The historical development of the Dutch posture-verb progressive construction: including a comparison with German

Okabe, A.

Chapter 2 Methodology

2.1 Introduction

As mentioned in 1.4., one of the main objectives of this research is to trace the development of the Dutch posture-verb progressive construction diachronically to see how it reached its current situation. Since this is a descriptive objective in a historical context, the research calls for the collection of historical data in which to examine the change of the construction over time. Therefore, this research relies on historical texts in the Dutch language as data sources, which are conveniently available in the form of corpora (such as the *Corpus Middelnederlands*). The nature and characteristics of these data sources are explained in detail in this chapter.

The structure of this chapter is as follows. First, the theoretical foundation for using corpora for the study of grammaticalization is discussed (2.2.1.). The subsequent sections describe how the corpora used in this study were selected (2.2.2.) and the criteria according to which the data were extracted (2.2.3.). Details of the corpora and the data extraction methods are provided in 2.3. Each of the corpora is presented in turn, since each corpus has different specifications and data access options. Section 2.4. presents an overview of how the corpora together cover the relevant period, and identifies some unavoidable limitations of the methods employed in this research. Subsequently, 2.5. presents the statistical methods used in the analysis and 2.6. the summary of the chapter.

2.2 Corpus data

2.2.1 The synergistic relationship between corpus linguistics and grammaticalization theory

On the one hand, this research is a corpus-based study, which naturally falls into the domain of corpus linguistics. On the other hand, the phenomenon in focus is grammaticalization. Therefore, this research spans two subfields of linguistics: corpus linguistics and grammaticalization theory.

Corpus linguistics and grammaticalization theory share considerable common ground, and collaboration between the two benefits both parties, according to Mair (2004, 2012). These benefits are aptly summarized by Lopez-Couso (2016) as follows:

> Corpus linguistics provides sound empirical methodology for the recognition and documentation of grammaticalization processes, by making use of computerized corpora and relying on established statistical practices […]. [G]rammaticalization theory helps to bring corpus linguistics beyond the purely statistical domain, "liberating" it from the stigma of being seen as nothing more than 'a cemetery of numbers, – an incoherent compilation of uninterpreted and hence pointless statistics' (Mair 2004: 139). (Lopez-Couso 2016: 7)

Corpus linguistics and grammaticalization theory can therefore take advantage of each other by providing data, and goals for which the data serves, respectively.

One of the major commonalities between these two subfields of linguistics is the importance of frequency (Mair 2004: 121). Studies of grammaticalization generally assume that a linguistic element becomes more frequent as it becomes grammaticalized (e.g. Hopper & Traugott 2003: 126-130, Bybee 2010: Chap. 6, Hoffman 2004: Chap. 5); meanwhile, corpus linguistics provides reliable methods to measure this quantitatively. The way in which frequency data obtained from a corpus can be used to evaluate the grammaticalization process of a construction is demonstrated in Hilpert & Koops (2008). Since that study is also directly relevant to the grammaticalization of the Dutch posture verbs as progressive auxiliaries, it is presented in some detail below.

The 2008 study by Hilpert and Koops investigates the grammaticalization of a pseudo-coordinate construction with the posture verb *sitta* 'to sit' in Swedish. This study assumes that as a particular form becomes more grammaticalized, the frequency of a given linguistic feature associated with the grammaticalized form will increase. This means that the grammaticalization process should be visible as an increase in the frequency of that linguistic feature over time. Table 1 summarizes the authors' predictions concerning the kind of features that would occur more frequently as the Swedish pseudo-coordinate construction became more grammaticalized.

Table 1. Hypotheses regarding more and less grammaticalized sentence patterns of Swedish pseudo-coordination (based on Hilpert & Koops 2008)

	Less grammaticalized	More grammaticalized
(i) sentence without locative elaboration	less frequent	more frequent
(ii) adverbials placed outside the verb sequence	less frequent	more frequent
(iii) object extraction	less frequent	more frequent

Since the study discusses a pseudo-coordinate construction with a posture verb, the hypotheses are related to degree to which the verb is used to express position (as in (i) in Table 1) and the independence of the two conjuncts (as in (ii) and (iii)). The first hypothesis is about the desemanticization of the posture verb over time. Specifically, the authors assume that the posture verb in its postural or locative use normally patterns with a locative modifier. With increasing grammaticalization, the postural/locative meaning becomes backgrounded while the temporal aspect of the verb is gradually foregrounded. As the verb is used as an aspectual marker, its locative meaning is less relevant, so the verb is less likely to occur with locative modification. In this way, the frequency of instances with locative modification could reflect the desemanticization of the posture verb.

The second hypothesis ((ii) in Table 1) concerns the cohesion of the verb sequence. As we saw in the grammaticalization path of the Bulgarian posture-verb progressive construction (cf. section 1.3.2.), the sequence consisting of a posture verb, a connector, and another verb gains syntactic and semantic cohesion as grammaticalization proceeds, which is reflected in the formal adjacency of the three elements. The rule of thumb can be stated as follows: 'the more intervening elements occur between the two verbs, the weaker the conceptual union appears to be' (Hilpert & Koops 2008: 245). An example of a non-cohesive verb sequence with intervening elements and an example of a cohesive verb sequence without intervening elements are given in (1a) and (b), respectively.

(1) a. Stock **satt** en stund tyst **och tänkte** över vad Marstrand hade sagt.
'Stock sat silent for a while and thought about what Marstrand had said'

b. Vi **satt och pratade** ett par timmar.
'we sat and talked for a few hours' (Hilpert & Koops 2008: 248)

In (1a), the adverbials *en stund* 'for a while' and *tyst* 'silently' are placed within the verb sequence and modify an individual conjunct, namely, the first one with *satt* 'sat'. If the adverbial is located outside the verb sequence, as is *ett par timmar* 'a few hours' in (1b), it modifies the whole event described by the two verbs. The latter variant, which supposes the integral interpretation of the two conjuncts, is expected to increase in frequency with a higher degree of grammaticalization.[1]

Third, Hilpert & Koops (2008) formulate a hypothesis on object extraction. Object extraction refers to the phenomenon that the element associated with the second verb appears in clause-initial position, as in (2).

(2) Den där artikeln har jag **suttit och läst** hela dagen.
'that article I have been reading all day' (*ibid.*: 245)

In this example, the noun phrase *den där artikeln* 'that article' is the direct object of the second verb *läsa* 'to read', but it is placed in clause-initial position. Extraction is not possible within regular coordination, as shown in (3).

(3) *Den där artikeln har jag **skrattat och läst** hela dagen.
'that article he has laughed and read all day' (*ibid.*: 245)

This example includes a verb *skratta* 'to laugh' and *läsa* 'to read', which do not form a pseudo-coordinate structure, and is thus ungrammatical with object extraction. The occurrence of object extraction thus indicates the unitary interpretation of the two-verb sequence and also its grammaticalized status. Therefore, with increasing grammaticalization of the posture-verb construction, instances of object extraction are expected to be more frequently observed.

These three hypotheses are verified in the study, meaning that each of the grammaticalized sentence patterns (i-iii) appears gradually more

[1] The same trend is also found in English. A pseudo-coordinate construction with *sit* does not allow an adverbial that intervenes the verb sequence, as shown in *What did the hermit **sit and** (*regularly/*never) **read**?* (De Vos 2005: 27, emphasis mine). Such a phenomenon is also expected to be observed for the Dutch posture-verb construction at its pseudo-coordinate stage (cf. section 3.3.2.).

frequently from around the 14th century up to the 20th century. This gradual increase in frequency is thought to reflect the syntactic and semantic development of the construction and can be regarded as correlating with increasing grammaticalization of the construction over the centuries.

In sum, the study by Hilpert & Koops (2008) on the Swedish pseudo-coordinate construction with the posture verb *sitta* demonstrates that frequency data can serve as a good indicator of how grammaticalized a construction is. At the same time, grammaticalization theory can provide the rationale for measuring the frequency of particular words or expressions. Likewise, the current research draws its rationale from grammaticalization theory and its data from historical corpora, thus further advancing the complementary relationship between these two subfields.

2.2.2 Corpus selection

Employing corpora as data sources implies finding appropriate corpora for the research. Fortunately, there are multiple historical corpora available for Dutch. In selecting the corpora, several points were taken into consideration, in particular, the period and the text type covered by the corpus, the size of the corpus, and the presence or absence of lemmatization and annotation.[2]

This research calls for data from the periods where the posture-verb progressive construction emerged, flourished, changed its form, and reached the state comparable with Modern Dutch. As mentioned in 1.3.3., the construction dates back to Early Middle Dutch (1200–1350); meanwhile, the older *en(de)* construction was still found in the 17th century, before becoming infrequent in the 18th century and being replaced by the modern *te* construction, which in turn started to become frequent from the 17th century. No significant developments are attested in the period after the *te* type became widespread. Therefore, it would be desirable to cover the period

[2] Lemmatization means that all inflectional forms related to one linguistic item are grouped under one lemma. For example, *stands*, *stood*, and *standing* are all tagged with the lemma *stand* in a lemmatized corpus. A corpus with lemmatization enables a search with a lemma, i.e. designating the lemma *stand* and extracting all the conjugated and unconjugated word forms from the corpus. Annotation refers to the information added to a linguistic unit, which is commonly provided in the form of tags. One of the most common types of information is word class, which is typically annotated with part-of-speech (PoS) tags.

from 1200 till around 1800 to trace the major diachronic developments of the construction.[3]

The ideal corpus would be a large, balanced corpus with lemmatization as well as annotation that covers the whole period from the 13th to 18th century. Unfortunately, however, none of the existing corpora meet all these criteria, which means that multiple corpora must be used to cover the relevant time period of time. Available corpora that cover part of the period between the 13th and 18th centuries include:

(4) a. *Deelcorpus (ambtelijke teksten)* of *Compilatiecorpus historisch Nederlands* (1250–1799) compiled by Coussé (2010);
b. *Corpus Gysseling* (13th century) offered by the *Instituut voor de Nederlandse Taal* (Dutch Language Institute, henceforth INT);
c. *Corpus 14de eeuw door Van Reenen & Mulder* (1300–1401) via *nederlab*;
d. *Corpus Middelnederlands* (1250–1550) via *nederlab;*[4]
e. *Corpus Laatmiddel- en Vroegnieuwnederlands* (15th and 16th centuries) via *nederlab*;
f. *Deelcorpus (narratieve teksten)* of *Compilatiecorpus historisch Nederlands* (1575–2000) compiled by Coussé (2010);
h. *Corpus literair Nieuwnederlands* (1600–1999) compiled by Geleyn & Colleman (2015); and
i. *KB Kranten* (1618–1900) offered by *Delpher*.

A pilot survey was conducted to see whether each corpus has sufficient occurrences of the construction in question. Based on this small-scale study, larger corpora are preferred, and ideally corpora with literary texts, since the frequency of the construction in official and legal documents seemed to be

[3] It could also be the case that no instances of the construction with posture verbs are attested in Old Dutch due to the limited amount of data from that period. Since it is difficult to determine whether the construction existed or was becoming grammaticalized in Old Dutch, I provisionally set the starting point of the timeframe to the beginning of Middle Dutch, i.e. the period for which there is plenty of data available and in which the posture-verb progressive construction is attested (Van der Horst 2008: 9.5.1.2.).

[4] This corpus is now also available via the web interface of the INT. (https://corpusmiddelnederlands.ivdnt.org/corpus-frontend/MNL/search/).

limited.[5] Finally, the three corpora that best meet the criteria were selected for this research:

(5) a. *Corpus Gysseling* (only the part with literary texts) for the 13th century
b. *Corpus Middelnederlands* for the 14th, 15th, and 16th centuries
c. *Corpus literair Nieuwnederlands* for the 17th and 18th centuries

Detailed descriptions of each corpus are given in 2.3.

2.2.3 Criteria for data extraction

Before providing detailed information on each corpus, this section describes the kind of sentences extracted from the corpora for this research. The first point to note is that the data are mainly obtained in a form-based manner. This means that no semantic distinctions were made in terms of whether a certain sentence has a progressive meaning or not. It is therefore possible that the database for this research includes sentences that could possibly be interpreted as mono- or bipredicative.

In terms of form, three types of word sequences are relevant: [PV *en(de)* V2], [PV *te* V2], and [PV V2]. As discussed in Chapter 1, the major forms of the construction are [PV *en(de)* V2] and [PV *te* V2]. Additionally, as seen in 1.2.2. and 1.3.3., the omission of the connector is possible, which results in the form [PV V2]. These three types of form are defined in detail below.

[5] Note that in historical linguistics, the texts that tend to be written and preserved (i.e. religious, legal, commercial, and literary texts) are not the kind of texts that reflect daily language (Janda & Joseph 2003: 17). This is due to the fact that writing tends to favor conservatism and reflects changes in spoken language with delay (Janda & Joseph 2003: 17, Andersen 2006: 66). It should also be pointed out that literary texts, mostly verses, show typical characteristics in terms of lexical choice and word order, deviating from spoken language (Nemoianu 1971). At the same time, as Janda & Joseph (2003: 17) put it, 'there is little we can do to change the circumstance that the texts which most often tend to be written and preserved are those which least reflect everyday speech. But we can at least admit our awareness of this situation, and concede that it obliges us to use extreme caution in generalizing from formal documents'. In the same spirit, the present analysis handles data from literary texts with discretion.

The first type consists of a posture verb (*staan* or its Middle Dutch form *staen*; *zitten* or its Middle Dutch form *sitten*; or *liggen*), a coordinating conjunction (either *ende* or *en*), and another verb (referred to as the second verb or V^2). A distance limit between the three elements was set in order to exclude long sentences with too many intervening elements, which would blur the formal and semantic cohesion of the verb sequence and, therefore, presumably fail to contribute to auxiliation or even impede it. This distance limit was set at zero to five intervening words, as shown schematically in (6).

(6) **PV** ($word_1$ $word_2$ $word_3$ $word_4$ $word_5$) ***en(de)*** ($word_1$ $word_2$ $word_3$ $word_4$ $word_5$) $\mathbf{V^2}$

Second, for the pattern [PV *te* V^2], sentences with (i) a posture verb and (ii) an infinitive clause with *te* (e.g. *te wachten* 'to wait') were collected. The *om te* construction was disregarded.[6] Again, there was a limit set on how many words could intervene in the sequence. In this condition, the upper limit was set at seven words between PV and *te* and one word between *te* and V^2.[7] The reason for the one-word limit between the connector and the following verb was that the [*te* V_{inf}] phrase usually allows maximally one intervening word (e.g. *aardappelen te schillen* lit. 'potatoes to peel', or *te aardappelen schillen* lit. 'to potatoes peel'). The structure searched for is presented schematically in (7a). In the clause-final verbal complex, the [*te* V^2]

[6] Infinitives with purpose meaning (cf. section 1.3.3.) did not always co-occur with *om* in Middle Dutch. The co-occurrence with *om* gradually increased in frequency in the 16th century and is almost always attested in the 18th century (Van der Horst 2008: 9.4.3.). This means that the co-occurrence of a posture verb and a *te* phrase without *om* does not necessarily imply that the verb is a progressive auxiliary. Rather, each case should be judged individually for whether the meaning is progressive or final, based on its semantics.

[7] The seven-word limit is set for two reasons. Firstly, since the construction with *te* obviously has a monoclausal structure and is less affected by the number of intervening elements that would blur the cohesion of the verbs, it is theoretically possible to have a large number of intervening words. However, the query with eight or more intervening words became too heavy for the *nederlab* system (used to extract data from the *Corpus Middelnederlands*), such that it could not return any search results. The second reason is that the most of the instances with *te* involve fewer than eight intervening elements according to the data extracted from the *Corpus literair Nieuwnederlands*.

phrase can be placed before the posture verb, as in (7b); this form of the construction is also investigated.[8]

(7) a. **PV** (word$_1$ word$_2$ word$_3$ word$_4$ word$_5$ word$_6$ word$_7$) **te** (word$_1$) **V^2**
b. **te** (word1) **V^2 PV**

Third, sentences with a two-verb sequence of a posture verb and a following verb in the infinitive without a connector (i.e. [PV V^2]) were also extracted. This form is seen in sentences like *hij moest **zitten wachten*** (lit. 'he had to sit wait'), where the posture verb is in the infinitive, and *als zij **liggen slapen*** (lit. 'when they lie sleep'), where the posture verb is in the present tense plural form (cf. section 1.2.2.). In this form, it is not required for the verbs to appear directly adjacent to each other, as intrusions in a clause-final verbal complex were not rare in Middle Dutch and Early Modern Dutch (Van der Horst 2008: 16.3.3). Considering this information, three intervening elements were allowed between the verbs, as illustrated schematically in (8).[9]

(8) **PV** (word$_1$ word$_2$ word$_3$) **V^2$_{inf}$**

The formal criteria represented in (6–8) are very loose and lead to the extraction of many sentences unrelated to the posture-verb progressive construction. Therefore, additional rules were set in order to restrict the selection, as summarized in (9).

(9) a. Both verbs have the same agent regardless of whether it is realized as an overt subject.
b. The second verb is not an auxiliary.

---

[8] Note that the word order in (7b) is not possible in Modern Dutch (**dat ik te wachten zit* lit. 'that I to wait sit') but is attested in my database (e.g. *na de wyze der vrouwen, die te broeijen zat op Labans afgoden* 'after the way of the woman, who sat to honor Laban's idols' [1597]). Considering the comparability of the connector *en(de)* with *te* in some cases (cf. section 1.3.3.), it may theoretically be possible to place the [*en(de)* V^2] phrase before the posture verb, as in [*en(de)* V^2 PV]. This sentence pattern was, however, not found in the data and hence is not included in the discussion.

[9] Ideally, the word limit between the verbs was set to 5, in line with (6), but the query with 4 or more intervening words became too heavy for the *nederlab*'s system that it could not return any search results. Hence, the maximum word limit is set to 3 for this case.

c. The second verb is not in the past unless the posture verb is in the past.
d. The verbs may be modified by the same auxiliary.
e. There is no indication of temporal sequence.
f. The posture verb is not a part of a multiword expression with a noncompositional meaning.

The first of these rules stipulates that a sentence must have the same agent for both of the verbs (i.e. *hij zat aan de tafel en ik bracht hem een kop koffie* 'he sat at the table, and I brought him a cup of coffee' is excluded, but *de man zat aan de tafel en hij las de krant* 'the man sat at the table and he read the newspaper' is permitted). Second, the first verb following the connector may not be an auxiliary (e.g. *staat en is gegaan* lit. 'stands and is gone' and *zit en kan lezen* lit. 'sits and can read' are both excluded). In addition, the second verb may not be in the past tense unless the posture verb is also in the past tense (e.g. *staat en wist* 'stands and knew' is excluded, but *stond en wist* 'stood and knew' is permitted). Furthermore, the verbs can be governed by an auxiliary including a modal verb, but they must be under the same verb (e.g. *zal staan en wachten* 'shall stand and wait' is permitted, but *zal staan en moet wachten* 'shall stand and must wait' is excluded). Additionally, all the sentences with an indication of temporal sequence (e.g. *zat en at* ***toen*** 'sat and ate then') were disregarded. Lastly, instances with multiword expressions including a posture verb with a noncompositional meaning (e.g. *in staden staen*, meaning 'to help') were excluded. This includes idiomatic expressions with an expletive syntactic subject (e.g. *het staat me (niet) vrij … te …* 'I am (not) at liberty to …', *het staat zo geschreven* 'it is written').

Additionally, for *staan*, the sentences in which the posture verb was used as a non-progressive auxiliary or quasi-auxiliary were excluded from the database. These sentences included *staan* used in the meaning of *zullen* 'shall', *moeten* 'must', and *kunnen* 'can' in Middle Dutch,[10] *staan* meaning 'to stop' in Middle Dutch in combination with *laten* 'to let' (e.g. *Laet staen u callen* 'stop your chitchatting' (the *Middelnederlandsch Woordenboek*, (henceforth MNW) headword *staen* I B 3 b α; translation mine), and *staan* in a 'gerundive' use, such as *de spijt staat op zijn gezicht te lezen* 'the regret can be read on his face', where the phrase *staan te lezen* has a meaning like 'can be read/is to be read'.[11] In these cases, the posture verb clearly does not retain

[10] See also the MNW, headword *staen* I C 3, 4a & b, e.g. *Doe* ***stont*** *hem daer niet meer te merrene,* 'then he could not wait anymore'.
[11] Cf. WNT, headword *staan* II A 12, which explains this usage as *in de beteekenis van*

its lexical meaning and is not compatible with a progressive interpretation. Hence, these kinds of instances were excluded from the database. The way in which the instances meeting these criteria are extracted differs between corpora and will be described in the next section.

2.3 Corpus description

In this section, the composition and characteristics of each corpus used are described. This section also elaborates on how the sentences that meet the criteria discussed in 2.2.3. were extracted.

2.3.1 Corpus Gysseling

The *Corpus Gysseling* is a complete collection of official and literary texts written in the 13th century. Since this study focuses on literary texts, the part containing official documents was not used (cf. section 2.2.2.). The literary texts can be further divided into two genres: prose and verse. The corpus is available online via a web application offered by the INT and is annotated with word classes and lemmas that have been manually verified. Since a query in Corpus Query Language (CQL) did not yield the expected results, I used the Simple search interface. I extracted data by entering the lemma of each posture verb, along with a part of speech (PoS) tag for verb (i.e. VRB.*). This search returned all the instances where the lemma or its associated forms occurred as verbs in the corpus, unless it was combined with a clitic. For the forms with a clitic (e.g. *enstaen* (= negator *en* + *staen*)), an additional word-form search was conducted. Subsequently, all the attestations were manually examined based on the criteria presented above in (6-9) and those that met the criteria were entered into the database.

2.3.2 Corpus Middelnederlands

The *Corpus Middelnederlands* is the most extensive corpus available for Middle Dutch and is based on the *CD-ROM Middelnederlands*. It consists of

---

een gerundium 'in the meaning of a gerund'.

literary texts from around 1250 to 1550, with some overlap with those in the *Corpus Gysseling*. The texts which are also included in the *Corpus Gysseling* were excluded in order to avoid double counts. The corpus also has some texts from after 1600, which were not taken into consideration in order to restrict the data source to one corpus per period. Texts with uncertain publication dates were also disregarded. The texts are divided into three genres: prose, verse, and a combination of the two. The corpus is available via the *nederlab* web interface with lemmatization and PoS tag annotation, enabling a CQL query.

The CQL queries used will be presented per formal pattern (6-8) of the construction. Firstly, for the [PV *en(de)* V^2] form as in (6), the CQL query shown in (10) was used (here, *staan* is used as an example).

(10) [lemma="staan"] []{0,5} [lemma="ende"] []{0,5} [pos="WW"]

This query returns a list of sentences with an item associated with the lemma *staan*, followed by an item associated with the lemma *ende* with zero to five intervening elements, and then an item tagged as "WW" (which means that the item is a verb), again with zero to five intervening elements.

In the posture-verb progressive construction, the coordinating conjunction sometimes appears in the reduced form *en* in Middle Dutch (cf. section 1.3.3.), which is incorrectly tagged as a negator in most cases in the corpus. To include these instances, the following query was used, which searches for the word form *en* instead of the lemma *ende*.

(11) [lemma="staan"] []{0,5} [t_lc="en"] []{0,5} [pos="WW"]

The instances with a connector *te* (i.e. the form [PV *te* V^2], see (7)) were extracted using the query shown in (12).

(12) a. [lemma="staan"] []{0,7} [t_lc="te"] []{0,1} [feat.wvorm="inf"]
b. [t_lc="te"] []{0,1} [feat.wvorm="inf"] [lemma="staan"]

As explained in 2.2.3., the query in (12a) allows zero to seven intervening elements between the posture verb and the connector *te* and zero to one intervening elements between the connector and the following infinitive verb. In the clause-final verbal complex, the [*te* V^2_{inf}] clause can be preposed, as in (12b).

Lastly, the cases without a connector (i.e. [PV V²], see (8)) were extracted using the CQL queries shown in (13), which search for a lemma of a posture verb followed by a verb in the infinitive with zero to three intervening elements.

(13) [lemma="staan"] []{0,3} [feat.wvorm="inf"]

Again, all instances were manually examined in terms of the criteria in (9) before being entered into the database.

2.3.3 Corpus literair Nieuwnederlands

The *Corpus literair Nieuwnederlands* is a corpus containing literary texts from the period 1600–1950 (Geleyn & Colleman 2015). The corpus is divided into subparts of 50 years, each including 1.5 to 2 million words from three genres, namely, drama, prose, and non-fiction. The subparts of the corpus covering the periods 1600–1649, 1650–1699, 1700–1749, and 1750–1799 were used for this research. These subparts of the corpus consist of texts written by authors from the northern part of the Dutch-speaking region. The corpus is not enriched with lemmatization or annotation. Therefore, it was necessary to search for each word form for each verb, taking spelling variations into consideration. Additionally, the forms with a clitic (e.g. *staeje* (= *sta* + *je* lit. 'stand + you')) were also searched for. All the sentences were manually inspected in terms of the criteria (6-9) before being included in the database.

2.4 Overview and limitations

As discussed in 2.2.2., the three corpora mentioned in 2.3. were chosen to cover the period from the 13th to the 18th century. This is organized as shown in Table 2, which also indicates the earliest and latest publication years of texts included in each corpus.

Table 2. The periods covered by the corpora

	Middle Dutch				Modern Dutch		
	13th		14th	15th	16th	17th	18th
Corpus Gysseling	1200	1300					
Corpus Middelnederlands			1300		1580		
Corpus literair Nieuwnederlands						1610	1799

The total word counts per century are given Tables 3 and 4.

Table 3. Corpus size (in number of words) for Middle Dutch

	Corpus Gysseling	*Corpus Middelnederlands*		
	13th	14th	15th	16th
prose	135,854	1,384,488	2,988,799	611,649
prose/verse	not applicable	0	0	20,484
verse	446,869	3,060,905	2,288,882	188,039

Table 4. Corpus size (in number of words) for Modern Dutch

	Corpus literair Nieuwnederlands	
	17th	18th
prose	636,043	1,679,791
drama	1,342,318	697,573
non-fiction	1,280,656	882,049

As is clear from Tables 3 and 4, the word counts for the 13th and 16th century are considerably lower compared to the other periods. Although this represents an imbalance of data from different time periods, it was considered important to include all relevant data from the selected corpora for the sake of data volume (cf. section 2.2.2.).

To enable meaningful comparison of the results from corpora of different sizes, the frequencies were normalized (cf. section 4.1.). The normalized frequency is called relative frequency, which is obtained by dividing the absolute frequency (actual count of the occurrences) by the total number of tokens in a corpus and multiplying it by the basis for normalization (for example, one million; Brezina 2018: 43). In addition, the corpora were divided into data sets per century to enable comparison between the periods. Hence, the relative frequency per century is one of the major heuristics adopted in this research.

The fact that corpora differ in size is only one of the problems that emerge from using several corpora as a data source. Corpora also vary in terms of annotation, regional distribution, and text types. The differences in these points can all influence the frequency of the linguistic phenomenon under investigation. From the descriptions of the corpora in sections 2.3.1.–2.3.3., it is evident that the three corpora employed in this research are not all annotated the same way. An effort has been made to minimize the potential influence of this difference by applying a single, uniform set of criteria (summarized in 6-9) to determine the instances from the three different corpora that should be included in the database.[12]

The corpora used in this research also differ in regional coverage. While the *Corpus Gysseling* and the *Corpus Middelnederlands* cover the whole Dutch-speaking area, the *Corpus literair Nieuwnederlands* only covers the northern dialects. Although the language was increasingly standardized in the 17th century and the regional differences were correspondingly decreasing, it should be borne in mind that the database for this research does not reflect the southern varieties in the 17th and 18th centuries. In addition to this difference in regional coverage, Coussé (2010) points out that the language after standardization is not necessarily comparable with the language in the Middle Ages, as the latter is significantly colored by the regional variety of the writer and/or the copyist. Although no regional differences are identified in the literature as influencing the development of the posture-verb progressive construction (except for the modern West Flemish dialects, cf. section 1.2.3.), attention should nonetheless be paid to possible influences of regional variation in the analysis.

In terms of the characteristics of texts, there are two inconsistencies between the *Corpus Gysseling* and the *Corpus Middelnederlands* on the one hand and the *Corpus literair Nieuwnederlands* on the other. The first is the identification of the place and the year of publication. While place and year of publication are determined unambiguously for the Modern Dutch texts, the information for Middle Dutch texts is not always exact and can be controversial.[13] Such inconsistencies are unavoidable, due to extralinguistic

[12] Aside from the different manner of annotation, the quality of annotation also deserves attention. While the annotation quality of the *Corpus Gysseling* is very good, probably because of manual examination, that of the *Corpus Middelnederlands* is not ideal. While the corpus was an invaluable source of data for this research, it should nonetheless be mentioned that some valid instances may have failed to appear in the search results due to inaccurate annotation.

[13] There is also a minor difference between the two Middle Dutch corpora in how the publication year of a given text is determined. The information on the publication

factors such as the popularization of letterpress printing from the mid-15th century onwards and accompanying changes in the manner of publication. The other inconsistency is the text genre. While the Middle Dutch corpora principally provide a bipartite classification of verse and prose, the Modern Dutch corpus consists of three text genres, namely, drama, prose, and non-fiction. The growth in the number of text genres is not merely a matter of classification but reflects a change in the literary world during the Renaissance. As part of the Renaissance, texts of more genres started to be written in Dutch instead of Latin. In a corresponding manner, this period also saw the emergence of new text genres, such as the picaresque novel, the travelogue, the epistolary novel, the historical novel, the novella, and the short story (Coussé 2010: 126). This development inevitably influenced the style and manner of writing, which may be reflected in the presence or absence of certain linguistic features.[14] Hence, attention will be paid to the possible gap between Middle Dutch and Modern Dutch in terms of text genres.

As is apparent from the discussion above, the method used to collect data for this research is not without limitations. Nonetheless, the methodology described here is designed to yield results that are as representative of actual language change as possible, given currently available corpora and technological tools. Furthermore, all methodological shortcomings mentioned here will be taken into consideration when analyzing and evaluating the data.

2.5 Statistical methods

The data extracted from the corpora are mainly analyzed using two statistical methods. The first is Fisher's exact test. Fisher's exact test examines whether there is a statistically significant difference between frequencies of two categories. The statistical significance is indicated as a p-

year in the *Corpus Gysseling* is determined based on the combination of the historical context, the script, and the language (Pijnenburg & Schoonheim 1996: 153f.). Texts in the *Corpus Middelnederlands* are dated for the time the text was handed down ('*handschriftenoverlevering*') and not for when it came into being ('*ontstaansperiode*'; Van Pottelberge 2002: 151).

[14] Paardekoper (1993) indeed points out some differences between formal and informal texts in terms of the occurrence of a certain type of structure with the posture-verb progressive construction.

value. When the p-value is smaller than the threshold (0.05), it is interpreted as reflecting a significant difference between the categories. This test is used to compare frequencies of instances with a certain linguistic feature across time periods.

Fisher's exact test can be used to compare two or more groups; however, in this case, it can only indicate that a significant difference exists, and not where this difference derives from. In other words, it cannot tell us between which groups a significant difference exists when there are more than two groups involved. To determine the origin of the significant difference, an additional test called the pairwise comparison using Fisher's exact test (adjusted using Holm's method) is conducted where necessary. This method compares the values of each group with that of all the other groups (i.e. if there are three groups, the test provides three outcomes).

The second method is Kendall rank correlation. This test is used to evaluate whether two series of values correlate with each other. For this analysis, two statistics are reported: Kendall's tau and a p-value. The former takes a value from -1.0 to 1.0, depending on whether there is negative correlation (-1.0 <= tau < 0.0), positive correlation (1.0 >= tau > 0.0), or no correlation (tau = 0.0). The magnitude indicates how strong the correlation is (e.g. 0.7 indicates a strong positive correlation; -0.07 a weak negative correlation). The p-value indicates whether the tau is statistically significant or not. As above, the p-value threshold for this research is 0.05, meaning that a value of 0.05 or larger is considered as not statistically significant. This test is used to compare frequencies of instances with a certain linguistic feature across time periods. All statistical tests were conducted using the programming language R version 3.6.3 (R Core Team 2018).

2.6 Summary

This chapter has described the data sources and the methods of data analysis used in this research. Since the research is concerned with the historical development of the Dutch posture-verb progressive construction, the data are collected from the three historical corpora: the *Corpus Gysseling*, the *Corpus Middelnederlands*, and the *Corpus literair Nieuwnederlands*. Of the data extracted from these three different corpora, only the instances that met a single, uniform set of criteria were entered into the database. These are further analyzed using two statistical tests: Fisher's exact test and Kendall rank correlation. In the next chapter, the putative grammaticalization path of

the posture-verb progressive construction and the accompanying hypotheses will be presented.