Policy versus Practice. Language variation and change in eighteenth- and nineteenth-century Dutch

Krogull, A.

CHAPTER 4

Corpus and methodology

1 Introduction

For the specific purpose of measuring and assessing the normative influence of the early nineteenth-century language policy on actual language usage, a new diachronic multi-genre corpus of more than 420,000 words was compiled. The *Going Dutch Corpus*, named after the research programme for which it was built, represents late eighteenth- and early nineteenth-century Dutch in the Northern Netherlands. Based on the assumption that linguistic changes affect different genres to different degrees, the corpus comprises data from three different types of authentic text sources, viz. (1) private letters (approx. 210,000 words), (2) diaries and travelogues (approx. 140,000 words), and (3) newspapers (approx. 70,000 words). In line with historical-sociolinguistic research and the *language history from below* approach (cf. Chapter 3), this corpus design takes into account handwritten and conceptually more 'oral'-like ego-documents, i.e. private letters, diaries and travelogues. On the other hand, the corpus also incorporates printed and published texts, in this case newspapers, which are commonly regarded as fairly standardised writing. The three genres of the *Going Dutch Corpus* and their use in historical-sociolinguistic research will be introduced in more detail in Section 3.1.

Considering the historical event of the official *schrijftaalregeling* 'written language regulation' in 1804 and 1805 as the main point of departure, two diachronic cross-sections of twenty years each were chosen. The nineteenth-century period of 1820–1840 represents the generation of language users *after* the introduction of Siegenbeek's (1804) official orthography and Weiland's (1805) official grammar, i.e. those writers who had (probably) received the national education with its corresponding language norms. Symmetrically, the eighteenth-century period of 1770–1790 represents the generation of language users *before* the officialised language norms were introduced. The diachronic dimension of the corpus will be discussed in Section 3.2.

With regard to the considerable degree of regional variation in the investigated language area, the *Going Dutch Corpus* covers seven regions of the Northern Netherlands, which are based on present-day provincial boundaries: Friesland, Groningen, North Brabant, North Holland, South Holland, Utrecht and Zeeland. This selection of regions comprises both the urbanised centre (i.e. North and South Holland, Utrecht) and more peripheral parts of the language area (i.e. Friesland, Groningen, North Brabant). The spatial dimension of the corpus, addressing variation between individual regions as well as between the centre and the periphery, will be discussed in Section 3.3.

Integrating a social dimension into the corpus design, the texts in the two sub-corpora of ego-documents were written by both men and women, mainly from

the (upper) middle to the upper classes. The variables of social class and gender will be introduced in Section 3.4.

In more general terms, the present chapter aims to give detailed insights into the compilation process of the *Going Dutch Corpus* and the methodology applied in this dissertation. Section 2 first outlines the collection and selection of corpus data (2.1), then describes the transcription procedure and conventions (2.2), and finally presents an overview of the size and structure of the final corpus (2.3). Section 3 introduces four variational dimensions and their independent variables investigated in all corpus-based case studies, viz. the genre dimension (3.1), the diachronic dimension (3.2), the spatial dimension (3.3), and the social dimension (3.4). The additional dimension of inter- and intra-individual variation and change will be discussed in Section 4, briefly presenting the specifically compiled *Martini Buys Correspondence Corpus*. Taking into account the developments in metalinguistic discourse, Section 5 introduces the *Normative Corpus of the Northern Netherlands*, a collection of eighteenth-century normative publications. Finally, Section 6 outlines the systematic methodological approach to the linguistic analyses in Chapters 5–12, followed by some final remarks on statistical methods.

2 Compiling the *Going Dutch Corpus*

Compiling a multi-genre corpus requires awareness of the specific characteristics and methodological challenges of its textual sources. In order to collect a representative and well-balanced sample of a certain genre on the one hand, and to meet the need of comparability with other genres on the other hand, customised approaches and selection criteria were developed for the *Going Dutch Corpus*. Section 2.1 discusses the collection and selection of corpus data. The transcription procedure and conventions will be presented in Section 2.2.

2.1 Collection and selection of data

The most crucial difference between the three genres included in the *Going Dutch Corpus* concerns the medium of texts, i.e. handwriting and print (cf. Rutkowska & Rössler 2012: 219). In fact, two of the three sub-corpora (i.e. private letters, diaries and travelogues) represent largely unpublished and handwritten ego-documents, whereas the third sub-corpus (i.e. newspapers) contains published and printed texts. Therefore, this section addresses the collection and selection of ego-documents and newspapers separately.

Ego-documents

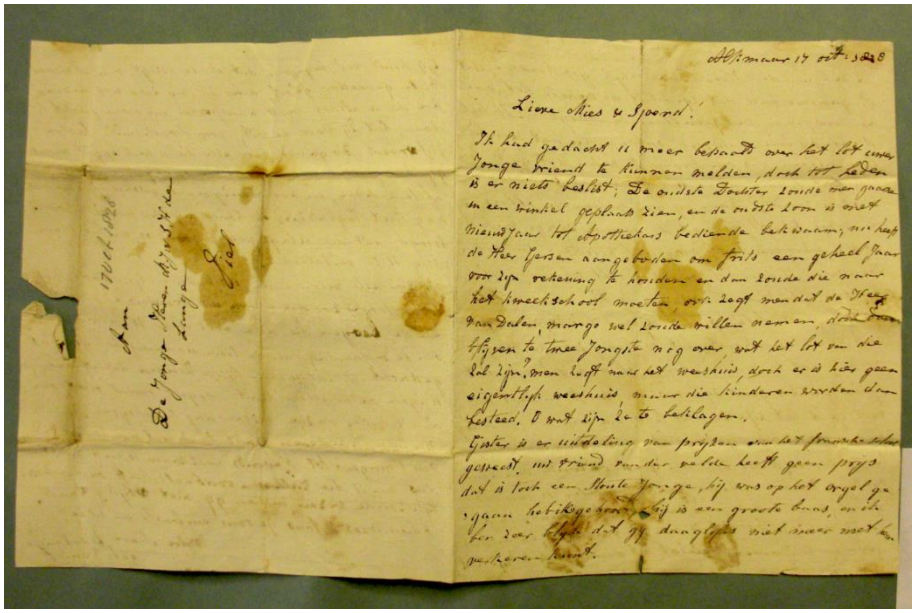
The compilation of the two sub-corpora of handwritten ego-documents, i.e. private letters as well as diaries and travelogues, started with a preparatory phase of thorough research and planning, in order to detect what kind of material was

actually available for the two investigated periods of 1770–1790 and 1820–1840, and where to find these sources. The ego-documents selected for the *Going Dutch Corpus* were collected from numerous municipal, regional and provincial archives spread all over the Netherlands, as listed below (categorised by province):

- Friesland: *Historisch Centrum Leeuwarden*, *Tresoar* (both Leeuwarden),
- Groningen: *Groninger Archieven* (Groningen),
- North Brabant: *Stadsarchief Breda* (Breda), *Brabants Historisch Informatie Centrum* ('s-Hertogenbosch), *Regionaal Archief Tilburg* (Tilburg),
- North Holland: *Regionaal Archief Alkmaar* (Alkmaar), *Stadsarchief Amsterdam* (Amsterdam), *Noord-Hollands Archief* (Haarlem), *Westfries Archief* (Hoorn),
- South Holland: *Archief Delft* (Delft), *Nationaal Archief* (The Hague), *Regionaal Archief Dordrecht* (Dordrecht), *Erfgoed Leiden en Omstreken* (Leiden), *Stadsarchief Rotterdam* (Rotterdam),
- Utrecht: *Het Utrechts Archief* (Utrecht),
- Zeeland: *Zeeuws Archief* (Middelburg).

As a rule, I visited the above-mentioned archives in order to request original manuscript sources and to take digital photographs (cf. Figure 1 for an example).

Figure 1. Private letter by Anna de Lange (17 October 1828, de Lange family archive, *Regionaal Archief Alkmaar*).



Only in a few exceptional cases, scans of the original documents were provided by staff members of the archives.

The digital images were inventoried according to a standardised format, for instance *Alkmaar_DeLange_79011_520_let12*. These file names include information about the place of the archives (i.e. *Regionaal Archief Alkmaar* in the town of Alkmaar), the name of the family archive (i.e. de Lange family). Furthermore, the file name contains the exact accession (79011) and inventory numbers (520), which makes it easy to trace back the origin of the documents. The abbreviation *let12* refers to the genre (i.e. private letter) and the individual code assigned to the document within a given inventory number (usually containing more than one document).

For the eighteenth-century data of private letters, the *Going Dutch Corpus* takes advantage of the extended *Letters as Loot* corpus, which had previously been compiled as part of the research programme *Letters as Loot. Towards a non-standard view on the history of Dutch* at Leiden University (2008–2013), directed by Marijke van der Wal (<http://brievenalsbuit.inl.nl>). The project investigated variation and change in the so-called sailing letters, confiscated during the wars fought between the Netherlands and England, and kept in the National Archives in Kew (London). This unique and linguistically highly valuable collection of Dutch private letters from the seventeenth and eighteenth centuries comprises texts from all ranks of the society, written by both men and women (Rutten & van der Wal 2014).

In order to achieve comparability with the letters specifically collected for the *Going Dutch Corpus*, a set of criteria had to be introduced. The selected texts from the extended *Letters as Loot* corpus comprise 104 private autograph letters (59,496 words) from the eighteenth-century period of 1776–1784. These letters were written by men and women from the upper middle class (UMC) and upper class (UC), which correspond with the social ranks predominantly represented in the *Going Dutch Corpus* (cf. Section 3.4.1). Geographically, the so-called 'regions of residence', i.e. the regions where letter writers were born and raised, largely match the regional categories of the *Going Dutch Corpus* (cf. Section 3.3.1), except for the region of North Holland. In this case, the *Letters as Loot* corpus distinguishes between 'North Holland (Amsterdam)' and 'North Holland (rest of the province)' as two separate categories (Rutten & van der Wal 2014: 11-12):

> Amsterdam is considered separately for geographical as well as demographic reasons. Geographically, the city of Amsterdam is located in the south of North Holland, separated from the northern parts of North Holland by water. Demographically, Amsterdam was a highly urbanized metropolis, attracting many immigrants from the rural areas of Holland and from other provinces of the Netherlands, as well as from abroad, mainly from German-speaking regions.

In contrast, the category of 'North Holland' in the *Going Dutch Corpus* comprises both Amsterdam and the rest of the province. Previous research, including a historical-sociolinguistic study of negation in seventeenth- and eighteenth-century Dutch, has shown that the metropolis of Amsterdam is "not exceptionally progressive compared to the other regions", and that "it perfectly fits

into the overall north-to-south pattern: it is less progressive than North Holland, and more progressive than South Holland" (Rutten & van der Wal 2013: 118). It did not seem fully justified to split North Holland into two separate categories again, which is why it is treated as one single regional category in the *Going Dutch Corpus*. The selected *Letters as Loot* texts, generally labelled as 'North Holland' here, thus actually comprise data from both Amsterdam and the rest of the province.

Table 1 provides an overview of the selected eighteenth-century private letters taken from the extended *Letters as Loot* corpus, distributed across the seven regions of the *Going Dutch Corpus* (i.e. FR = Friesland, GR = Groningen, NB = North Brabant, NH = North Holland, SH = South Holland, UT = Utrecht, ZE = Zeeland) and across genders.

Table 1. Selection of eighteenth-century private letters taken from the extended *Letters as Loot* corpus.

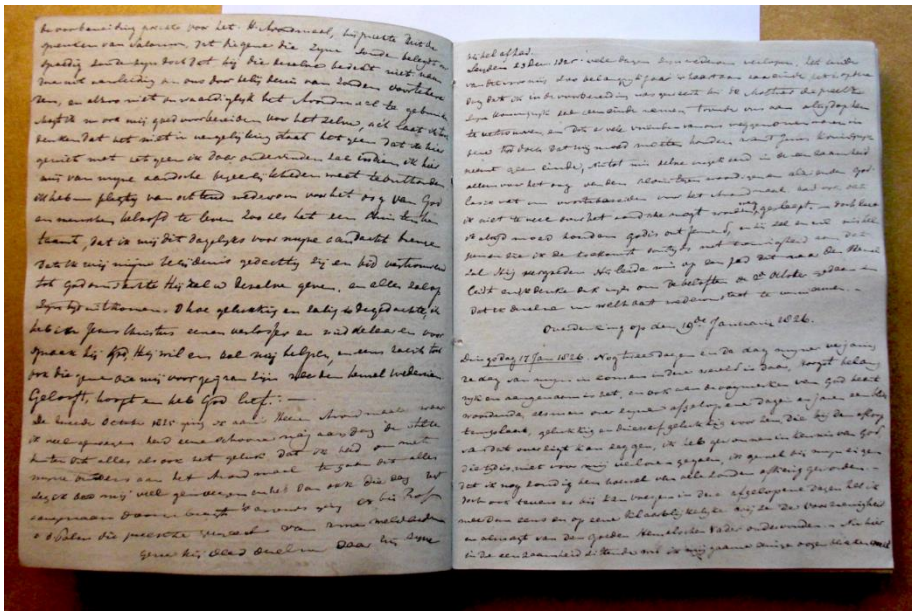
	FR	**GR**	**NB**	**NH**	**SH**	**UT**	**ZE**	**Total**
Male	5	4	2	18	13	8	6	56
Female	8	–	3	16	11	2	8	48
Total	**13**	**4**	**5**	**34**	**24**	**10**	**14**	**104**

It shows that the regions are not equally represented in quantitative terms. Most texts, in fact, stem from the western coastal regions of North Holland, South Holland and, to a lesser extent, Zeeland. Although the texts from *Letters as Loot* form the basis of the eighteenth-century data (104 out of 200 private letters), additional material had to be collected from various Dutch archives in order to fill the gaps in the under-represented regions, especially Groningen, North Brabant and Utrecht, but also Friesland and Zeeland.

Similar to the collection of private letters, the sub-corpus of diaries and travelogues also started out with exploratory research on the availability of suitable material. The fundamental works by Lindeman et al. (1993; 1994), providing a comprehensive inventory of ego-documents in the Northern Netherlands, as well as the corresponding and highly valuable website of the *Center for the Study of Egodocuments and History* <http://www.egodocument.net> by Arianne Baggerman and Rudolf Dekker, served as a starting point for the compilation of this sub-corpus. However, whereas private letters were relatively numerous and easy to find in Dutch archives, it was considerably more challenging to collect an appropriate amount of diaries and travelogues. Not surprisingly, there were far fewer eighteenth- and nineteenth-century diarists than letter writers, which affected the availability of suitable texts for the periods under investigation.

The actual procedure to collect and inventory diaries and travelogues was similar to the collection of private letters. Again, I visited various archives in order to take digital photographs of original manuscript documents (cf. Figure 2 for an example).

Figure 2. Diary by Pieter Glaudius Hubrecht (1825/1826, Hubrecht family archive, *Erfgoed Leiden en Omstreken*).



The selected archival sources were inventoried according to the same standardised format, e.g. *Leiden_Hubrecht_529_457_dia01*, with the abbreviation *dia* referring to the genre of diaries and travelogues.

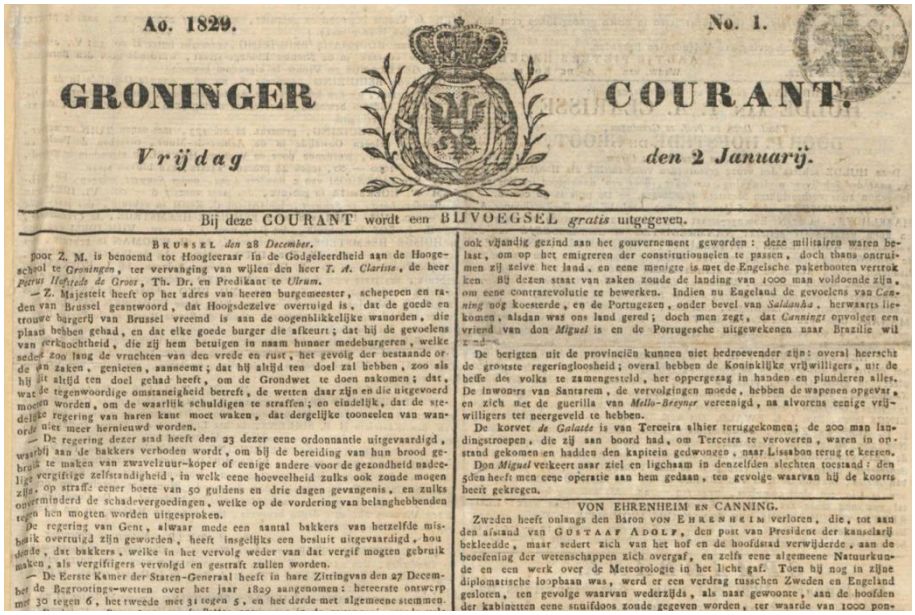
Newspapers

Unlike the sub-corpora of handwritten ego-documents, which are based on original archival sources, the sub-corpus of newspapers was compiled on the basis of digital scans taken from the *Delpher* website (<http://www.delpher.nl>) (cf. Figure 3 for an example).

Delpher is an online service, which gives free access to a vast amount of Dutch newspapers, magazines and books from the seventeenth to the twentieth century. For the *Going Dutch Corpus*, digitised newspapers were selected from the eighteenth-century period of 1770–1790, and from the nineteenth-century period of 1820–1840.

Taking into account regional variation, a representative newspaper was selected for each of the seven regions, preferably published in both periods and accessible on the *Delpher* website. Only in the case of North Brabant, data had to be taken from two different newspapers, both of which were published in the city of 's-Hertogenbosch, viz. the *'s Hertogenbossche courant* (1770–1790) and the *Noord Brabander* (1820–1840).

Figure 3. Title page of the *Groninger courant* (2 January 1829, *Delpher.nl*).



Ao. 1829. No. 1.

GRONINGER COURANT.

Vrijdag den 2 Januarij.

Bij deze COURANT wordt een BIJVOEGSEL *gratis* uitgegeven.

BRUSSEL *den* 28 *December*.

door Z. M. is benoemd tot Hoogleeraar in de Godgeleerdheid aan de Hoogeschool te *Groningen*, ter vervanging van wijlen den heer *T. A. Clarisse*, de heer *Petrus Hofstede de Groot*, Th. Dr. en Predikant te *Ulrum*.

— Z. Majesteit heeft op het adres van heeren burgemeester, schepenen en raden van Brussel geantwoord, dat Hoogstdezelve overtuigd is, dat de goede en trouwe burgerij van Brussel vreemd is aan de oogenblikkelijke wanorden, die plaats hebben gehad, en dat elke goede burger die afkeurt; dat hij de gevoelens van verknochtheid, die zij hem betuigen in naam hunner medeburgeren, welke sedert zoo lang de vruchten van den vrede en rust, het gevolg der bestaande orde en zaken, genieten, aanneemt; dat hij altijd ten doel zal hebben, zoo als hij dit altijd ten doel gehad heeft, om de Grondwet te doen nakomen; dat, wat de tegenwoordige omstandigheid betreft, de wetten daar zijn en die uitgevoerd moeten worden, om de waarlijk schuldigen te straffen; en eindelijk, dat de stedelijke regering van haren kant moet waken, dat dergelijke tooneelen van wanorde niet meer hernieuwd worden.

— De regering dezer stad heeft den 23 dezer eene ordonnantie uitgevaardigd, waarbij aan de bakkers verboden wordt, om bij de bereiding van hun brood gebruik te maken van zwavelzuur-koper of eenige andere voor de gezondheid nadeelige vergiftige zelfstandigheid, in welk eene hoeveelheid zulks ook zonde mogen zijn, op straffe eener boete van 50 guldens en drie dagen gevangenis, en zulks onverminderd de schade-vergoedingen, welke op de vordering van belanghebbenden tegen hen mogten worden uitgesproken.

De regering van Gent, alwaar mede een aantal bakkers van hetzelfde misbruik overtuigd zijn geworden, heeft insgelijks een besluit uitgevaardigd, houdende, dat bakkers, welke in het vervolg weder van dat vergift mogten gebruik maken, als vergiftigers vervolgd en gestraft zullen worden.

— De Eerste Kamer der Staten-Generaal heeft in hare Zitting van den 27 December de Begrootings-wetten over het jaar 1829 aangenomen: hetzeerste ontwerp met 30 tegen 6, het tweede met 31 tegen 5, en het derde met algemeene stemmen.

ook vijandig gezind aan het gouvernement geworden: deze militairen waren belast, om op het emigreren der constitutionelen te passen, doch thans ontruimen zij zelve het land, en eene menigte is met de Engelsche pakerbooten vertrokken. Bij dezen staat van zaken zoude de landing van 1000 man voldoende zijn, om eene contrarevolutie te bewerken. Indien nu Engeland de gevoelens van *Canning* nog koesterde, en de Portugezen, onder bevel van *Saldanha*, herwaarts liekomen, alsdan was ons land gered; doch men zegt, dat *Canningt* opvolger een vriend van don *Miguel* is en de Portugesche uitgewekenen naar Brazilie wil zenden.

De berigten uit de provinciën kunnen niet bedroevender zijn: overal heerscht de grootste regeringloosheid; overal hebben de Koninklijke vrijwilligers, uit de heffe des volks te zamengesteld, het oppergezag in handen en plunderen alles. De inwoners van Santarem, de vervolgingen moede, hebben de wapenen opgevat, en zich met de guerilla van *Mello-Breyner* vereenigd, na alvorens eenige vrijwilligers tot neergeveld te hebben.

De korvet *de Galatée* is van Terceira alhier teruggekomen; de 200 man landingstroepen, die zij aan boord had, om Terceira te veroveren, waren in opstand gekomen en hadden den kapitein gedwongen, naar Lissabon terug te keeren.

Don *Miguel* verkeert naar ziel en ligchaam in denzelfden slechten toestand: den 5den heeft men eene operatie aan hem gedaan, ten gevolge waarvan hij de koorts heeft gekregen.

VON EHRENHEIM EN CANNING.

Zweden heeft onlangs den Baron VON EHRENHEIM verloren, die, tot aan den afstand van GUSTAAF ADOLF, den post van President der kanselarij bekleedde, maar sedert zich van het hof en de hoofdstad verwijderde, aan de beoefening der wetenschappen zich overgaf, en zelfs eene algemeene Natuurkunde en een werk over de Meteorologie in het licht gaf. Toen hij nog in zijne diplomatische loopbaan was, werd er een verdrag tusschen Zweden en Engeland gesloten, ten gevolge waarvan wederzijds, als naar gewoonte, aan de hoofden der kabinetten eene snuifdoos zoude gegeven worden, ter waarde van 1000 pon-

The final sub-corpus contains text samples from the following newspapers (categorised by province):

- Friesland: *Leeuwarder courant*,
- Groningen: *Groninger courant*,
- North Brabant: *'s Hertogenbossche courant*, *Noord Brabander*,
- North Holland: *Oprechte Haarlemsche courant*,
- South Holland: *Leydse courant*,
- Utrecht: *Utrechtsche courant,*
- Zeeland: *Middelburgsche courant*.

2.2 Transcription procedure and conventions

In order to make the selected material machine-readable and analysable for corpus-linguistic software tools such as *WordSmith*, all texts were manually transcribed and saved as electronic text files. As the genres included in the *Going Dutch Corpus* represent two fundamentally different types of data, i.e. handwritten and printed texts, the transcription procedures and conventions will be discussed separately for ego-documents and newspapers.

Ego-documents

Based on digital images of archival sources, the selected ego-documents were diplomatically transcribed, which means that the original spelling, punctuation and word boundaries were retained and not normalised according to contemporary standards. With regard to the handwritten nature of private letters, diaries and travelogues, detailed guidelines were essential in order to guarantee a consistent transcription process (cf. Appendix I for the full transcription conventions).

Aspects that were taken into consideration include ambiguities (<ambig>*word*</ambig>)[17], illegibilities (<illeg/>), deletions (~~*word*~~), insertions (<ins>*word*</ins>), underlining (<u>*word*</u>), hyphenation (<reg orig="*wo|rd*">*word*</reg>), capitalisation and intra-word spacing, as well as line and page breaks. Illustrating the use of these tags, the example below provides the extract of a transcribed private letter taken from the *Going Dutch Corpus*:

> *Waarde Dogter! Breda den 16 augustus* <u>*1825*</u>
> *UEd brief met couvert er om, heb ik wel ontvangen*
> *daar ik hem vrÿdags s'avons heb ontvangen konde*
> *ik s'zaterdags niets meer van UEd goed laten* <reg orig="*was|sen*">*wassen*</reg>
> *omdat* ~~*ik*~~ <ins>*er*</ins> *twee leegen*<ins>*dagen*</ins> *op volgen*
> *en ik zend*
> *UEd nu maar zo veel als ik in den Groote* <reg orig="*Lesse|naar*">*Lessenaar*</reg>
> *kan bergen en waar van het zÿsje dat*
> *ik UEd zend in UEd roode jas gespeld is, de jas*
> *is wel vuil maar om dat er nog plaats in was*
> *heb ik er hem in gedaan UEd kund hem dan maar*
> *met u ander goed laten wassen […]*
>
> 'Dear daughter! Breda the 16th of August 1825
> I have well received your letter with [the] envelope around it.
> As I received it on Friday evening, I could not
> have your laundry washed on Saturday
> because two vacant days follow and now I send
> you as much as I can store in the big lectern
> and of which the siskin that
> I send you is tacked on your red jacket. The jacket
> is dirty but because there was still space in it
> I put it in there. You could have it
> washed with your other laundry then […]'

The representation of *ij/y*, i.e. one of the central orthographic variables under investigation, needed special attention before and during the transcription process. In order to be able to investigate the use of this variable in detail, four main variants were distinguished (cf. also Chapter 9):

---

[17] The tag <ambig> was used to indicate unclear or ambiguous spellings and words. The suggested transcriptions are too uncertain to be taken into consideration for the corpus-based analysis of orthographic features in particular, but may still be useful for the study of morphosyntactic features and further context-related matters.

(1) <ij>, i.e. double-dotted <ij> (*lange ij*) with <i> and <j> written as two separate characters,

(2) <ÿ>, i.e. double-dotted <y>,

(3) <y>, i.e. (undotted) <y> (*Griekse y*),

(4) <°y>, i.e. other variants, e.g. single-dotted <y>, <y> with accent marks or other diacritics (positions of dots and accents are irrelevant here).

The introduction of clearly defined conventions, ensuring that the transcriptions are as consistent as possible, is crucial especially with regard to the number of people involved in the transcription process. For the most part, the transcriptions of texts were carried out by the project's research assistants Christa Bouwmans and Hielke Vriesendorp, as well as by myself[18]. After the first transcription phase, each document was thoroughly double-checked either by one of the assistants or myself, comparing the first transcription to the corresponding digital images in order to detect and fix possible transcription errors. Even though a few remaining inconsistencies cannot be excluded, the final transcriptions can be considered as accurate and reliable.

Within the sub-corpus of private letters, the transcription of eighteenth-century data marked a special case. As mentioned in Section 2.1.1, a substantial number of private letters was taken from the *Letters as Loot* corpus. The applied conventions, however, were slightly different from the transcription guidelines followed here (cf. Nobels 2013 and Simons 2013 for a detailed description). Consequently, even minor deviations would have resulted in an inconsistent use of tags and, in some cases, to different transcriptions of specific variants[19]. For the sake of consistency and comparability, all transcriptions from the *Letters as Loot* corpus were modified according to the conventions of this dissertation. New text files were created for each of these external transcriptions, applying the same tags as for all other private letters in the *Going Dutch Corpus*.

Enriching the transcriptions with a basic set of metadata, headers were added to each text file, both for the newly collected letters and for those taken from the *Letters as Loot* corpus. These headers contain information on the provenance of the original archival documents, the genre (either 'letter' or 'diary'), as well as the date and place of writing. Furthermore, all headers provide information on the transcriber, the number of words, and (optionally) notes about the transcriptions. A header example is given below:

[18] I also want to thank our MA students Brenda Assendelft, Anne Rose Haverkamp and Marlies Reitsma for contributing some initial transcriptions as part of their Master's theses.

[19] In the *Letters as Loot* corpus, the highly variable representations of the *ij/y* variable were transcribed in a less complex manner, only taking into account the two variants <ij> and <y>. All occurrences were manually revised in order to consider the four variants distinguished in the *Going Dutch Corpus* (i.e. <ij>, <ÿ>, <y>, <°y>).

```
<header>
DOCUMENT: Alkmaar_DeLange_79011_520_let12
ARCHIVE: Regionaal Archief Alkmaar
GENRE: letter
DATE: 1828-10-17
PLACE: Alkmaar
TRANSCRIPTION: HV
NOTES:
WORD COUNT: 592
</header>
```

For the sake of a convenient corpus analysis, each transcription selected for the *Going Dutch Corpus* was given a file name, for example *LET-2-NH-F_Alkmaar_DeLange_79011_520_let12*. In addition to the name of the source document (*Alkmaar_DeLange_79011_520_let12*), these file names contain the standardised codes for genre[20] (LET), period[21] (2), region[22] (NH) and gender[23] (F).

One final remark from a more practical point of view concerns the condition of archival documents and the legibility of handwriting. Both factors, at least in some cases, could influence the selection of data. In fact, (parts of) ego-documents in very poor condition and/or with hardly legible handwritings had to be neglected. As Van Bergen & Denison (2007: 4) rightly note, "deciphering the letters could be at least as time-consuming as the actual transcription".

Newspapers

While the transcription of handwritten ego-documents was indeed a time-consuming and, depending on the legibility of handwriting, challenging procedure, the transcription of newspapers turned out to a comparatively straightforward task. First of all, the texts selected for this sub-corpus were manually transcribed[24] based on digitised newspapers on the *Delpher* website. All transcriptions were provided by research assistant Hielke Vriesendorp and double-checked by myself. Again, all texts were transcribed diplomatically, intended to be as close to the original as possible. The legibility of the newspaper texts was generally unproblematic, although ambiguous and even illegible readings did occur, mostly due to scan quality or folds in the paper. Those instances were explicitly marked in the transcriptions with the corresponding tags for ambiguities (`<ambig>`*word*`</ambig>`) or illegible words (`<illeg/>`).

---

[20] The codes for genres are: LET (= private letters), DIA (= diaries and travelogues), NEW (= newspapers).

[21] The codes for periods are: 1 (= period 1, 1770–1790), 2 (= period 2, 1820–1840).

[22] The codes for regions are: FR (= Friesland), GR (= Groningen), NB (= North Brabant), NH (= North Holland), SH (= South Holland), UT (= Utrecht), ZE (= Zeeland).

[23] The codes for genders are: F (= female), M (= male).

[24] Initial attempts to use OCR software (optical character recognition) resulted in transcriptions which were too unreliable and would have required a fair amount of manual correction.

In terms of typographic variation, capitalisation and other types of emphasis, i.e. usually words or passages in italics (<emph>*word*</emph>), were also transcribed in their original form. Furthermore, line and page breaks were taken into account as two fundamental aspects of layout[25]. This was mainly done for practical reasons in order to find back specific passages in the original scans more conveniently.

Similar to the transcriptions of ego-documents, all files in the sub-corpus of newspapers contain a header with a basic set of metadata, including the name and source of the document, the genre (i.e. 'newspaper'), the date(s) and place of publication, and the word count (i.e. a standardised amount of 5,000 words). An example is given below:

```
<header>
DOCUMENT: Groninger Courant
ARCHIVE: Delpher
GENRE: newspaper
DATE: 1828-01-01, 1829-01-02, 1830-01-01, 1831-01-04
PLACE: Groningen
TRANSCRIPTION: HV
NOTES:
WORD COUNT: 5064
</header>
```

Each text file in the final *Going Dutch Corpus* was assigned a code, for instance *NEW-2-GR.txt*, comprising information on the relevant independent variables, i.e. genre (NEW), period (2) and region (GR).

2.3 Size and structure of the final corpus

The aim was to compile a diachronic multi-genre corpus, which comprises 420,000 words, and can be divided into three sub-corpora, representing three different genres, viz. (1) private letters (approx. 210,000 words), (2) diaries and travelogues (approx. 140,000 words) and (3) newspapers (approx. 70,000 words). Table 2 provides an overview of the intended corpus size and structure, serving as a starting point for the actual collection and selection of data for the *Going Dutch Corpus*.

Table 2 also illustrates that the tripartite division of the *Going Dutch Corpus* into private letters, diaries and travelogues, and newspapers involves different sizes of sub-corpora. The underlying consideration behind the definition of (sub-)corpus sizes was the degree of (expected) uniformity and linguistic 'standardness', or, to put it the other way round, the degree of linguistic variation that is expected to be found in these texts.

[25] Catch words (so-called *custoden* in Dutch), i.e. words which are inserted at the end of a page and repeated on the following page, were only transcribed once.

Table 2. General corpus design and structure of the *Going Dutch Corpus*.

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**
Seven regions		**Seven regions**		**Seven regions**		**Seven regions**		**Seven regions**	**Seven regions**
♂	♀	♂	♀	♂	♀	♂	♀		

Firstly, the sub-corpus of private letters comprises approximately 210,000 words in total, with 105,000 words per period, 15,000 words per region per period, and so on. It has repeatedly been demonstrated in historical-sociolinguistic research that private letters are a particularly useful genre of ego-documents to investigate usage patterns, given the expectedly high degree of linguistic variation. Moreover, the wide availability of such texts in Dutch archives was a decisive factor in defining the intended size of the sub-corpus of private letters.

Second, the sub-corpus of diaries and travelogues comprises approximately 140,000 words. Although these texts, like private letters, belong to the group of ego-documents, they tend to be written in more 'standard'-like language (Schneider 2013: 66). Therefore, in comparison with private letters, less linguistic variation has to be expected here, which is why this sub-corpus contains a lower number of words than the linguistically more heterogeneous letter sub-corpus. The limited availability of suitable diaries and travelogues in Dutch archives was another, more practical reason to reduce the total number of words in this sub-corpus.

Third, as the only printed sources in the *Going Dutch Corpus*, newspapers are expected to display the highest degree of linguistic uniformity. Rutten & van der Wal (2014: 3) point out that the printed language from the eighteenth-century (onwards) can be characterised as considerably uniform. Therefore, even a comparatively limited amount of data, i.e. in this case approximately 70,000 words, was considered to be sufficient for a representative sample of contemporary newspaper writing, in particular with regard to the focus on pervasive orthographic and morphosyntatic features.

Table 3 gives an overview of the actual sizes of the three sub-corpora in the final version of the *Going Dutch Corpus*, distributed across the two diachronic cross-sections and across the entire corpus. As shown, the initially intended corpus size of approximately 420,000 words was reached. In fact, the intended sizes were practically reached for all three sub-corpora and for both periods.

Table 3. Sizes of the sub-corpora in the *Going Dutch Corpus* (in absolute numbers and percentage of the total corpus).

Period	Private letters	Diaries and travelogues	Newspapers	Total
	N words (%)	**N words (%)**	**N words (%)**	**N words (%)**
1770–1790	105,427 (25.0)	71,157 (16.9)	35,323 (8.4)	211,907 (50.2)
1820–1840	105,299 (25.0)	69,350 (16.4)	35,322 (8.4)	209,971 (49.8)
Total	**210,726 (50.0)**	**140,507 (33.3)**	**70,645 (16.8)**	**421,878 (100)**

3 Variational dimensions of the *Going Dutch Corpus*

This section introduces the variational dimensions integrated in the *Going Dutch Corpus*. In Chapters 5–12, four dimensions will be considered in the corpus analyses of orthographic and morphosyntactic variables, viz. (1) the genre dimension (Section 3.1) with its sub-corpora of private letters (3.1.1), diaries and travelogues (3.1.2) and newspapers (3.1.3), (2) the diachronic dimension (Section 3.2) with its two twenty-year periods before and after the *schrijftaalregeling*, (3) the spatial dimension (Section 3.3), which considers variation across regions (3.3.1) as well as centre and periphery (3.3.2), and (4) the social dimension (Section 3.4) with its variables of social class (3.4.1) and gender (3.4.2). The individual dimension of inter- and intra-writer variation will be addressed separately in Section 4.

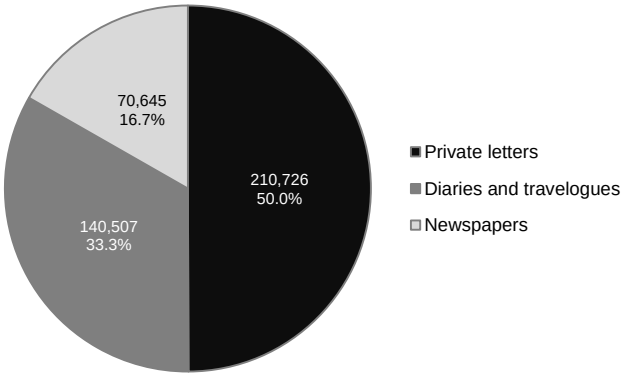
3.1 Genre dimension

Based on the assumption that linguistic changes affect different genres to different extents, the *Going Dutch Corpus* was designed as a diachronic multi-genre corpus, representing the following three genres: (1) private letters, (2) diaries and travelogues, and (3) newspapers. Figure 4 shows the relative distribution of these three genres in the final corpus.

As can be seen, the sub-corpus of private letters (50.0%) makes up one half of the entire *Going Dutch Corpus*, whereas the second half of the corpus comprises the sub-corpora of diaries and travelogues as well as newspapers with one-third (33.3%) and one-sixth (16.7%) of the data, respectively.

In terms of corpus design and, more specifically, the selection of genres, this dissertation aims to take an integrated multi-genre perspective on Dutch language history. On the one hand, it follows the historical-sociolinguistic approach *from below* by utilising handwritten ego-documents (i.e. private letters, diaries and travelogues). On the other hand, the corpus also incorporates more standard-like printed and published texts (i.e. newspapers).

Figure 4. Genre distribution in the *Going Dutch Corpus* (absolute number of words and percentage).



As outlined in Chapter 3 (cf. Section 3 in particular), traditional language histories often had a strong focus on the standardisation process and were first and foremost based on printed language, mainly literary and formal texts from the higher registers. These sources represent a fairly standardised form of writing, which often fails to fully reflect the variation found in authentic language usage. Suggesting an alternative approach to language history, historical sociolinguists therefore introduced the *language history from below* (e.g. Elspaß et al. 2007). This change of perspective involves a shift from relatively uniform printed texts to more informal handwritten sources from the private sphere, such as letters, diaries and travelogues. These ego-documents are conceptually more 'oral' and closer to the 'language of immediacy' than the sources traditionally studied in language historiography (cf. also Section 2 of Chapter 3 for Koch & Oesterreicher's (1985) conceptual continuum).

While many historical-sociolinguistic studies have criticised the teleological view on (primarily printed) language histories for being "one-sided, partial, biased" accounts of the linguistic past (Rutten et al. 2014b: 1-2), it has also been argued that the alternative approach *from below* "may run the risk of presenting another one-sided view of language history" (ibid.). Therefore, the selection of genres in the *Going Dutch Corpus* offers an integrated perspective, considering the study of handwritten ego-documents from the 'language of immediacy' and printed texts from the 'language of distance' as complementary rather than contradictory (cf. also Elspaß & Niehaus 2014). In fact, this multi-genre corpus design allows for a direct comparison of two conceptually more 'oral' genres of ego-documents (i.e. private letters, diaries and travelogues) in relation to a conceptually more 'literate' and standardised printed genre (i.e. newspapers). At the same time, this design enables to compare manuscript to print sources, investigating possible differences on the level of the medium.

Introducing the three genres of the *Going Dutch Corpus* individually, Sections 3.1.1 and 3.1.2 focus on the two sub-corpora of ego-documents, i.e. private letters as well as diaries and travelogues, respectively. Section 3.1.3 provides an outline of the sub-corpus of newspapers.

3.1.1 Private letters

In the field of historical sociolinguistics, ego-documents have been attested a special and particularly valuable role in gaining access to authentic language use in the past (cf. Chapter 3). According to Elspaß (2012: 156), they are "as close to speech as non-fictional historical texts can possibly be and therefore cast light on the history of natural language". Among the group of ego-documents, letters, and private letters in particular, are "the best possible data for studying everyday men and women in society, their linguistic knowledge and behaviour, as well as their social inscription" (Marquilhas 2012: 31). Similarly, Martineau (2013: 133) argues that "private family letters are the best documents for historical sociolinguistics because they are the closest written documents to language of immediacy".

As part of the *Going Dutch Corpus*, a sub-corpus of private letters of approximately 210,000 words was compiled. During the careful selection phase of Dutch letters from the late eighteenth- and early nineteenth centuries, the following criteria were considered and applied:

(1) The selected private letters should primarily include personal matters (i.e. exclusion of business letters).

(2) The selected private letters should be written in a symmetrical communicative setting (i.e. correspondence between family members and friends).

(3) The selected private letters should represent 'everyday communication' (i.e. exclusion of letters of condolence, thanks and congratulations).

First of all, all letters had to be classified as 'personal letters' or 'private letters', implying that their primarily contain personal matters, written in the private sphere. This also means that business letters and other types of non-private letters were excluded. It should be noted, though, that the dividing line between private and business letters is often very fuzzy, especially in historical letter writing (e.g. Van Bergen & Denison 2007: 4; Włodarcyk 2013: 205). Therefore the classification of letters as private letters applied in this dissertation mainly follows the rule of thumb suggested by Nobels (2013: 27-28):

> if the sender and addressee of the letter were closely related to each other (e.g. husband and wife, father and son, cousin and cousin, nephew and uncle) the letter was classified as private, even if it contained information about business. If the sender and intended receiver of the letter were not closely related and if the letter

> did not contain any private message other than greetings for the addressee's family and wishes for the addressee's good health, the letter was classified as a business letter.

In fact, the vast majority of letters in the *Going Dutch Corpus* represents correspondence between family members. The extraordinary value of this specific type of correspondence in historical-sociolinguistic investigation is also emphasised by Martineau (2013: 132), who considers private family letters as the best possible way to gain access to authentic language use in historical contexts:

> Private correspondence, especially letters to family members, are a valuable primary source of information for reconstituting the nature of exchange, and the language used in former times. Despite the use of writing as a medium, family letters reflect a fairly close relationship between the writer and the recipient in a manner similar to exchanges between friends, not always found in such oral materials as folktales or plays featuring popular characters, or even some modern sociolinguistic interviews.

Whenever family archives provided a substantial amount of suitable texts, preference was given to the more intimate relationships such as spouses, parent–child and siblings(-in-law) rather than, for instance, uncle/aunt–nephew/niece or cousin–cousin.

The second criterion is closely related to the personal content of letters, taking into account the symmetry in communication. With regard to the relationship between senders and addressees, all private letters selected for the *Going Dutch Corpus* should be written in a symmetrical communicative setting. Elspaß (2005: 55) discusses the unsuitability of sources in institutional and thus asymmetrical contexts:

> Wenig tauglich sind Quellen, die in einem institutsbezogenen Zusammenhang stehen, also Bitt-, Petitions-, Beschwerdebriefe oder andere Schreiben an Behörden. Erstens kann die Autorschaft solcher Briefe sowie der Einfluss vorgefertigter Briefmuster nicht eindeutig geklärt werden, und zweitens repräsentieren sie asymmetrische Kommunikation, d. h. dass durch die geforderte Anpassung an spezialisierte und routinierte Kommunikationsformen die ‚natürliche' Ausdrucksweise der Alltagssprache in hohem Maße verfremdet erscheint.

He argues that in composing "letters in asymmetrical communicative settings (letters of appeal), […] writers usually draw on discourse traditions with highly formalized discourse" (Elspaß 2012: 158), making them less suitable for the study of authentic language use. On the other hand, in symmetrical communication, "grammatical correctness, spelling or particular sets of formulae were not crucial to a successful communicational act, so that even barely literate people would take up

pen or pencil to write down texts of private interest" (ibid.). This makes them far more authentic in terms of 'historical orality'[26] than, for example, letters of request.

The third criterion considered during the selection phase refers to what Elspaß (2005) calls *geschriebene Alltagssprache* 'written everyday language'. In order to meet the criterion of 'everyday communication' as much as possible, more formal and formalised types of communication such as letters of condolence, letters of thanks and letters of congratulations, were preferably avoided. Nevertheless, it is important to relativise the 'everydayness' of these texts. In fact, Elspaß (2005: 66-67) remarks a striking contradiction with regard to 'everyday language' in private letters:

> Ein gewisser Widerspruch zwischen der Bezeichnung "Alltagsbriefe" und der Zuordnung der Briefe zur alltagssprachlichen Kommunikation scheint darin zu bestehen, dass diese Briefe eben nicht alltäglich geschrieben wurden. […] Obwohl durch verloren gegangene Briefe sicherlich Lücken in der Chronologie der überlieferten Briefserien bestehen, kann man doch feststellen, dass zwischen den Schreiben eines Briefwechsels oft Monate, sogar Jahre liegen. Es fällt auf, dass viele Briefe an Sonn- und Feiertagen […], geschrieben wurden, also gerade nicht im Alltag im Sinne von 'Wochen- und Arbeitstag'. […] Entscheidend für die Bestimmung von Textsorten wie den Privatbriefen als Alltagstextsorte ist nicht ihre Frequenz im alltäglichen Leben, sondern die Tatsache, dass sie überhaupt nur den Lebensbereichen und Gebrauchsdimensionen des Alltags zugeordnet werden können.

Ultimately, the fact that the letters were written within the sphere of everyday life was considered as more crucial than the actual frequency and moment of writing.

To sum up, the selected letters for the *Going Dutch Corpus* contain primarily personal content, are written in a symmetrical communicative setting as found in family correspondence, and, ideally, represent everyday language. Certain writers contributed more than one letter to the corpus, which met the defined selection criteria. However, in order to avoid overrepresentation of prolific writers (cf. Nobels 2013: 51), as well as to guarantee the comparability of texts (cf. Wegera 2013: 63), the number of words per writer was restricted to a maximum of approximately 2,000 words. This limit was based on the longest letter selected for this sub-corpus, which contains 2,078 words[27]. In practice, the data of individual letter writers may thus comprise either one long letter or a number of shorter letters.

As summarised in Table 4, the sub-corpus of private letters consists of 210,726 words in total, equally distributed across the two diachronic cross-sections (i.e. 105,427 words for 1770–1790; 105,299 words for 1820–1840) and more or less equally distributed across all seven regions (i.e. ideally 15,000 words per region per

[26] For a critical discussion on the notion of 'historical orality', see Zeman (2013).

[27] Similar limits were defined during the compilation of the *Letters as Loot* corpus, restricting the number of words per individual writer to a maximum of 2,000 words in the seventeenth-century cross-section (Nobels 2013: 50) and to a maximum of 2,500 words in the eighteenth-century cross-section (Simons 2013: 86).

period). The entire sub-corpus comprises 400 texts[28] (200 in each period), which were written by 298 different letter writers[29].

Table 4. General distribution of data in the sub-corpus of private letters.

Period	N texts (%)	N words (%)	N writers (%)
1770–1790	200 (50.0)	105,427 (50.0)	154 (51.7)
1820–1840	200 (50.0)	105,299 (50.0)	144 (48.3)
Total	**400 (100)**	**210,726 (100)**	**298 (100)**

Aiming at a well-balanced gender representation, the sub-corpus of private letters comprises data from 181 male and 117 female writers. The actual number of words, however, gives a more accurate overview of the achieved gender balance: 54.5% of the letter data was written by men, 45.5% written by women. Although male letters writers are thus slightly more prevalent in the corpus, this can be considered as a well-balanced gender representation, especially for a historical corpus.

In fact, filling the grid cells of the intended corpus design presented in Section 2.3 largely depended on the availability of archival material. This turned out to be the case for less urbanised provinces like Friesland and North Brabant. It was not possible to find an equal amount of texts written by men and women for each region and period. Therefore, the selection criteria were slightly loosened in order to reach the intended corpus size and, at the same time, not to neglect valuable data. It was decided to compensate for the gaps in some gender grid cells by adding more data from the other gender. To give an example, the eighteenth-century data from Friesland only comprise 4,645 words written by women (out of the intended 7,500 words). However, additional male data from the same period was available for this region, which was ultimately used to reach the intended number of 15,000 words. This does not mean, though, that this compensation strategy resulted in an overly male-dominated letter corpus. In the case of nineteenth-century North Brabant, for instance, the gaps in the male grid cell (only 4,333 words) were compensated by additional female data. This modification was considered for the benefit of a larger dataset and should not skew the corpus analyses considerably.

Tables 5a and 5b provide a detailed overview of the sub-corpus of private letters[30].

---

[28] Some letters were written by more than one hand. The transcriptions of each hand were saved as separate Text files (indicated by the codes *hand1*, *hand2* etc.) and treated as different texts in this overview, even though they were originally taken from the same archival document.

[29] In a few exceptional cases, letter writers contributed data for both periods. Since the two periods represent two distinct generations of language users, these writers are counted as two different individuals in this overview.

[30] The actual numbers of words per grid cell generally deviate from the (exact) intended numbers of 7,500/15,000 words. This is mainly due to the decision to include complete

Table 5a. Distribution of data in the sub-corpus of private letters across region and gender (P1 = 1770–1790).

P1	Male			Female			Total		
	Texts	*Words*	*Writers*	*Texts*	*Words*	*Writers*	*Texts*	*Words*	*Writers*
FR	23	10,889	19	15	4,645	9	38	15,534	28
GR	25	10,282	15	5	3,398	4	30	13,680	19
NB	21	10,286	13	11	5,187	7	32	15,473	20
NH	18	7,517	17	16	7,579	13	34	15,096	30
SH	13	7,503	13	11	7,548	10	24	15,051	23
UT	10	9,514	7	4	5,771	3	14	15,285	10
ZE	14	7,522	13	14	7,786	11	28	15,308	24
Total	**124**	**63,513**	**97**	**76**	**41,914**	**57**	**200**	**105,427**	**154**

Table 5b. Distribution of data in the sub-corpus of private letters across region and gender (P2 = 1820–1840).

P2	Male			Female			Total		
	Texts	*Words*	*Writers*	*Texts*	*Words*	*Writers*	*Texts*	*Words*	*Writers*
FR	18	10,121	14	9	5,560	6	27	15,681	20
GR	16	7,574	12	20	7,624	12	36	15,198	24
NB	13	4,333	10	18	11,620	10	31	15,953	20
NH	15	8,149	12	12	7,958	11	27	16,107	23
SH	17	7,949	16	14	7,644	9	31	15,593	25
UT	13	7,507	13	13	7,771	7	26	15,278	20
ZE	14	5,622	7	8	5,867	5	22	11,489	12
Total	**106**	**51,255**	**84**	**94**	**54,044**	**60**	**200**	**105,299**	**144**

3.1.2 Diaries and travelogues

In addition to the sub-corpus of private letters (3.1.1), the *Going Dutch Corpus* comprises a second type of handwritten ego-documents, viz. diaries and travelogues. Although these texts are often mentioned in the same breath as private

letters rather than cut-off samples. Whenever texts had to be shortened, transcriptions were continued until the end of the sentence.

letters, it has to be kept in mind that they represent two distinct types of ego-documents, differing in various respects. First and foremost, they represent opposite poles of the monologicity–dialogicity continuum, as pointed out by Elspaß (2012: 162):

> Whereas private letters are characterized by dialogue and 'a social practice' between the correspondents […], private diaries are strictly monologic by nature. Such texts may be as informal in style and unplanned in their conception as private letters, but they are usually less 'oral'.

Compared to private letters, which are characterised by their interactive purpose, diaries and travelogues are generally further away from the side of Koch's & Oesterreicher's (1985) 'language of immediacy' (van der Wal & Rutten 2013: 2; cf. also Chapter 3). Both terminologically and methodologically, the genre referred to as 'diaries and travelogues' needs some further clarification. Elspaß (2012: 163) outlines that

> the term 'diary' covers different types of monological texts, such as personal diaries (with mostly private content), family books (recording events of family life), account books and private chronicles with irregular entries (thus hardly 'journals' in the strict sense) that comprise events of family and village life, interspersed with weather reports and news about wars and accidents.

In addition to these types, there is yet another type of diaries, written in travel settings and fairly inconsistently labelled as *reisdagboeken* 'travel diaries', *reisjournalen* 'travel journals' or *reisverslagen* 'travelogues'. In their comprehensive inventory of Dutch travelogues, Lindeman et al. (1994: 10) address the vague character of these categories:

> De grenzen met sommige andere genres kunnen vaag zijn. Een dagboek kan bijvoorbeeld overgaan in een reisverslag, en omgekeerd.
>
> 'The boundaries with certain other genres can be vague. A diary, for example, can blend into a travelogue, and the other way round.'

Interestingly, the 3.2 version of ARCHER (*A Representative Corpus of Historical English Registers*), a multi-genre historical corpus of British and American English, introduced a split of the previous single genre 'journals-diaries' into two separate genres 'diaries' and 'journals' (mostly travel journals) (*ARCHER* website; cf. also Yáñez-Bouza 2011, 2016):

> Following the original design of the corpus, the defining criterion for the classification of the materials in ARCHER 3.2 is topic and purpose of the text: diaries record private matters, domestic affairs, everyday activities and routines; journals report on a journey or a task associated with travel (including sea travel and war campaigns) and with political matters. In ARCHER 3.2 there are 122 diaries and 122 journals, of which 105 are travel journals and 17 are political journals.

In the *Going Dutch Corpus*, no such distinction between travel and non-travel settings of diary writing is made. The crucial selection criterion for the categorisation as 'diaries and travelogues' was the personal character of these texts, comprising the writer's own experiences and commentary. Following the approaches in Dekker (1995) and Lindeman et al. (1994), impersonal accounts such as cash account books (*kasboeken*) and ship's log books (*scheepsjournalen*, *logboeken*) were not included in the *Going Dutch Corpus*, as they cannot be regarded as ego-documents. However, Dekker (1995: 277) admits that it is "not always easy to draw the line, which will come as no surprise for a time when the personal and the public spheres were still strongly intertwined". What is more, we have to be aware that the 'personal' character is very often limited to a fairly factual account of daily activities without a considerably high degree of attention given to introspection and intimacy. In fact, texts from the period under investigation can hardly be compared to our present-day understanding of personal diary writing. Baggerman (2011: 465) rightly remarks that many diaries "provide more thorough information about the outside temperature than about the author's inner life".

Apart from their varying terminology, diaries and travelogues also tend to differ in length and layout, ranging from concise telegram-style notes to more comprehensive narrations, as well as from daily to more irregular entries. Generally, text samples of 2,500 words per writer (usually taken from one single document) were randomly selected in order to avoid an overrepresentation of certain writers[31]. For practical reasons, also keeping in mind the limited availability of eighteenth- and nineteenth-century diaries and travelogues, the maximum number of words per writer had to be slightly extended to 2,500 words (as opposed to 2,000 words per writer for private letters) in order to reach the intended corpus size.

As summarised in Table 6, the sub-corpus of diaries and travelogues consists of 140,507 words in all, comprising 71,157 words for the eighteenth-century period and 69,350 words for the nineteenth-century period.

Table 6. General distribution of data in the sub-corpus of diaries and travelogues.

Period	N texts (%)	N words (%)	N writers (%)
1770–1790	26 (52.0)	71,157 (50.6)	25 (50.0)
1820–1840	24 (48.0)	69,350 (49.4)	25 (50.0)
Total	**50 (100)**	**140,507 (100)**	**50 (100)**

All regions are represented by approximately 10,000 words per period. The sub-corpus contains 50 different diaries and travelogues, which were written by 50

[31] Only in some exceptional cases (i.e. for regions where the amount of suitable texts was limited), 5,000 words per writer were transcribed in order to reach the intended corpus size.

different writers[32]. The detailed distribution of data across periods, regions and gender is given in Tables 7a and 7b.

Table 7a. Distribution of data in the sub-corpus of diaries and travelogues across region and gender (P1 = 1770–1790).

P1	**Male**			**Female**			**Total**		
	Texts	*Words*	*Writers*	*Texts*	*Words*	*Writers*	*Texts*	*Words*	*Writers*
FR	4	10,198	4	0	0	0	4	10,198	4
GR	3	10,144	3	0	0	0	3	10,144	3
NB	3	10,156	2	0	0	0	3	10,156	2
NH	3	7,680	3	1	2,392	1	4	10,072	4
SH	4	10,126	4	0	0	0	4	10,126	4
UT	3	7,662	3	1	2,601	1	4	10,263	4
ZE	3	7,633	3	1	2,565	1	4	10,198	4
Total	**23**	**63,599**	**22**	**3**	**7,558**	**3**	**26**	**71,157**	**25**

Table 7b. Distribution of data in the sub-corpus of diaries and travelogues across region and gender (P2 = 1820–1840).

P2	**Male**			**Female**			**Total**		
	Texts	*Words*	*Writers*	*Texts*	*Words*	*Writers*	*Texts*	*Words*	*Writers*
FR	4	10,250	4	0	0	0	4	10,250	4
GR	3	10,061	3	0	0	0	3	10,061	3
NB	1	5,009	1	0	0	0	1	5,009	1
NH	3	6,067	3	3	5,101	3	6	11,168	6
SH	3	7,807	3	2	5,120	2	5	12,927	5
UT	1	5,056	1	2	4,727	2	3	9,783	3
ZE	3	10,152	3	0	0	0	3	10,152	3
Total	**18**	**54,402**	**18**	**7**	**14,948**	**7**	**25**	**69,350**	**25**

[32] The apparent 1:1 ratio needs some further explanation, though. On the one hand, one nineteenth-century travelogue from North Holland (*Amsterdam_Backer_172_663_dia01*) was actually written by two distinct hands, most probably by a husband (first part) and his wife (second part). A diarist from North Brabant, on the other hand, contributed two different texts (samples) for the eighteenth-century period. Like in the sub-corpus of private letters, writers who contributed data for both time periods were counted as two different persons.

Like the sub-corpus of private letters, this sub-corpus was initially planned as gender-balanced. Unfortunately, research in the visited archives has shown that the distribution of male and female diary writers from the periods under investigation is not balanced at all, which is why the intended gender representation could not be achieved. The final sub-corpus does contain data from at least ten female diarists, though. 22,506 words were written by females, which roughly correspond to 16.0% of the total sub-corpus. However, it must be taken into account that these texts are not equally distributed across periods (i.e. mainly nineteenth century) and regions (i.e. mainly Holland and Utrecht).

3.1.3 Newspapers

In addition to two types of handwritten ego-documents, the multi-genre design of the *Going Dutch Corpus* also incorporates printed and published texts. Unlike private letters (Section 3.1.1) and diaries and travelogues (Section 3.1.2), the genre of newspapers is typically associated with more standardised writing, more closely representing the 'language of distance' in Koch & Oesterreicher's (1985) terms.

With their broad readership and, especially compared to formal and literary texts, a more popular and accessible style of writing, newspapers can certainly be considered as a valuable linguistic source in order to examine variation and change in language practice. Rademann (1998: 49) argues that with regard to the "considerably large target audiences, the language used in newspaper articles is often assumed to be characteristic of the respective period and society they are published in", which makes this genre particularly suitable for diachronic studies.

Another methodological advantage of newspapers and a decisive factor to include them in the *Going Dutch Corpus* is their geographical spread across the language area. In the late eighteenth and early nineteenth centuries, newspapers were still locally produced and distributed, and thus primarily catered to regional readerships. This makes them a particularly interesting printed genre for a (historical-)sociolinguistic approach. In fact, for each of the seven regions in the corpus, a regional newspaper could be selected. Therefore, the sub-corpus of newspapers is as regionally balanced as the two sub-corpora of ego-documents, covering the same seven regions (cf. Section 3.3).

The use of newspapers for a systematic comparison with ego-documents has also been attested before. Percy (2012: 194) argues that "[t]he register of news reportage has an interesting if indirect relationship with everyday language". Notably, the documentation of the *GerManC* corpus (Durrell et al. 2012: 1; cf. also Chapter 3) even classifies newspapers as orally oriented registers, alongside personal letters[33]. In this respect, newspapers are probably best considered as a genre which, on the one hand, displays a printed, edited and fairly standardised

[33] The orally oriented registers in the *GermanC* corpus comprise drama, newspapers, sermons and personal letters, as opposed to more print-oriented registers like narrative prose, scholarly, scientific and legal texts.

form of writing, but on the other hand, represents authentic 'everyday' language and the 'language of immediacy' more closely than, for instance, academic prose or literary works. Elspaß & Niehaus (2014: 51-52) suggest a similar corpus design for German, considering regional newspapers as suitable historical data *from above* as opposed to private letters as historical data *from below*.

The selected newspaper texts, as a rule, comprise news reports only, following Niehaus' (2016: 48) criterion to take into account proper newspaper language, representing the language of editors and correspondents:

> Ich habe außerdem darauf geachtet, möglichst nur Texte zu berücksichtigen, die ‚Zeitungssprache' i.e.S., also die Sprache der für eine Zeitung schreibenden Redakteure und Korrespondenten, wiedergeben.

Therefore, official government announcements, advertisements as well as extensive lists of names, for instance lists of decedents, were categorically excluded from the corpus.

As summarised in Table 8 below, the sub-corpus of newspapers consists of 70,645 words in all, comprising an equal number of words for both periods (i.e. approximately 35,000 words for 1770–1790 and 1820–1840 each) and all seven regions (i.e. 5,000 words per region)[34].

Table 8. Distribution of data in the sub-corpus of newspapers across period and region.

Region	Period 1	Period 2	Total
Friesland	5,025	5,018	10,043
Groningen	5,051	5,064	10,115
North Brabant	5,018	5,036	10,054
North Holland	5,088	5,093	10,181
South Holland	5,048	5,027	10,075
Utrecht	5,040	5,033	10,073
Zeeland	5,053	5,051	10,104
Total	**35,323**	**35,322**	**70,645**

3.2 Diachronic dimension

The diachronic dimension of the *Going Dutch Corpus* is closely linked to the diachronically oriented approach of this dissertation, investigating the possible influence of top-down language policy measures on actual language practice. The

[34] The minor deviations from the limit of 5,000 words are due to the methodological decision not to cut off sentences but to fully transcribe them until the next full stop.

historical event of the Dutch *schrijftaalregeling* in the early 1800s, with Siegenbeek's official orthography and Weiland's official grammar being published in 1804 and 1805, respectively (cf. Chapter 2), serves as a starting point for defining the diachronic cross-sections of the *Going Dutch Corpus*.

In order to gain access to language use before and after this landmark in the history of Dutch standardisation, two periods of twenty years each were defined, with a gap of approximately one generation between these cross-sections. The late eighteenth-century period, spanning the years 1770–1790, represents the generation of language users *before* the national language policy was introduced. Symmetrically, the early nineteenth-century period, i.e. *after* the introduction of Siegenbeek (1804) and Weiland (1805), spans the years 1820–1840, representing the generation of language users which had (probably) been exposed to the language policy measures, as envisaged by the government. For the main research objectives of this dissertation, the diachronic dimension is the most important independent variable of the *Going Dutch Corpus*.

3.3 Spatial dimension

Addressing the importance of space as an external factor, Elspaß (2012: 313) argues that when dealing with "languages […] with considerable regional variation, it is also imperative to consider texts from different regions". This is certainly the case for late eighteenth- and early nineteenth-century Dutch. Ultimately aiming at a regionally balanced representation of all three genres and both time periods, the *Going Dutch Corpus* comprises data from a variety of regions in the Northern Netherlands.

Previous historical-sociolinguistic research on this language area, most notably the *Letters as Loot* programme (Rutten & van der Wal 2014: 11-13; cf. also Nobels 2013: 28-30; Simons 2013: 104-106;), mainly focused on the regions on the western coast of the Northern Netherlands, viz. Holland and Zeeland with their main cities of Amsterdam, Rotterdam, Middelburg and Vlissingen[35]. However, as Rutten et. al (2014b: 12) point out, it is important to avoid the emphasis on specific regions, usually demographic and socio-economic centres and their surroundings, and to consider demographically less important regions as well. For this reason, the *Going Dutch Corpus* adds a new layer of four rather under-studied regions to the three westernmost regions of North Holland, South Holland and Zeeland, expanding the previously investigated language area to the north (Friesland, Groningen), to the east (Utrecht) and to the south (North Brabant).

At the same time, the remaining provinces in the eastern part of the Northern Netherlands were not included in the *Going Dutch Corpus*, mainly but not exclusively for practical reasons. While it was not feasible to compile a corpus that

[35] The focus of the *Letters as Loot* corpus on the western regions of the Nothern Netherlands is due to the prevailing origin of the confiscated letters. The vast majority of letters was sent to and from the provinces of Holland and Zeeland (Rutten et al. 2012: 329; Rutten & van der Wal 2014: 11).

covers the entire language area, the eastern border provinces certainly offer intriguing points of departure for future research.

The spatial dimension of the *Going Dutch Corpus* incorporates two different perspectives. Section 3.3.1 takes into account regional variation on the basis of provincial boundaries. Another distinction is based on demographic and socio-economic differences, focusing on variation between the urbanised centre and the less urbanised periphery, which will be addressed in Section 3.3.2.

3.3.1 Regions

The first variable of the spatial dimension investigates variation across different regions of the Northern Netherlands. For practical purposes, these regional categories were based on present-day provinces and provincial boundaries (cf. also Simons 2013: 104), which, in some cases, deviate from the historical boundaries in the late eighteenth and early nineteenth centuries. The present provinces of North Holland and South Holland, for instance, were part of the province of Holland until its split in 1840. However, previous studies on seventeenth- and eighteenth-century Dutch (e.g. Rutten & van der Wal 2014), have revealed distinct regional patterns in North and South Holland, the latter of which being characterised as a "transitional zone between Holland and Zeeland" (Rutten & van der Wal 2014: 341). Therefore, it seemed both logical and necessary to consider the Holland area before 1840 as two distinct regions.

The following seven regions of the Northern Netherlands are covered in the *Going Dutch Corpus* (listed in alphabetical order): Friesland (FR), Groningen (GR), North Brabant (NB), North Holland (NH), South Holland (SH), Utrecht (UT) and Zeeland (ZE). See Figure 5 for a map of the investigated language area indicating the regions represented in the corpus.

As mentioned before, a balanced representation of all selected regions was envisaged. While this aim was easily achieved in the compilation of the sub-corpus of newspapers (with 10,000 words per region), the compilation of the sub-corpora of handwritten ego-documents (private letters: ideally 30,000 words per region; diaries and travelogues: ideally 20,000 words per region) largely depended on the availability of suitable archival sources.

As a consequence, some regions like North Brabant and also Zeeland comprise slightly less words than socio-economically and demographically more dominant regions like North Holland and South Holland with various big cities and, from a practical point of view, more archives to visit.

Figure 5. Map of the Northern Netherlands indicating the regions represented in the *Going Dutch Corpus*.

Figures 6 and 7 illustrate the well-balanced distribution of data across regions in the sub-corpora of private letters (cf. Tables 5a and 5b) as well as diaries and travelogues (cf. Tables 7a and 7b).

Figure 6. Distribution of data across region in the sub-corpus of private letters.

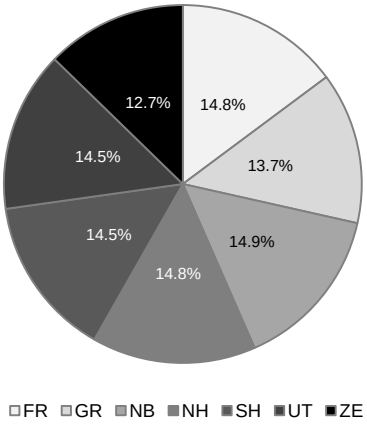
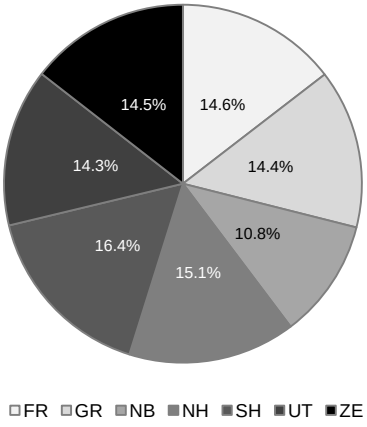


Figure 7. Distribution of data across region sub-corpus of diaries and travelogues.



Another methodological consideration concerns the regional categorisation of texts. Whereas it was a straightforward task to categorise newspapers according to their place of publication[36], ego-documents could not be categorised that easily. The major challenge was to determine the starting point for a reliable categorisation into regions. In the case of private letters, the place from which a letter was sent might serve as an indication, but at the same time, would have been to inaccurate and even misleading. To give an example, a writer who sent a letter from Middelburg (Zeeland) to a close relative in Utrecht was not necessarily a citizen of Zeeland. Moreover, any letter from outside the language area (i.e. outside the seven selected regions and even outside the Northern Netherlands), would have been rejected, even though their writers were actually based in one of the investigated regions. Issues such as travelling, migration and inter-regional marriages further complicate the regional categorisation. Ideally, the so-called 'region of residence', i.e. "the region where a sender was born and raised or where he or she spent most of his or her life" (*Letters as Loot* corpus website) could be traced and identified.

In order to assign letter writers to one of the seven regional categories, the following procedure was applied, listed in descending order of importance:

(1) Based on the names and information given in the letters, basic genealogical and biographical research was conducted online in order to determine the places of birth and death, also taking into account mobility across the lifespan. It was not possible, though, to trace back the background of every single writer. Generally, in-depth genealogical and biographical research of individual writers was beyond the scope of this project.

(2) If the background of a writer, and most importantly the region of residence, could not be identified, the general regional association of the family (and the corresponding family archive) was considered.

(3) Only in a few exceptional cases, i.e. when very little or even nothing was known about a writer, the place from which the letter was sent (as given on the document itself) was used as a tentative indication.

The regional categorisation of diaries and travelogues was based on the first two criteria of the procedure mentioned above.

[36] It should be emphasised, though, that this categorisation takes no account of the regional background of the actual writers of news reports. Given the lack of information about correspondents in eighteenth- and nineteenth-century newspapers, it is impossible to determine who contributed the texts, let alone where these writers came from.

3.3.2 Centre versus periphery

While the categorisation of regions described in Section 3.3.1 was based on provincial boundaries, the spatial dimension of the *Going Dutch Corpus* also integrates a second categorisation on the basis of demographic and socio-economic factors (cf. also Vosters 2011: 207-208). Utilising the variety of regions in the corpus, a distinction was made between the centre (i.e. the more urbanised, demographically and economically stronger regions) and the periphery (i.e. the far less urbanised regions outside the centre). Considering the fact that the "biggest sociogeographical contrast in the Republic was […] not that between town and countryside, but between more and less urbanized provinces" (Kloek & Mijnhardt 2004: 48), the degree of urbanisation was considered as the crucial factor for the grouping of provinces into the two categories of 'centre' and 'periphery'.

Kloek & Mijnhardt (2004: 32) outline the demographic situation in the Northern Netherlands at the turn of the century as follows:

> Around 1800, the contours of what would far later come to be known as the "Randstad," the urban agglomeration of western Holland, were already becoming clear. More than 625,000 people – 30% of the country's population – lived within the area bounded by Amsterdam, Haarlem, Leiden, The Hague, Rotterdam, and Utrecht. Even beyond this perimeter, the next town was never far away, and the countryside was relatively densely populated. Cities were the natural habitat of the average Dutch man or woman of the day.

Based on this outline, the regions of North Holland, South Holland and Utrecht, in which all of the above-mentioned cities are located, make up the centre of the language area. The regions of Friesland and Groningen (in the north) as well as North Brabant (in the south) can be regarded as peripheral with respect to this centre.

In terms of the binary centre–periphery distinction, the seventh province in the corpus, i.e. Zeeland, takes a more ambiguous position. Historically, it clearly belonged to the demographically and economically leading regions in the sixteenth and seventeenth centuries, i.e. during the Golden Age of the Northern Netherlands (together with Holland). However, Zeeland's importance declined in the course of the subsequent centuries, ultimately losing its status as a centre. As Kloek & Mijnhardt (2004: 49) point out, "the Republic's center of economic gravity shifted to the Amsterdam-Rotterdam axis", which left the once flourishing region of Zeeland as one of the victims of this development (ibid.: 33). Therefore, I decided to leave Zeeland out of consideration and to treat it separately. However, the corpus-based case studies in Chapters 5–12 might shed more light on the position of Zeeland, i.e. whether it is linguistically closer to either the centre or the periphery, or whether the empirical investigation actually confirms the ambiguous intermediate position.

3.4 Social dimension

Studying the relation between linguistic variation and its social significance has always been central to sociolinguistic research ever since the emergence of this academic field of study (Nevalainen & Raumolin-Brunberg 2003: 16). In this section, two major social variables will be briefly discussed: social class (3.4.1) as well as gender (3.4.2), the latter of which will be further investigated in the corpus analyses of this dissertation.

3.4.1 Social class

In both present-day and historical sociolinguistics, the variable of social class has often been regarded as "one of the major – if not *the* major – external constraints" (Nevalainen & Raumolin-Brunberg 2003: 133). Investigating seventeenth- and eighteenth-century Dutch, the findings presented in Rutten & van der Wal (2014), based on the *Letters as Loot* corpus, confirmed social class as one of the central independent variables affecting patterns of language variation and change.

On the basis of the well-established historians' model of social stratification in the seventeenth- and eighteenth-century Republic of the Seven United Provinces (1581-1795), letter writers were classified into four social categories: the upper class (UC), the upper-middle class (UMC), the lower-middle class (LMC) and the lower class (LC) (cf. Rutten & van der Wal 2014: 9-10). The classification presented in Table 9 was primarily based on the writers' professions or, in the case of women, on the profession or social position of their husbands or fathers (Rutten & van der Wal 2014: 10; cf. also Nevalainen & Raumolin-Brunberg 2003: 37 for an English example).

Table 9. Social stratification of the seventeenth- and eighteenth-century Republic of the Seven United Provinces in the *Letters as Loot* corpus (cf. Rutten & van der Wal 2014: 10).

	Historians' stratification	***Letters as Loot* corpus**
(1)	Nobility and the non-noble ruling classes	
(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 storekeepers, 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	

Although this model needs to be modified according to the changing social stratification in the late eighteenth- and early nineteenth-century period under investigation, I maintain the suggested four-partite division into upper, upper-middle, lower-middle and lower classes. Initially, it was considered to integrate social class variation in this dissertation as well. However, throughout the exploratory preparation phase and the actual collection of data, it became evident that a representative amount of ego-documents written by lower- and lower-middle-class writers in the periods 1770–1790 and 1820–1840 is practically unavailable in Dutch archives. Whereas the eighteenth-century cross-section of private letters, at least to some extent, could have been covered with data from the *Letters as Loot* corpus, suitable material for the nineteenth-century period turned out to be sparse.

These limitations only emphasise the unique character of the collection of Dutch sailing letters used for the *Letters as Loot* corpus. At the same time, they confirm the arbitrariness of written sources preserved and stored in municipal and regional archives. Schneider (2013: 65, originally quoted from Montgomery 1997: 227) describes them as "products of the 'vagaries and accidents of history (such as which family chose to preserve letters, whether letters survived decay)'". Not surprisingly, those documents which *have* been preserved and kept in the archives to the present day, are more likely to derive from relatively well-to-do families of the middle to the upper classes rather than from the lower ranks of society.

In order to avoid a far too small and therefore hardly representative sample of lower-class and lower middle-class writing, I preferred to compile a well-balanced and socially more homogeneous corpus of eighteenth- and nineteenth century writers from the (upper-)middle to the upper classes. Most importantly, the very highest rank of Dutch society was excluded from the corpus. In fact, even the upper-*middle* class has to be regarded as a proper middle class, which is why these texts do not necessarily contradict the historical-sociolinguistic tradition *from below*. Furthermore, the ego-documents and particularly the sub-corpus of private letters represent a wide range of family archives, often comprising more texts than just from the 'influential' main branch only. The selected texts also represent less central family members of 'minor' or in-law branches of the extended family (Martineau 2013: 141).

With respect to the comparatively homogeneous representation of social ranks in the *Going Dutch Corpus*, the variable of social class will not be considered in this dissertation. Instead, the focus will be on the equally significant social variable of gender (Section 3.4.2).

3.4.2 Gender

Within and across the two sub-corpora of ego-documents (i.e. private letters, diaries and travelogues), it is possible to investigate social variation by focusing on the independent variable of gender. In sociolinguistic research, gender has repeatedly emerged as "one of the most robust social variables" (Nevalainen &

Raumolin-Brunberg 2003: 110) in order to identify and explain patterns in language use by men and women. Even though the categorisation of men and women in the *Going Dutch Corpus* is purely based on their biological sexes, I prefer to use the term *gender* rather than *sex*, taking into account that this social variable primarily focuses on variation based on a social roles and practices rather than on a biologically or physiologically-based distinction (Meyerhoff 2011: 201; Nevalainen & Raumolin-Brunberg 2003: 110). Kiełkiewicz-Janowiak (2012: 313) points out that "linguistic patterns distributed according to the sex of the speaker are to be accounted for by reference to the social characteristics of the speakers (their social roles, their attitudes, their preferences) in the larger societal context". She further argues that the idea of gender as a socio-cultural concept should also be considered when studying language variation in historical contexts (Kiełkiewicz-Janowiak 2012: 313):

> For historical sociolinguistics too it became obvious that, rather than simply indicating the sex of the speaker, researchers should define gender in terms of a set of social roles and characteristics usually ascribed to, and accepted by, women and men within a given society.

Irrespective of this terminological choice, traditional language histories are almost exclusively based on texts by male writers, mostly from the elite and socio-economically leading regions (cf. Chapter 3). Women, on the other hand, "are, as a rule, under-represented" (Kiełkiewicz-Janowiak 2012: 308).

In order to investigate language variation based on gender differences, the *Going Dutch Corpus* initially aimed at a balanced representation of men and women, by including a more or less equal number of words written by male and female writers in the two sub-corpora of ego-documents. Table 10 shows that male writers are overrepresented in the *Going Dutch Corpus* with a total share of two-thirds. However, with respect to the near-absence of female writers in traditional language histories, the gender representation in the *Going Dutch Corpus*, with one-third of the data being written by women, is still a considerable change.

Table 10. General distribution of data across gender and time.

Gender	Period 1	Period 2	Total
Male	127,112 (72.0%)	105,657 (60.5%)	232,769 (66.3%)
Female	49,472 (28.0%)	68,992 (39.5%)	118,464 (33.7%)
Total	**176,584 (100%)**	**174,649 (100%)**	**351,233 (100%)**

Zooming in on the gender representation in the sub-corpus of private letters, Figure 8 shows that gender balance was, in fact, almost achieved for this genre, with a relative distribution of 54.5% (male) versus 45.5% (female) (cf. also Tables 5a and 5b).

Figure 8. Gender distribution in the sub-corpus of private letters.

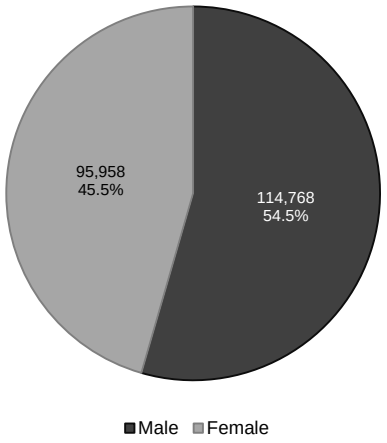
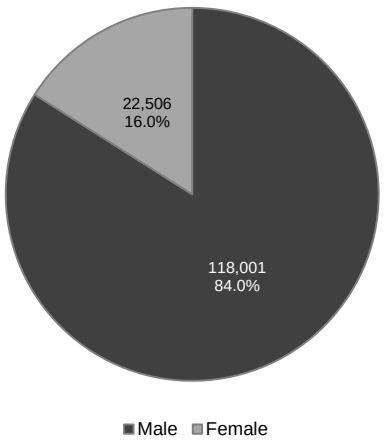


Figure 9. Gender distribution in the sub-corpus of diaries and travelogues.



Unfortunately, a similarly balanced gender representation could not be achieved in the sub-corpus of diaries and travelogues, due to the sparsity of diaries and travelogues written by women. Whereas the major part, namely 84.0% of the words, was written by men, merely 16.0% was written by women, as shown in Figure 9 (cf. also Tables 7a and 7b).

4 Individual dimension and the *Martini Buys Correspondence Corpus*

The fifth variational dimension addresses variation and change on the level of individual language users. Whereas the variables presented in Section 3 considered the community at large, or specific groups of language users (e.g. writers from North Holland versus writers from Friesland, or social groups such as men versus women), this dimension zooms in on the behaviour of individual language users, examining variation *between* each other and *within* their own language practices. In modern sociolinguistics in particular, these variables have commonly been referred to as *inter-speaker* and *intra-speaker variation* (e.g. Meyerhoff 2011: 17). However, working with historical and thus written data implies that one can hardly refer to language users as 'speakers' in the strict sense. In this dissertation, the modified terms of *inter-writer* and *intra-writer* variation will be used in order to refer to variation between and within individuals, respectively.

By taking a micro-level perspective on the linguistic behaviour of individual writers, a number of (partly interrelated) questions will be addressed. Assuming that language users possibly alternate between different realisations of the same linguistic variable, it will be examined how consistent or inconsistent

individual writers were in the use of particular features, both from a synchronic and diachronic point of view. Furthermore, a close comparison of individual patterns and preferences can shed more light on inter-individual differences between close family members.

Specifically for the study of inter- and intra-individual variation and change, a separate, stand-alone corpus of family correspondence was designed and compiled. The *Martini Buys Correspondence Corpus* comprises approximately 64,000 words, distributed across 102 private letters written by eleven family members. Sixteen of these letters are also included in the *Going Dutch Corpus*. Spanning three generations of male and female informants from the second half of the eighteenth and the first half of the nineteenth century, the correspondence corpus takes into account the additional factors of age and individual lifespan changes. Most interestingly, however, it also allows to take a micro-level perspective on the effects of language policy measures on the linguistic behaviour of individual family members. The *Martini Buys Correspondence Corpus* will be presented in more detail in Chapter 13.

5 The *Normative Corpus of the Northern Netherlands*

In addition to the multi-genre *Going Dutch Corpus* as the main corpus for the linguistic analyses of this dissertation (Sections 2 and 3, cf. also Chapters 5–12), and the *Martini Buys Correspondence Corpus* for the study of individual variation and change (Section 4, cf. also Chapter 13), a third corpus of eighteenth-century normative works was compiled, referred to here as the *Normative Corpus of the Northern Netherlands*. In order to determine the normative influence of spelling and grammatical prescriptions on language practices, the quantitative analyses of orthographic and morphosyntactic variables will be complemented by a qualitative study of contemporary metalinguistic discourse. As outlined in Chapter 2 (Section 2), there was a vivid normative tradition in the Northern Netherlands throughout the eighteenth century, i.e. before the official *schrijftaalregeling* of Dutch. Rather than to focus on the direct influence of these two officialised publications (i.e. Siegenbeek 1804, Weiland 1805) alone, I will also take into account the normative preferences and prescriptions laid down in metalinguistic discourse of the preceding eighteenth century. In fact, the codifying character of Siegenbeek's orthography and Weiland's grammar can certainly be regarded as a 'conclusion' of the eighteenth-century normative tradition (van de Bilt 2009: 192). Based on a wide range of eighteenth-century normative publications, gradually paving the way for the national language policy in the early 1800s, developments in actual language use can be related to the possible influence of norms and prescriptions.

The *Normative Corpus of the Northern Netherlands* compiled for this dissertation comprises 31 normative publications on orthographic and grammatical issues, such as spelling guides, grammar books and more general linguistic treatises. The selection of texts can be considered as a (more or less) exhaustive account of normative works published in the Northern Netherlands in the course of the

eighteenth-century, spanning the period of 1699–1805. The texts listed in Table 11 are available either in print or in digital form.

Table 11. The *Normative Corpus of the Northern Netherlands* (1699–1805).

Year	Author	Title [Place of publication]
1699	Francius, Petrus	*Greogorius Nazianzenus, Van de mededeelzaamheidt* [Amsterdam]
1700	van Hoogstraten, David	*Aenmerkingen over de geslachten der zelfstandige naemwoorden* [Amsterdam]
1703	Nylöe, Jakobus	*Aanleiding tot de Nederduitsche taal, om goedt en zuiver Nederduitsch te spreken of te schryven* [Amsterdam]
1705	Hilarides, Johannes	*Nieuwe taalgronden der Neederduitsche taal* [Franeker]
1706	Moonen, Arnold	*Nederduitsche spraekkunst* [Amsterdam]
1707	Verwer, Adriaen	*Linguae Belgicae idea grammatica, poëtica, rhetorica* [Amsterdam] (Translation *Letterkonstige, dichtkonstige en redenkonstige schetse van de Nederduitsche tale*)
1708	Sewel, Willem	*Nederduytsche spraakkonst* [Amsterdam]
1712	Sewel, Willem	*Nederduytsche spraakkonst* (Second edition) [Amsterdam]
1723	ten Kate, Lambert	*Aenleiding tot de kennisse van het verhevene deel der Nederduitsche sprake* [Amsterdam]
1730	Huydecoper, Balthazar	*Proeve van taal- en dichtkunde* [Amsterdam]
1743	van Niervaart, Cornelis	*Oprecht onderwijs van de letter-konst* [Purmerend]
1746	Hakvoord, Barend	*De nieuwe Nederduitse spel-, lees- en schryf-kunst* [Deventer]
1748	van Belle, Jan	*Korte wegwyzer, ter spel- spraak- en dichtkunden* [Haarlem]
1755	van Belle, Jan	*Korte schets der Néderduitsche spraakkonst* [Haarlem]
1758	van Rhyn, Leonard	*Kort begryp der Nederduitsche spel-konst* [Amsterdam]
1761	Elzevier, Kornelis	*Drie dichtproeven […] benevens een proef van een nieuwe Nederduitsche spraekkonst* [Haarlem]
1763	Kluit, Adriaan	*Eerste vertoog over de tegenwoordige spelling der Nederduitsche taal* [Leiden]
1763	Heugelenburg, Martinus	*Klein woordenboek, zijnde een kort en klaar onderwijs in de Nederlandze spel, en leeskonst* [Amsterdam]
1764[37]	de Haes, Frans	*De nagelaten gedichten, en Nederduitsche spraekkunst* [Amsterdam]
1769	van der Palm, Kornelis	*Nederduitsche spraekkunst, voor de jeugdt* [Rotterdam]
1770	Kunst wordt door arbeid verkreegen	*Nederduitsche spraekkunst* [Leiden]
1774	Zeydelaar, Ernst	*Nederduitsche spelkonst* [Dordrecht]

[37] De Haes'*Nederduitsche spraekkunst* was published posthumously in 1764, but had probably been written before or around 1740 (Dibbets 1999: 44).

1776	Tollius, Herman	*Proeve eener Aanleiding tot de Nederduitsche Letterkunst*
1777	Kluit, Adriaan	*Vertoog over de tegenwoordige spelling der Nederduitsche taal* [Leiden]
1776	Stijl, Klaas & Lambertus van Bolhuis	*Beknopte aanleiding tot de kennis der spelling, spraakdeelen, en zinteekenen van de Nederduitsche taal* [Groningen]
1793	van Bolhuis, Lambertus	*Beknopte Nederduitsche spraakkunst* [Leiden]
1799	Wester, Hendrik	*Bevatlyk onderwys in de Nederlandsche spel- en taalkunde, voor de schooljeugd* [Groningen]
1799	Maatschappij tot Nut van 't Algemeen [van Varik, Gerrit]	*Rudimenta, of gronden der Nederduitsche spraake* [Leiden, Deventer & Utrecht]
1799	Weiland, Petrus	*Nederduitsch taalkundig woordenboek* (Introduction) [Amsterdam]
1804	**Siegenbeek, Matthijs**	***Verhandeling over de Nederduitsche spelling, ter bevordering van eenparigheid in dezelve*** **[Amsterdam]**
1805	**Weiland, Petrus**	***Nederduitsche spraakkunst*** **[Amsterdam]**

The overview of normative publications in Table 11 is based on a number of previous studies on eighteenth-century language norms, most notably van de Bilt (2009), Vosters et al. (2010), Rutten (2011) and Simons & Rutten (2014).

6 Procedure and methodological remarks

6.1 Systematic methodological procedure for linguistic analyses

Investigating language norms and language usage in late eighteenth- and early nineteenth-century Dutch, the following Chapters 5–12 present eight corpus-based case studies of five orthographic and three morphosyntactic features, all of which can be considered relevant linguistic issues in the context of the Dutch *schrijftaalregeling*. The official regulations in Siegenbeek's (1804) orthography and Weiland's (1805) grammar, in fact, serve as the starting points for the case studies in this dissertation. Ultimately striving for a sophisticated assessment of the effectiveness of these concrete language policy measures, each linguistic variable will be investigated systematically by following the methodological procedure described below.

In the first part of each chapter, the linguistic variables under investigation will be introduced by providing a summary of the normative discussion by either Siegenbeek (1804) or Weiland (1805). Moreover, this section also introduces the relevant variants that were mentioned and possibly evaluated by Siegenbeek and Weiland. Did they take into account language variation and acknowledge the existence of alternative forms? If so, how explicitly (or implicitly) do they prescribe the officialised variant(s), and on which principles were these choices grounded? Given the high complexity of morphosyntactic variables, the corresponding

chapters briefly outline the variable and its history more generally, before moving on to the discussion of Weiland's (1805) preferences and choices.

In the second part, the officialised norms by Siegenbeek (1804) and Weiland (1805) are placed in the wider context of the eighteenth-century normative tradition. By providing an outline of the preceding discussions and developments in metalinguistic discourse, making use of the *Normative Corpus of the Northern Netherlands* (cf. Section 5), a more fine-grained assessment of Siegenbeek's and Weiland's choices is possible. It will be examined how eighteenth-century variation was represented and commented on, also in comparison to Siegenbeek and Weiland, and which alternative forms were mentioned. Furthermore, this section also discusses whether the officialised choices by Siegenbeek and Weiland were innovative and even radical, or rather grounded on existing preferences, i.e. continuing the eighteenth-century normative tradition.

In the third part, I provide an overview of previous research on the linguistic variable under investigation, establishing links and identifying gaps with regard to the research objectives of the present dissertation.

After having outlined the investigated feature by taking into consideration the corresponding discussions in Siegenbeek or Weiland, as well as in eighteenth-century metalinguistic discourse, the focus shifts to the empirical investigation of actual language usage. Based on the multi-genre *Going Dutch Corpus*, each of the eight linguistic variables will be investigated quantitatively, taking into account the variational dimensions of the corpus (i.e. genre, time, space, gender, cf. Section 3) and, whenever relevant, internal factors potentially conditioning the use and distribution of variants.

In the final section of each chapter, the findings drawn from the corpus analyses will be discussed with reference to the official prescriptions of 1804/1805 as well as the eighteenth-century metalinguistic discourse, aiming to assess the normative influence on variation and change in the use of linguistic variables. By systematically following this methodological procedure in each case study, I seek to measure and determine the overall effectiveness of the national *schrijftaalregeling*.

6.2 Final remarks on statistical methods

Throughout the corpus-linguistic analyses in this dissertation (Chapters 5–12), I will make use of descriptive statistics, presenting quantitative results in the form of cross tabulation and column graphs. With regard to the multidimensionality of the *Going Dutch Corpus*, which takes into account genre, time, space and gender as independent variables, monofactorial statistical tests, such as chi-square tests, t-tests or correlation tests, would hardly do justice to the complex corpus design and the variety of external factors under investigation.

On the other hand, multifactorial approaches, making use of more advanced mixed-effect regression models (e.g. conducted with software tools like *R*) offer intriguing possibilities, such as the representation of systematic interaction patterns between independent variables. In recent years, these more advanced

statistical models have primarily been applied to modern (socio)linguistic data. In the field of historical sociolinguistics, however, the employability of these methods is still being explored. While quantitative (present-day) sociolinguistic studies commonly rely on perfectly balanced data sets, even well-balanced historical-sociolinguistic corpora have natural inconsistencies, which, in turn, present new challenges for statistical methods. Only in the last couple of years, historical sociolinguists have started to further explore whether and in what ways statistical methods and tests can be applied to historical data. Balancing the advantages and disadvantages of new quantitative methods, a number of corpus-linguistic case studies demonstrate the possibilities for future research (e.g. Tagliamonte & Baayen 2012; Mannila et al. 2013; Krug & Schlüter 2013; Nevalainen & Raumolin-Brunberg 2016: ch. 9). At this point, though, methods of statistical data analysis are yet to be established as obligatory parts of historical-sociolinguistic research.

From a more practical point of view, the amount of time that needs to be invested in statistical data analysis would have considerably reduced the amount and variety of linguistic variables investigated in this dissertation. However, in order to assess the effects of language policy on patterns of actual language usage, a substantial number of both orthographic and morphosyntactic case studies appeared to be essential for a sophisticated assessment. Therefore, I have chosen a wider range of linguistic variables over a statistically more advanced method. In fact, I would argue that a thoroughly designed and compiled corpus, aiming at a well-balanced representation of authentic language use, as well as a systematic procedure of both quantitative and qualitative data analysis, can, to a large extent, counterbalance the lack of a statistically advanced method.