van Wetenschappen.

van der Wal, M. (2021): The black box of delegated writing: Early Modern scribes and female literacy in The Netherlands. *Journal of Historical Sociolinguistics* 7 (2), 303–330.

van der Wal, M., Rutten, G. & Simons, T. (2012): Letters as loot: Confiscated letters filling major gaps in the history of Dutch. In: M. Dossena & G. Del Lungo Camiciotti (eds.), *Letter Writing in Late Modern Europe*. Amsterdam, Philadelphia: John Benjamins, 139–161.

van der Wal, M. & Rutten, G. (2013): Ego-documents in a historical sociolinguistic perspective. In: M. J. van der Wal & G. Rutten (eds.), *Touching the Past: Studies in the Historical Sociolinguistics of Ego-Documents*. Amsterdam, Philadelphia: John Benjamins, 1–17.

Watts, R. & Trudgill, P. (2002): *Alternative Histories of English*. London, New York: Routledge.

Whitt, R. J. (2018): Using diachronic corpora to understand the connection between genre and language change. In: R. J. Whitt (ed.), *Diachronic Corpora, Genre, and Language Change*. Amsterdam, Philadelphia: John Benjamins, 1–15.

Leiden University Centre for Linguistics (LUCL), Faculty of Humanities, Leiden University
Willem de Zwijgerlaan 1, 2341 EG Oegstgeest, The Netherlands
<m.j.van.der.wal@hum.leidenuniv.nl>