Historical corpus of Dutch: a new multi-genre corpus of Early and Late Modern Dutch

Van de Voorde, I.; Rutten, G.J.; Vosters, R.; Wal, M.J. van der; Vandenbussche, W.

Amsterdam University Press

TAAL & TONGVAL 75.1 (2023) 114-132
HTTPS://DOI.ORG/10.5117/TET2023.1.006.VAND

Historical Corpus of Dutch: A new multi-genre corpus of Early and Late Modern Dutch

Iris Van de Voorde
Vrije Universiteit Brussel & Universiteit Leiden
Iris.Van.De.Voorde@vub.be

Gijsbert Rutten
Universiteit Leiden
g.j.rutten@hum.leidenuniv.nl

Rik Vosters
Vrije Universiteit Brussel
Rik.Vosters@vub.be

Marijke van der Wal
Universiteit Leiden
m.j.van.der.wal@hum.leidenuniv.nl

Wim Vandenbussche
Vrije Universiteit Brussel
Wim.Vandenbussche@vub.be

Abstract

In this contribution, we present the Historical Corpus of Dutch (HCD), a new multi-genre, diachronic corpus of Early and Late Modern Dutch (ca. 1550-1850). It consists of a digitised collection of handwritten administrative texts (e.g. town council meeting reports), handwritten ego-documents (e.g. diaries and travelogues), and printed pamphlets (e.g. of a political or religious nature). The corpus is also balanced between northern and southern material, with data from the provinces of Holland and Zeeland for the North, and from Flanders and Brabant for the South. After having discussed its structure and composition, we will illustrate the value of the new corpus with a number of smaller case studies. Based on our experiences

with the corpus, we will conclude by launching a plea for historical corpus building not to focus too much on the quantity of data ('big data'), but rather shift attention to data quality.

Keywords: historical corpus building, corpus linguistics, history of Dutch, northern and southern Dutch, spelling of long *a*, *d*- and *w*-forms

1 Introduction

In a key publication in historical-sociolinguistic research, Elspaß (2012, 156) argues that a 'complete account of language history, viewed from the perspective of its agents, can only be achieved if we attempt to consider as many text sources from as many different times, varieties, regions, domains, and text types as possible'. In this spirit, we recently compiled the Historical Corpus of Dutch (HCD), a new multi-genre, diachronic corpus of Early and Late Modern Dutch.[1] The corpus was compiled within a project on the historical pluricentricity of Dutch. For this project we examined the emergence and evolution of supralocal written Dutch, testing the idea that innovations spread from a linguistic 'centre' to the 'periphery' (e.g. from Holland to Zeeland), and from the North of the language area (e.g. Holland and Zeeland) to the South (e.g. Brabant and Flanders) (Rutten et al. 2023). In addition, the project also aimed to remedy the relative absence of the southern Netherlands in traditional histories of Dutch (Van de Voorde 2022).

The Historical Corpus of Dutch consists of a digitised collection of handwritten administrative texts, handwritten ego-documents, and printed pamphlets. The corpus is also balanced between northern and southern materials, i.e. originating from the northern and southern Netherlands respectively. Northern data stem from the regions of Holland and Zeeland, and southern data are connected to Flanders and Brabant. In addition, the corpus is also built up around different time periods, gathering data from around the middle of the sixteenth, seventeenth, eighteenth and nineteenth centuries. We hope to expand the corpus in the future by adding data from additional regions (particularly the eastern peripheries) and genres (e.g. literary texts), and perhaps also time periods.

We will first explain the need for a new corpus in Section 2. The structure and composition of the Historical Corpus of Dutch will be discussed in Section 3. In that section we will also reflect on the compilation process, and on the different genres included in the corpus. In Section 4 we will then

proceed to illustrate the value of the new corpus with two case studies. We selected an orthographical feature (spelling of long *a* in closed syllables), alongside a morphosyntactic feature (*d*- and *w*-forms in relativisers). The case studies are focused on linguistic changes spreading through time and space, and in the case of the relativisers, we also look at the different genres. This section is followed by a discussion (Section 5), leading to a plea concerning data and corpus compilation for historical language studies (Section 6).

2 The need for a new corpus

Diachronic multi-genre corpora exist for various languages. English is well-served with the Helsinki Corpus of English Texts and A Representative Corpus of Historical English Registers (ARCHER). The Helsinki Corpus was already compiled in the 1980s. It comprises approximately 1.5 million words and 450 texts, spanning ten centuries from about 730 to 1710.[2] The texts are diverse, covering a wide range of genres, such as religious treatises, philosophical and scientific texts, travelogues, letters, sermons, and so on. The development of ARCHER began in the 1990s and is still ongoing.[3] ARCHER covers the period 1600-1999, and it comprises various genres such as advertising, sermons, journals, news, letters and diaries. Version 3.2 has circa 3.3 million words, distributed over British English (2 million words) and American English (1.3 million words). Similar diachronic multi-genre corpora have been compiled for other languages such as German, Spanish and French, where often also regional variation is built into the corpus design.[4]

Dutch is less well-served, especially when we consider the sixteenth to the nineteenth century. Many historical texts are available through websites such as the Digitale Bibliotheek voor de Nederlandse Letteren (DBNL) 'Digital Library for Dutch Literature'[5], which focuses on – but is not restricted to – literary language from the Middle Ages to the present day. Digitisation initiatives and research projects have resulted in even more online available textual data.[6] The Instituut voor de Nederlandse Taal 'Dutch Language Institute' hosts the historical dictionaries of Dutch including some of the datasets underlying these dictionaries.[7] Especially for the Medieval period, some excellent resources are available, such as the Corpus Oudnederlands 'Corpus of Old Dutch' (sixth to twelfth centuries), the Corpus Gysseling (thirteenth century), the Corpus Van Reenen-Mulder (fourteenth century) and the Corpus Middelnederlands 'Corpus of Middle Dutch' (thirteenth

to sixteenth centuries). For the Early and Late Modern periods, we can also draw on some excellent corpora, but these are mostly focused only on northern varieties of Dutch (e.g. the socially stratified Letters as Loot Corpus contains letters from the seventeenth and eighteenth centuries originating from the coastal regions of Holland and Zeeland – Rutten and Van der Wal 2014), and built around a single genre, often drawing from already available material online (e.g. the Dutch Corpus of Contemporary and Late Modern Periodicals – Piersoul, De Troij, and Van de Velde 2021). One exception is the Compilatiecorpus Historisch Nederlands (Coussé 2010), which is in fact a diachronic multi-genre corpus. It comprises texts from Holland, Flanders and Brabant, with a component of administrative texts (1250-1799) and a component of narrative texts (1575-2000). However, the corpus is for a large part based on scans of nineteenth-century text editions which have not been checked against the original sources for transcription accuracy. Given their intended use for linguistic analysis, and given the frequent practice of making changes to orthography as well as phrasing in such older text editions, this material cannot be used uncritically.

In terms of available material, it is also worth mentioning NederLab, which is a large digital infrastructure project aiming to bring together in one search environment all freely available historical Dutch texts.[8] The material for the period of interest to us mostly originates from the DBNL, but other existing corpora and texts have also been added to the collection. However, in spite of the large amount of material involved, we must stress that NederLab was not conceived as a balanced linguistic corpus founded on (predefined) principles concerning data selection with respect to genre, period, region, and so on. Comparing genres across the ages is therefore still difficult in the case of Dutch, and the need for a reliable, balanced, multi-genre corpus, covering different centuries and regions, is what motivated us to compile the Historical Corpus of Dutch.

3 Structure and composition of the Historical Corpus of Dutch

The Historical Corpus of Dutch (HCD) was developed at the Vrije Universiteit Brussel and Leiden University. It is a new diachronic corpus with text material from four centuries, four regions, and three genres. These three dimensions are discussed below.

First of all, the corpus covers the sixteenth to the nineteenth century. Similar to corpora such as ARCHER and GerManC, we have opted to focus

first on the Early and Late Modern periods, for which it can be assumed that textual sources from various genres are sufficiently available. Textual material was chosen from around the middle of each century: 1550, 1650, 1750, and 1850. For each of these dates, a margin of 20 years before and 20 years after the date was built in in order to find sufficient sources, resulting in four time periods: 1530-1570, 1630-1670, 1730-1770, and 1830-1870. A corpus spanning several centuries allows mapping language change in real time (Nevalainen and Raumolin-Brunberg 2017, 53).

We also included a regional dimension, with textual material from four regions in the northern and southern Netherlands. The regional dimension comprises various levels. We distinguish between the northern and the southern Netherlands, roughly corresponding to the present-day Netherlands and Belgium. In our corpus, we chose the regions of Holland and Zeeland in the North, and the regions of Brabant and Flanders in the South. Note that the southern region Brabant includes the present-day Dutch-speaking provinces of Flemish Brabant and Antwerp, and that Flanders refers to the present-day provinces of East and West Flanders (so not to the entire Dutch-speaking area in Belgium). This leads to a corpus with four smaller regions that can be grouped into northern and southern regions. Furthermore, within the North and the South of the language area, Holland and Brabant can be considered as central regions, while Zeeland and Flanders occupy a more peripheral position so that the corpus can also be used to investigate centre-periphery dynamics. This dimension was also included in view of pluricentric theory. Many texts originate from larger cities such as Amsterdam, Antwerp, Middelburg, and Ghent, but smaller towns and villages (e.g. Arnemuiden, Strijpen) are also represented in the corpus.

Finally, the corpus comprises administrative texts, ego-documents, and pamphlets. The administrative texts in our corpus are handwritten, formal texts, such as town council meeting reports and resolutions. The authors of these texts were generally used to writing because of their profession. The sources for this genre were related to guilds or to industry on the one hand, and to the general administration on the other. These documents are similar because they always concern legislation or decisions made by higher authorities. We selected, for example, sources that were part of the compilation by N.W. Posthumus[9]; the original documents are kept in the archives of Erfgoed Leiden en Omstreken. When we used existing transcriptions, these were checked against the original archival materials (see below).

Ego-documents, on the other hand, are less formal, handwritten texts, assumed to be conceptually closer to the everyday vernacular (Elspaß

2012; Koch and Oesterreicher 1985). The ego-documents in our corpus are travelogues, diaries and chronicles of local events or family history. As for the travelogues and chronicles, we made sure that the events were perceived by the author himself. Documents were collected from various libraries and archives including the University Library of Amsterdam, the Zeeuws Archief, the Koninklijke Bibliotheek van België (KBR), and the University Library of Ghent.

These two handwritten genres are supplemented by a printed genre, namely pamphlets. These are published texts, mostly commentaries or polemics about current affairs, politics or religious topics. This genre also covers public ordinances and regulations. Due to the variety of documents, printed pamphlets may vary on the continuum between more and less formal. This variety reflects the heterogeneity associated with the genre of pamphlets. Most of our northern sources are part of the so-called Knuttel collection of Dutch pamphlets. This collection, assembled by W.P.C. Knuttel, comprises about 32,000 pamphlets (Van der Hoeven 1978). The pamphlets from the Knuttel collection are electronically available via Dutch Pamphlets Online.[10] For the periods and regions that were not covered by this existing collection, we found additional pamphlets in, among others, the FelixArchief (city archives Antwerp) and the University Library of Ghent.

In order to create an electronically searchable corpus, we needed transcriptions of the original texts. The texts were first collected in the form of photographs. Thanks to the increasing digitisation of archives and libraries, a considerable part of the documents (mainly administrative texts and pamphlets) had already been scanned and could be consulted online. All other documents were photographed on site in archives and libraries.

The majority of the texts were manually transcribed, using the diplomatic method. This means that we kept the original form of the texts, in that – among other things – spelling and spacing were not normalised. For this purpose, we drew up specific transcription guidelines, defining a 'header' and several 'tags'. The header of each transcription contains all the available metadata of the text (cf. Example 1).

(1) <header>
DOCUMENT: 1652_VanderVinne_07-15
ARCHIVE: GA Haarlem (a) hss verz. 172/Top. atlas no.52
GENRE: travelogue
NAME: Vincent Laurensz. van der Vinne
DATE: 1652
PLACE: Haarlem

TRANSCRIBER: ES, controle MW
NOTES:
WORD COUNT: 2028
</header>

The header in Example 1 belongs to a seventeenth-century ego-document from the region of Holland. In the header, we noted the name of the author, as well as when and where the text was produced. Note that the name of the author was only included in the header of ego-documents, since the authors of administrative texts and pamphlets were mostly unknown.

In addition, in the transcription guidelines, we defined so-called tags to be used for named entities (personal and place names), ambiguous words, illegible words, and so on. These tags were manually added while transcribing a text. Example 2 illustrates the use of 'place' tags with a sentence from the seventeenth-century ego-document presented in Example 1:

(2) *Ick vervorderde mijn reijs en ginck te voet nae* <pl>*aernem*</pl>
'I continued my journey and went by foot to <pl>Arnhem</pl>'

The full tag set and documentation will be released alongside the corpus. For the time-consuming task of transcribing, we benefited from the help of various student assistants, students and volunteers from the crowdsourcing project Wikiscripta Neerlandica II.[11] Given the difficulty of sixteenth- and seventeenth-century texts (especially in terms of handwriting), the aforementioned group mostly transcribed eighteenth- and nineteenth-century texts, which are easier in terms of readability. The more difficult texts were transcribed by a project member, or by an experienced student assistant. All transcriptions were checked by hand by a project member, who had to be someone else than the person who had made the transcription. In the case of already existing transcriptions, for example transcriptions of administrative texts that had been published in a text edition, these transcriptions were also checked by hand against the original (photographed) text. The final transcriptions were saved as plain text files, but will be made available for research purposes in TEI-compliant XML files. The whole compilation process, from photographs to final transcriptions, took about three years (2017-2020).

All in all, the corpus consists of 209 different texts, amounting to 463,248 words. More specifically, it comprises 58 administrative texts, 60 ego-documents, and 91 pamphlets. Figure 1 shows the word count per genre.

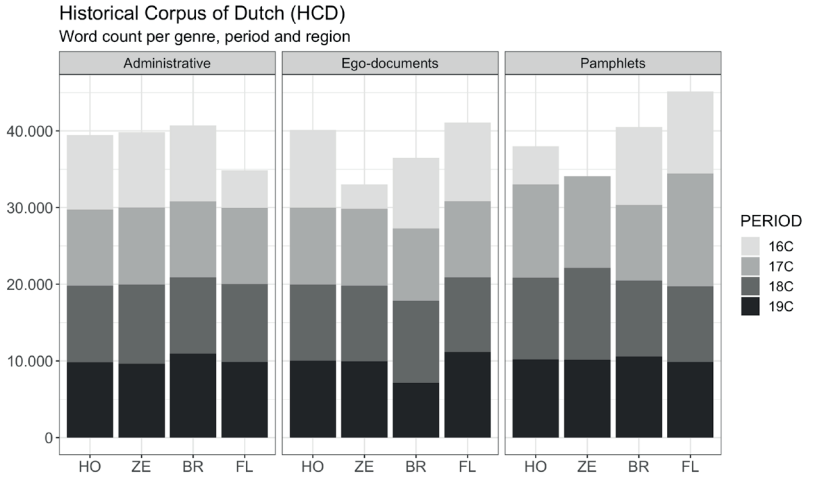

Figure 1: Visual representation word count per dimension (HO = Holland, ZE = Zeeland, BR = Brabant, FL = Flanders)

We aimed for 10,000 words per region and per period for each genre. For reasons of representativeness, these 10,000 words were preferably spread over multiple documents. For the administrative texts we aimed for two texts of 5000 words each, and for the ego-documents and pamphlets we aimed for five texts of 2000 words each. This implies that in most cases we are dealing with fragments, and not with complete texts. As far as possible, the beginning of the transcription does match the beginning of the text. The different spread of the number of texts is related to the genre itself: for the administrative component, we were able to collect very similar texts, resulting in a fairly homogeneous component. The other two genres, on the other hand, are rather heterogeneous: ego-documents due to individual differences between scribes (Nevalainen and Raumolin-Brunberg 2012, 32–3), and pamphlets due to the variety of documents that can be categorised as 'pamphlets'.

From Figure 1, we can deduce that most of the deviations from the intended 10,000 words can be found in the sixteenth century. A smaller lacuna can be noted for the nineteenth-century ego-documents from Brabant, but overall, we were able to construct a solid, well-balanced historical corpus.

4 Case studies

The case studies discussed below serve to show how the Historical Corpus of Dutch may be used to investigate variation and change in historical Dutch sources from the sixteenth to the nineteenth centuries. Our prime aim is to demonstrate the suitability of the corpus for this type of research rather than discuss the linguistic features analysed in great detail.

4.1 Spelling of long *a* in closed syllables

The first feature that we want to discuss is the orthographical representation of long *a* in closed syllables. This sound is realised as /a:/ in modern Standard Dutch. In Middle Dutch (approximately from the twelfth to the sixteenth century), the traditional practice in the areas under scrutiny was to indicate vowel length by adding an <e> (or less often an <i>) to the original vowel <a>, resulting in spellings such as *maend* ('month') and *daer* ('there') (Van der Sijs 2004, 228–29). From the seventeenth century onwards, however, this older writing practice was increasingly being replaced by the doubling of the original vowel, largely irrespective of the pronunciation of the long *a*. This resulted in <aa> rather than <ae> spellings, for example *maand* ('month') and *daar* ('there'). Note that we did not include examples of long *a* before *r*+dental (e.g. *paard* 'horse'), since these forms are derived from an original short *e* (Van Loon 2014, 225).

The shift in writing practices from <ae> to <aa> took several centuries, with the newer variant <aa> already being used in the Middle Ages, and the earlier variant <ae> being used well into the nineteenth century (see for example Puttaert 2019 for a recent historical sociolinguistic study of nineteenth-century sources).[12] The spelling <ae> even developed into a shibboleth of southern Dutch writing practice in the eighteenth and nineteenth centuries, with <aa> being taken as a northern variant (Rutten 2011, 185–89).

We searched the HCD for all occurrences of <aa>, <ae> and <ai> using regular expressions with the programming language *R*. The results of the regular expressions (one per spelling variant) were merged into one Excel file, upon which the results were manually filtered. In this stage, we removed so-called 'false positives', which include words such as *bataille* 'battle' (no long *a*), occurrences of long *a* in open syllables, and the aforementioned occurrences of long *a* before *r*+dental. After manual filtering, we retained 16,017 tokens of <ae>, 9174 tokens of <aa>, and only 182 tokens of <ai>. These results were analysed in *R*. As the spelling variant <ai> accounts for less than one per cent of the tokens in our corpus, we limit our attention to the

Figure 2: The diachronic change from <ae> to <aa>

two main variants <ae> and <aa>. Figure 2 presents the proportion of <aa> and <ae> across the centuries.

From Figure 2, we can deduce that the incoming <aa> forms first appear in the seventeenth century. At that time, <aa> was still a minority variant. The modern variant breaks through in the eighteenth century, and has become the dominant form by the nineteenth century. The new <aa> spelling thus spread with an S-curve-like pattern across time. The S-shaped curve is a model used by sociolinguists to describe 'the spread of linguistic innovations' (Nevalainen and Raumolin-Brunberg 2017, 53). The S-shape represents the rate of the linguistic change: the new form is adopted slowly at the beginning, with a rapid change in the middle stage, followed by a slower final stage (Nevalainen and Raumolin-Brunberg 2017, 53–4). In the case of the spelling of long *a*, we recognise the slow spread in the seventeenth century and the rapid change in the eighteenth century. The final stage of the curve, however, is situated later in the nineteenth century, or perhaps only in the twentieth.

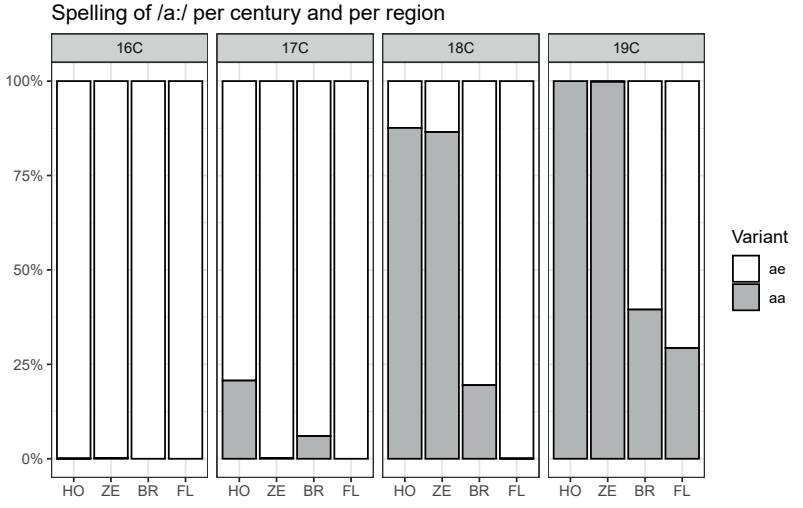

Figure 3: The change from <ae> to <aa> across centuries and regions

Splitting up the data according to century and region (Figure 3), we notice that in the seventeenth century, the change first took off in Holland, and – to a lesser extent – in Brabant. By the eighteenth century, the <aa> spellings have become the dominant forms in Holland and Zeeland, and thus have become the clear norm in the North. In the South, on the other hand, <ae> is still the dominant variant in Brabant at that time, while the new <aa> forms are practically absent in Flanders.[13] It is not until the nineteenth century that the Flemish writers in the corpus start to adopt the <aa> forms. Both in Brabant and in Flanders, however, the progressive variant <aa> is used alongside the older <ae> forms. These older forms have been almost completely lost in the northern provinces by the nineteenth century.

In other words, the innovation starts slowly in Holland and Brabant, with Holland as the leading region. From the eighteenth century onwards, however, the North and the South develop their own dynamics: the change quickly nears completion in the North, but spreads more slowly in the South, seemingly from Brabant to Flanders.

4.2 *D*- and *w*-forms in relativisers

For our second case study, we focus on a morphosyntactic change in the domain of relativisers, where forms with initial *d*- (e.g. *daar* 'there') are gradually replaced by forms with initial *w*- (e.g. *waar* 'where'). We particularly

looked at relative adverbs and relative pronominal adverbs. In the Early Middle Dutch period, mainly *d*-forms were used for this kind of relative clauses (Van der Horst 2008, 476–77). This resulted in relative clauses such as *de plaats daar wij sliepen* ('the place where we slept') and *de mand daar men mee te markt gaat* ('the basket with which one goes to the market'). During the Late Middle Dutch period, these *d*-forms were increasingly replaced by *w*-forms (Van der Horst 2008, 703–04; Van der Wal 2003). The *w*-forms in relative clauses such as *de plaats waar wij sliepen* ('the place where we slept') and *de mand waarmee men naar de markt gaat* ('the basket with which one goes to the market') have become the norm in present-day Standard Dutch. The change from *d*- to *w*-forms in relative adverbs and relative pronominal adverbs is complete in present-day Standard Dutch. It is part of a wider ongoing change in the relativisation system, where also the pronoun *dat* 'that' changes into *wat*, and where also the form *die* 'who, that' may have begun to change into *wie* (Van der Wal 2002).

In this case, too, the change already began in the medieval period, but forms of the conservative variant with initial *d*- can be found well into the nineteenth century. In the normative tradition, little attention is devoted to this issue in the Early Modern period, and it is only from the early nineteenth century onwards that more and more normative injunctions are made about *d*- and *w*-relativisers (Van der Wal 2003). The sociolinguistic profile of the variants at various points in time is not clear. De Schutter and Kloots (2000) suggest formal differences in the seventeenth century, with *w*-forms being less formal, or more likely to occur in informal contexts. Rutten and Van der Wal (2014) show that the change from *d*- to *w*-forms in seventeenth- and eighteenth-century private letters was a change from above in the social sense, i.e. with male writers and upper-ranked writers using more *w*-forms than female writers and lower-ranked writers.

Using regular expressions in *R*, we searched the corpus for all occurrences of *daar* 'there' and *waar* 'where', including spelling variants of long *a* (i.e. <ae>, <ai> and <a>). We did not include a word boundary in the regular expressions, so that besides the relative adverbs *daar* 'there' and *waar* 'where', all potential relative pronominal adverbs (e.g. *waarmee* 'with which') were also detected. We then again conducted a manual filtering in Excel and removed false positives, including occurrences of *daar* 'there' used in an expletive construction (e.g. *daar waren…* 'there were…') and occurrences of *waar* 'where' used as an interrogative adverb. We eventually retained 681 tokens of the original *d*-forms and 1012 tokens with *w*-. Figure 4 gives the proportions of *d*- and *w*-forms across the centuries.

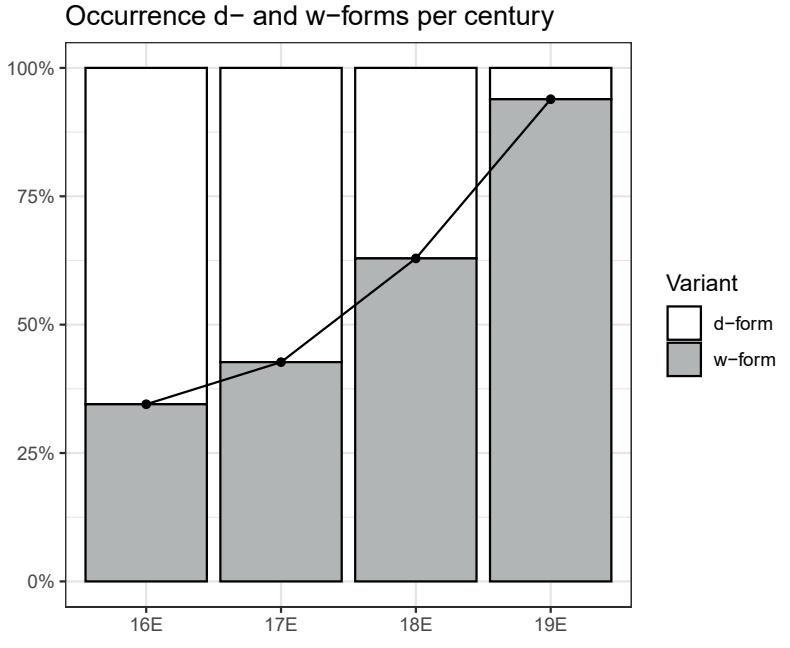

Figure 4: The diachronic change from d- to w-forms

Figure 4 shows that the shift from *d-* to *w*-forms was already ongoing in the sixteenth century. After a slight increase in the number of *w*-forms in the seventeenth century, this newer variant becomes dominant in the eighteenth century. By the nineteenth century, the change is almost complete. Unlike the spelling of long *a*, we find a partial S-curve for this feature that is almost complete, with 94 per cent *w*-forms in the nineteenth century.

When we split up our results with regard to century and region (Figure 5), it becomes clear that the *w*-forms occur in every region in the sixteenth century. Holland and especially Brabant seem to be leading the change. When we take a look at the seventeenth and eighteenth centuries, however, this leading role is not confirmed. Note, for example, the strong increase in the number of *w*-forms in Zeeland, and the slight decrease in Holland in the seventeenth century. In the eighteenth century, we again see a different picture, until usage in all regions is more or less equal in the nineteenth century. To sum up, the incoming variant does not seem to spread following a clear regional pattern.

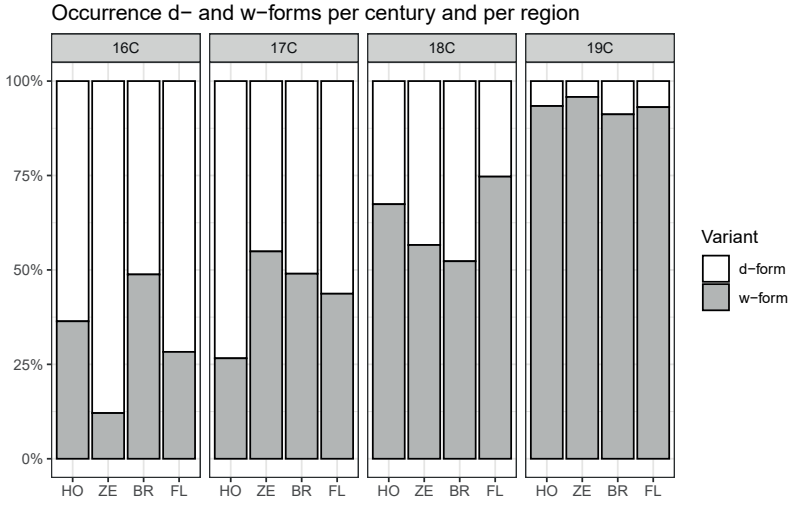

Figure 5: The change from d- to w-forms across centuries and regions

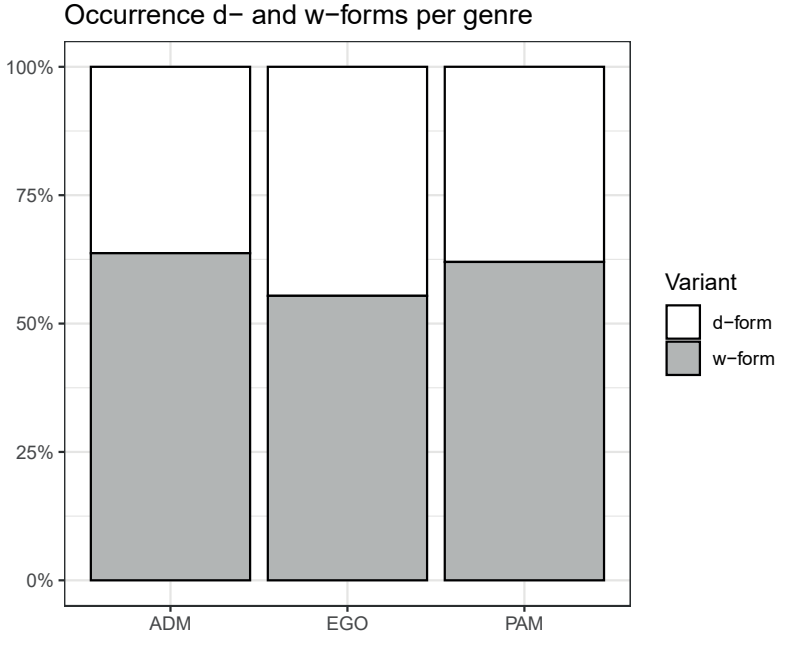

Figure 6: The change from d- to w-forms across genres

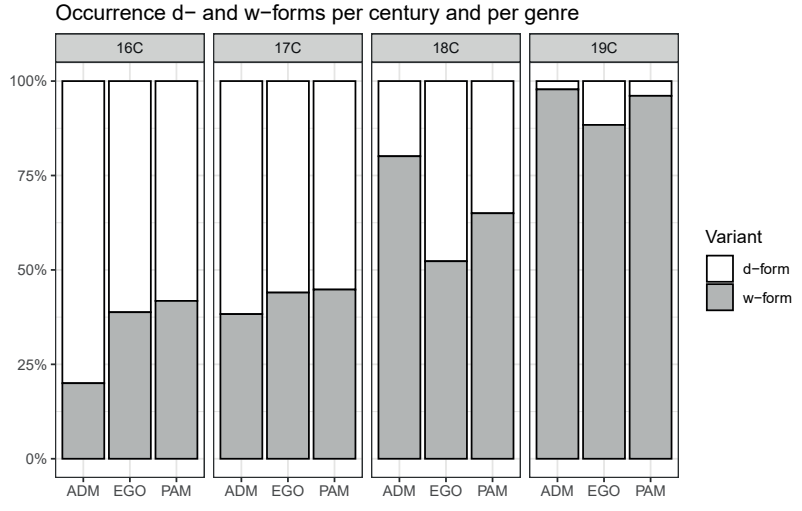

Figure 7: The change from d- to w-forms across centuries and genres

For this morphosyntactic feature, we would also like to draw attention to genre differences. Looking at the occurrence of *d*- and *w*-forms with regard to genre (Figure 6), it turns out that the administrative texts and pamphlets were slightly more progressive than the ego-documents in adopting the w-forms.

If we combine the diachronic dimension with genre (Figure 7), it becomes clear that especially in the sixteenth century, and to a lesser extent also in the seventeenth century, the administrative texts in fact displayed more *d*-forms than the other two genres. It is not until the eighteenth century that the more formal administrative documents adopt the incoming *w*-forms in large numbers. This effect is observed in many different texts and among multiple writers. By the nineteenth century, the older *d*-forms hardly occur in the administrative texts, or in the printed pamphlets. This may indicate that the *w*-forms were initially avoided in the more formal administrative documents, but when they had become the majority forms in general, in the eighteenth century, they were adopted at the highest rate in the administrative texts. Most of the *d*-forms in the nineteenth century were found in the ego-documents. Considering the varying tendencies across centuries, we could state that the shift from *d*- to *w*-forms took place at a different pace within each genre.

5 Discussion

In the previous section we illustrated how the Historical Corpus of Dutch may be used for historical-sociolinguistic research. The Historical Corpus of Dutch can help chart out changes *in real time*, for example by visualizing S-curve patterns of language change, *across regions* by investigating geographical patterns including North-South tensions (as in the case of the spelling of long *a*), or centre-periphery dynamics, and *across genres*, for example by showing how the change from *d*- to *w*-forms affects different genres at a different rate. We observed different patterns of diffusion for different features in that incoming variants did not spread following the same pattern each time. Whereas we observed a clear regional pattern for the long *a*, the regional pattern for *d*- and *w*-forms was very disparate. The spelling of long *a* is obviously a written phenomenon, although regional variation in pronunciation, for example differing degrees of palatalisation, cannot be excluded. It is as yet unknown to what extent the change from *d*- to *w*-forms was linked to the base dialects, although it may seem likely that such a morphosyntactic variable largely follows the grammar of the spoken language. The present-day distribution of *d*-forms does not show a clear regional pattern (SAND 1, map 88b). The case studies therefore also demonstrate that supposedly written phenomena such as spelling may display regional patterning, whereas morphosyntactic changes that could suggest a dialectal base may lack clear regional patterns in writing.

The case studies show that S-curves can be established for changes in Dutch, especially when considering changes in the corpus as a whole, but also for separate regions (see for example Figure 5, where each region follows its own S-like pattern). Below this highest level, however, patterns of variation and change are often less clear, or even disparate. In the case of long *a*, the results across region for the eighteenth and nineteenth centuries show a neat North-South difference in line with a pluricentric view of language history. The regional results for the relativisers are more difficult to interpret in that regions gradually show more *w*-forms, but there are no indications that one or more of the regions are leading this change, nor is it clear whether any North-South or centre-periphery dynamics play a role. At the same time, the changing role of administrative documents, i.e. from conservative in the earlier periods to progressive in the later centuries, points to genre differences that may reflect the changing status of *d*- and *w*-relativisers in terms of writing norms. At an even more microscopic level, i.e. at the level of individual writers and texts, patterns of variation and change often become even less clear (see Van de Voorde 2022, 111–20 for examples taken from the HCD).

6 Conclusions

The Historical Corpus of Dutch (HCD) was created in order to fulfil a long-standing wish, i.e. to have a diachronic multi-genre corpus of Dutch covering various regions. The HCD spans four centuries, three genres and four regions. In the future, more genres can be added in order to get a fuller picture of the genre-varietal spectrum, for example private letters, business letters, novels, plays, sermons, and so on. It is also highly desirable to include more regions in the HCD, in particular also eastern and northern regions which are still underrepresented in the datasets available for historical Dutch (e.g. Friesland, Groningen, Gelderland, Limburg). An extension back into time could also be considered, although it turned out to be quite difficult already to find sufficient sources for the sixteenth century. For some regions not included yet, data collection may be quite difficult even for the Early and Late Modern period.

We believe the HCD constitutes a solid foundation for future projects on the history of Dutch. In this paper, we aimed to introduce the corpus and explain its composition. We also illustrated the usefulness of the corpus through two short empirical explorations, focusing on spelling and morphosyntax. The rationale behind the HCD is our conviction that the best results can be obtained on the basis of carefully constructed corpora, which not only incorporate various variational dimensions, such as time, genre and region, but which are also based on reliable transcriptions of historical sources, including extensive metadata. Rather than bringing together large amounts of unstructured and poorly documented data with significant issues of representativeness, we would argue that there is a lot of potential, if not more, in the construction of smaller but balanced and well-structured corpora of different genres, across the literacy/orality continuum, and enriched with sufficient metadata to allow for sociolinguistically informed analyses of language variation and change.

Bibliography

Coussé, Evie. 2010. "Een digitaal compilatiecorpus historisch Nederlands." *Lexikos* 20: 123–42.

De Schutter, Georges, and Hanne Kloots. 2000. "Relatieve woorden in het literaire Nederlands van de 17e eeuw." *Nederlandse Taalkunde* 5: 325–42.

Elspaß, Stephan. 2012. "The Use of Private Letters and Diaries in Sociolinguistic Investigation." In *The Handbook of Historical Sociolinguistics*, edited by Juan

Manuel Hernández-Campoy, and Juan Camilo Conde-Silvestre, 156–69. Chicester: Wiley-Blackwell.

Hoeven, H. van der. 1978. "Verzamelaars en pamfletten." In *Catalogus van de pamfletten-verzameling berustende in de Koninklijke Bibliotheek* (herdruk van de oorspronkelijke uitgave van 1889–1920), edited by Willem Pieter Cornelis Knuttel, v– xxiii. Utrecht: HES.

Horst, Joop van der. 2008. *Geschiedenis van de Nederlandse syntaxis*. 2 vols. Leuven: Leuven University Press.

Koch, Peter, and Wulf Oesterreicher. 1985. "Sprache der Nähe – Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte." *Romantistisches Jahrbuch* 36: 15–43.

Loon, Jozef Van. 2014. *Historische fonologie van het Nederlands* (2nd ed.). Deurne: Universitas.

Marynissen, Ann. 2011. "Namen." In *Dialectatlas van het Nederlands,* edited by Nicoline van der Sijs, 300–53. Amsterdam: Bert Bakker.

Nevalainen, Terttu, and Helena Raumolin-Brunberg. 2012. "Historical Sociolinguistics: Origins, Motivations, and Paradigms." In *The Handbook of Historical Sociolinguistics*, edited by Juan Manuel Hernández-Campoy, and Juan Camilo Conde-Silvestre, 22– 40. Chicester: Wiley-Blackwell.

Nevalainen, Terttu, and Helena Raumolin-Brunberg. 2017. *Historical Sociolinguistics: Language Change in Tudor and Stuart England* (2nd ed.). Londen: Routledge.

Piersoul, Jozefien, Robbert De Troij, and Freek Van de Velde. 2021. "150 years of written Dutch: The construction of the Dutch Corpus of Contemporary and Late Modern Periodicals." *Nederlandse Taalkunde* 26, no. 3: 339–62.

Puttaert, Jill. 2019. *Vergeten stemmen van onderop: Een sociolinguïstische analyse van briefwisseling van de lagere klassen in de Lage Landen in de lange negentiende eeuw*. Brussels: Vrije Universiteit Brussel.

Rutten, Gijsbert. 2011. With the cooperation of Rik Vosters. *Een nieuwe Nederduitse spraakkunst: Taalnormen en schrijfpraktijken in de Zuidelijke Nederlanden in de achttiende eeuw*. Brussels: VUBPRESS.

Rutten, Gijsbert, and Marijke van der Wal. 2014. *Letters as Loot: A sociolinguistic approach to seventeenth- and eighteenth-century Dutch*. Amsterdam/Philadelphia: John Benjamins.

Rutten, Gijsbert, Iris Van de Voorde, and Rik Vosters. 2023. "Transmission and Diffusion." In The Cambridge Handbook of Historical Orthography, edited by Marco Condorelli, and Hanna Rutkowska, 596–616. Cambridge: Cambridge University Press.

SAND 1: Barbiers, Sjef, Hans Bennis, Gunther De Vogelaer, Magda Devos, and Margreet van der Ham. 2005. *Syntactische Atlas van de Nederlandse Dialecten: Deel 1*. Amsterdam: Amsterdam University Press.

Sijs, Nicoline van der. 2004. *Taal als mensenwerk: Het ontstaan van het ABN*. Den Haag: Sdu.

Voorde, Iris Van de. 2022. *Pluricentriciteit in de taalgeschiedenis: Bouwstenen voor een geïntegreerde geschiedenis van het Nederlands (16de-19de eeuw)*. Amsterdam: LOT.

Wal, Marijke van der. 2002. "Relativisation in the History of Dutch: Major Shift or Lexical Change?" In *Relativisation on the North Sea Littoral*, edited by Patricia Poussa, 27–36. München: Lincom.

Wal, Marijke van der. 2003. "Relativiteit in de grammaticale traditie: Tussen norm en descriptie?" In *Bon jours Neef, ghoeden dagh Cozyn! Opstellen aangeboden aan Geert Dibbets*, edited by Els Ruijsendaal, Gijsbert Rutten, and Frank Vonk, 361–75. Münster: Nodus.

Notes

1. We intend to publish the corpus in the near future, for example, in cooperation with the Instituut voor de Nederlandse Taal 'Dutch Language Institute'.
2. https://varieng.helsinki.fi/CoRD/corpora/HelsinkiCorpus/
3. https://varieng.helsinki.fi/CoRD/corpora/ARCHER/updated%20version/introduction.html
4. For German, the GerManC corpus was compiled at the University of Manchester, see https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2544. For Spanish: the Corpus de Documentos Españoles Anteriores a 1800 or CODEA (https://corpuscodea.es). A French example is the corpus FRAN with texts from North America, see https://www.usherbrooke.ca/crifuq/recherche/corpus/corpus-heberges/corpus-fran.
5. https://www.dbnl.org
6. See e.g. https://www.bijbelsdigitaal.nl for manually transcribed versions of historical Bibles from 1477 to 1648. Or see https://brievenalsbuit.ivdnt.org/ for the Letters as Loot Corpus, which mainly comprises private letters from the seventeenth and eighteenth centuries.
7. https://www.ivdnt.org/historisch-nederlands/
8. https://www.nederlab.nl/onderzoeksportaal/?action=verkennen
9. http://resources.huygens.knaw.nl/leidsetextielnijverheid
10. https://primarysources.brillonline.com/browse/dutch-pamphlets-online
11. https://www.zooniverse.org/projects/wikiscriptaneerlandica/wikiscriptaneerlandica-ii (with financial support from Algemeen Nederlands Verbond (ANV))
12. Note that the <ae>-spelling has been retained in Belgian surnames up to the present, such as *Adriaens* (Marynissen 2011, 317).
13. We found one single token spelled with <aa> in Flanders in the eighteenth century.