Een digitaal compilatiecorpus historisch Nederlands

Evie Coussé, *Vakgroep Nederlandse Taalkunde, Universiteit Gent, Gent, België (evie.cousse@ugent.be)*

Samenvatting: In deze bijdrage wordt een digitaal compilatiecorpus historisch Nederlands voorgesteld dat historische teksten uit verschillende bestaande bronnenverzamelingen samenbrengt tot een methodologisch verantwoorde steekproef die de gehele geschiedenis van het Nederlands overspant. Het compilatiecorpus bestaat uit twee deelcorpora die elk een stuk van de gehele tijdspanne bestrijken en die elk één bepaald type taalgebruik weerspiegelen. Het deelcorpus ambtelijke teksten beoogt het regionaal gekleurde taalgebruik te representeren dat typisch is voor de geschreven taal uit de middeleeuwen en bevat een selectie van lokale ambtelijke teksten uit de drie dialectstreken Vlaanderen, Brabant en Holland voor de tijdspanne van 1250 tot 1800. Het deelcorpus narratieve teksten beoogt dan weer het geschreven taalgebruik in Holland evenwichtig te representeren vanaf het einde van de zestiende eeuw tot vandaag. De evenwichtige samenstelling van het compilatiecorpus maakt het mogelijk om taalveranderingen ononderbroken in de tijd en ruimte te volgen vanaf het vroegste Nederlands tot vandaag.

Sleutelwoorden: CORPUS, CORPUSVERZAMELING, CORPUSMETHODOLOGIE, DIGITALISERING, AMBTELIJKE TEKSTEN, NARRATIEVE TEKSTEN, HISTORISCHE TAALKUNDE, NEDERLANDS, MIDDELNEDERLANDS, NEDERLANDSE DIALECTEN

Abstract: A Digital Compilation Corpus Historical Dutch. In this article, a digital compilation corpus of historical Dutch is presented that brings together historical texts from different source collections in a methodologically motivated sample spanning the whole history of Dutch. The compilation corpus consists of two subcorpora which each covers a part of the complete time span and which each reflects one type of specific language use. The subcorpus of chancellery texts aims to represent the regionally coloured language use typical of the written language of the Middle Ages and contains a selection of local chancellery texts from the three dialect regions Flanders, Brabant and Holland for the time span from 1250 to 1800. The subcorpus of narrative texts aims to give a balanced representation of the written language use in Holland from the end of the sixteenth century to the present day. The balanced composition of the compilation corpus enables the user to follow language changes uninterruptedly in time and space from the earliest Dutch to the present day.

Keywords: CORPUS, CORPUS COLLECTION, CORPUS METHODOLOGY, DIGITALIZING, CHANCELLERY TEXTS, NARRATIVE TEXTS, HISTORICAL LINGUISTICS, DUTCH, MIDDLE DUTCH, DUTCH DIALECTS

1. Inleiding

De taalkundige studie van het historische Nederlands staat of valt met de beschikbaarheid van betrouwbare bronnen die ons toegang geven tot het taalgebruik van vroegere generaties. De voorbije decennia zijn steeds meer historische bronnen digitaal ontsloten, wat hun toegankelijkheid en bruikbaarheid voor taalkundig onderzoek aanzienlijk heeft verhoogd. Het klassieke voorbeeld van een gebruiksvriendelijke collectie met elektronische historische teksten is de cd-rom Middelnederlands, die naast het gedigitaliseerde Corpus Gysseling (dat alle overgeleverde Middelnederlandse teksten uit de dertiende eeuw verzamelt) ook een selectie van Middelnederlandse rijm- en prozateksten bevat uit latere eeuwen. Daarnaast is de laatste jaren ook het belang van het internet gegroeid bij de digitale ontsluiting van historische bronnen. Ik denk hierbij bijvoorbeeld aan de Digitale Bibliotheek voor de Nederlandse Letteren (www.dbnl.org), die toegang biedt tot een steeds groeiende verzameling klassiekers uit de Nederlandse literatuur in digitaal formaat. Voor een exhaustiever overzicht van andere digitale bronnen verwijs ik naar Coussé (2007).

Hoewel tegenwoordig steeds meer historische bronnen digitaal ter beschikking zijn, is het niet zo dat we een taalkundig fenomeen met wat eenvoudige muisklikken zomaar kunnen onderzoeken vanaf het vroegste Middelnederlands tot vandaag. Om te beginnen is er het praktische probleem dat het bronnenmateriaal zich verspreid over verschillende cd-roms en internetlocaties bevindt. Als we die bronnen allemaal in ons historisch onderzoek van het Nederlands willen betrekken, moet dat materiaal in de eerste plaats dus op één plek gecentraliseerd worden. Daarnaast is er de bijkomende praktische complicatie dat de teksten op de cd-roms en het internet doorgaans niet op dezelfde manier geannoteerd werden (als er al een annotatie beschikbaar is) en/of niet in hetzelfde formaat gedigitaliseerd zijn. Als we met een paar simpele muisklikken dat materiaal willen doorzoeken, moet er dus gestreefd worden naar een uniforme annotatie en codering van de digitale gegevens. Ten slotte is er nog het methodologische probleem dat het voor ernstig taalhistorisch onderzoek doorgaans niet volstaat om zomaar wat historische teksten samen te brengen. Voor betrouwbare resultaten is het van belang dat er een evenwichtig corpus samengesteld wordt, waarbij de invloed van buitentalige variatie (bv. uiteenlopende herkomst of datering van de teksten) uit het corpus is geweerd of op een gecontroleerde manier in de tekstverzameling is geïntegreerd. We moeten dus met andere woorden een betrouwbare steekproef samenstellen van historische teksten die een welbepaald soort historisch Nederlands representeert.

In wat volgt, zal ik een digitaal compilatiecorpus historisch Nederlands (verder kortweg compilatiecorpus) voorstellen dat een antwoord wil bieden op de bovenstaande praktische en methodologische problemen. Het corpus is oorspronkelijk samengesteld in het kader van mijn doctoraatsonderzoek naar woordvolgordepatronen in tweeledige werkwoordclusters (Coussé 2008). Ik zal de lezer stap voor stap de selectiecriteria voorleggen die gehanteerd zijn bij de

samenstelling van het corpus. Deze bijdrage kan dan ook in de eerste plaats gelezen worden als een theoretische reflectie over de samenstelling van een — zo ideaal mogelijk — longitudinaal corpus historisch Nederlands. Daarnaast kan deze bijdrage ook gebruikt worden als een praktische handleiding bij het compilatiecorpus dat gratis te downloaden is via de tst-centrale (www.tst.inl.nl). Door het nieuw samengestelde corpus online aan te bieden wil ik andere historische taalkundigen de kans bieden onmiddellijk aan de slag te gaan met een betrouwbaar corpus historische teksten zonder zelf eerst al te veel tijd te verliezen met het vele praktische en methodologische voorbereidende werk dat komt kijken bij het samenstellen van een eigen corpus. Ik wil evenwel benadrukken dat het corpus dat in deze bijdrage wordt voorgesteld slechts een methodologisch verantwoorde compilatie is van reeds uitgeven historische bronnen in digitale en papieren vorm. Het samenstellen van een echt nieuw corpus dat de beschikbare digitale bronnen systematisch aanvult met het vele bronmateriaal dat nog in de archieven ligt, blijft werk voor een veel groter opgezet corpusproject in de toekomst.

2. Methodologische uitdagingen

Het samenstellen van een corpus historische teksten dat de hele geschiedenis van het Nederlands bestrijkt van de vroegste doorlopende teksten uit de dertiende eeuw tot het moderne Standaardnederlands is beslist geen sinecure. In wat volgt, zal ik ingaan op een tweetal ontwikkelingen in de geschiedenis van het Nederlands die het bijzonder lastig maken een longitudinaal corpus te compileren dat de hele vooropgestelde tijdspanne bestrijkt én dat voldoende evenwichtig samengesteld is qua taalgebruik en tekstgenre (vgl. ook Van der Horst 1997).

Om te beginnen zorgt de ontwikkeling van de geschreven standaardtaal ervoor dat het taalgebruik uit de middeleeuwse bronnen moeilijk te vergelijken is met de schrijftaal uit latere tijden (o.a. Van der Wal 1992, Van der Sijs 2004). In de middeleeuwen wordt de geschreven taal nog sterk gekleurd door het persoonlijke dialect van de auteur en/of de kopiist. In de loop van de zestiende en zeventiende eeuw wordt de schrijftaal onder impuls van de renaissance en de ontwikkelingen in de boekdrukkunst beregeld en gestandaardiseerd naar het voorbeeld van de klassieke talen (Van der Wal 1995). Hierbij is de geschreven standaardtaal vooral geënt op de Hollandse volkstaal, hoewel talige invloeden van zuidelijke en ook oostelijke immigranten niet onderschat mogen worden (voor discussie, zie Hendriks 1998, Van der Sijs 2004). We hebben dus te maken met een ontwikkeling van een weinig geüniformeerde schrijftaal waarbij de regionale herkomst van de auteur en/of de kopiist niet weg te denken is naar een meer gestandaardiseerde schrijftaal waarin regionale variatie zo veel mogelijk geëlimineerd is. Bij de selectie van de teksten voor het compilatiecorpus moet met die externe geschiedenis van het geschreven Nederlands voldoende rekening gehouden worden.

Naast het ontstaan van een geschreven standaardtaal zorgen ook enkele literaire ontwikkelingen tijdens de renaissance voor een breuk in het aanbod teksten uit de middeleeuwen en latere eeuwen. Zo gaat het Nederlands vanaf de renaissance het Latijn steeds meer vervangen als de prestigieuze taal voor wetenschap en cultuur (Van der Wal 1995). Dat betekent dat naast de tekstgenres die in de middeleeuwen al in de volkstaal geschreven werden (bv. epische gedichten, liederen, kluchten, receptenboeken, ambtelijke teksten) ook veeleer elitaire tekstgenres (bv. wetenschappelijke, politieke en religieuze traktaten, klassieke tragedies) steeds meer opgesteld worden in het Nederlands. Daarnaast zien we vanaf de renaissance naast de traditionele berijmde literatuur ook heel wat nieuwe tekstgenres ontstaan in prozavorm zoals de schelmenroman, het reisverhaal en later ook de briefroman, de historische roman, de novelle, het kortverhaal, enz. Beide literaire ontwikkelingen tonen aan dat het niet vanzelfsprekend is om één type literaire teksten te selecteren dat al acht eeuwen lang bestaat én dat al die tijd ook in het Nederlands werd geschreven.

3. Samenstelling van het compilatiecorpus

De geschetste ontwikkelingen in de geschiedenis van het Nederlands tonen dat het bijzonder moeilijk is — misschien zelfs in principe onmogelijk — om een longitudinaal corpus historisch Nederlands samen te stellen dat homogeen is op het vlak van taalgebruik en teksttype. Toch hoeft die methodologische uitdaging ons niet voor een volstrekte impasse te plaatsen. In wat volgt, zal ik uit de doeken doen hoe ik met de samenstelling van het Compilatiecorpus Historisch Nederlands een antwoord heb proberen te bieden op de geschetste complexe uitgangssituatie. Ik wil evenwel benadrukken dat de samenstelling van het compilatiecorpus slechts één mogelijke oplossing biedt voor de geschetste methodologische uitdagingen. Concreet heb ik in plaats van tevergeefs naar één homogeen corpus te streven voor de hele vooropgestelde tijdspanne twee deelcorpora verzameld die maar een stuk van de tijdspanne bestrijken en die elk één bepaald type taalgebruik weerspiegelen. Het eerste deelcorpus beoogt meer bepaald het regionaal gekleurde taalgebruik te representeren dat typisch is voor de geschreven taal uit de middeleeuwen. Het tweede deelcorpus richt zich dan weer op documenten in de standaardtaal vanaf het einde van de zestiende eeuw tot vandaag. In wat volgt, zal ik de precieze samenstelling van beide deelcorpora verder toelichten.

Het eerste deelcorpus wil zoals gezegd een representatieve steekproef vormen van de regionaal gekleurde schrijftaal uit de middeleeuwen. Ik herinner eraan dat er in de middeleeuwen nog geen sprake is van een verregaande standaardisatie in spelling, woordkeuze en grammatica zoals in de moderne schrijftaal, maar dat de geschreven taal nog sterk beïnvloed wordt door het persoonlijke dialect van de auteur. Voor een regionaal gediversifieerde steekproef is het nu van essentieel belang dat de verzamelde teksten betrouwbaar in de ruimte gelokaliseerd kunnen worden. Hier wringt helaas het schoentje bij

een aanzienlijk deel van de overgeleverde teksten uit de middeleeuwen. Veelal kennen we de identiteit van de middeleeuwse auteurs niet waardoor het gissen blijft naar de precieze herkomst van de schrijver of de plaats waar de tekst neergeschreven is. Toch zijn er wel degelijk een aantal middeleeuwse tekstgenres die betrouwbaar te lokaliseren zijn op basis van hun expliciete vermelding van de plaats en tijdstip van het schrijven in het document zelf (Van Loon 2002). Het gaat om egodocumenten zoals brieven of dagboeken en om ambtelijke documenten zoals oorkonden en processtukken die normaal gezien steeds een precieze vermelding van de tijd en plaats van het schrijven bevatten.

Het ligt voor de hand om van die goed gelokaliseerde bronnen de meest spontaan geformuleerde tekstgenres te kiezen zoals brieven of processtukken, waarvan we kunnen verwachten dat ze het dichtst aanleunen bij de spreektaal. Helaas zijn precies die spontane tekstgenres jarenlang relatief verwaarloosd bij de ontsluiting van de archieven door historici en taalkundigen. We mogen dan wel beschikken over enkele waardevolle uitgaven van de dagboeken of de briefwisseling van een aantal historische figuren, het aanbod echter is verre van groot én gediversifieerd genoeg om een representatieve steekproef van het taalgebruik uit de middeleeuwen samen te stellen.

Gelukkig is het met de ontsluiting van goed gelokaliseerde kanselarijdocumenten zoals oorkonden en statuten beter gesteld. Voor de hele middeleeuwse periode (en zelfs voor de daaropvolgende eeuwen tot het einde van het Ancien Régime in 1795) is de tekstproductie van de belangrijkste stedelijke kanselarijen uitgegeven in de vorm van grote tekstcollecties in zowel papieren als digitaal formaat. Het spreekt voor zich dat dergelijke rijke verzamelingen ambtelijke teksten een erg toegankelijke en gemakkelijk aan te boren bron vormen bij het samenstellen van het compilatiecorpus. Helaas zijn ambtelijke teksten dan weer in een formele en zelfs archaïsche stijl geschreven die mogelijk ver van het gewone taalgebruik uit die tijd staat. Toch kan de taalkundige studie van ambtelijke teksten ons wel degelijk iets leren over de geschiedenis van de Nederlandse taal. In onderzoek van Coussé (2010) is bijvoorbeeld aangetoond dat zelfs de erg frequente en bijzonder stereotiepe wending *zoals voorzeid is* onderhevig is aan diachrone veranderingsprocessen zoals grammaticalisatie. Soortgelijke bevindingen zijn door Marynissen (1999) en Boonen (2005) gepresenteerd voor de begin- en eindformules van oorkonden, waar heel wat meer vormvariatie optreedt dan doorgaans verwacht wordt van de stereotiepe delen van ambtelijke teksten.

Na heel wat wikken en wegen blijken ambtelijke teksten het best geschikt te zijn voor het invulling voor het eerste deelcorpus. Concreet heb ik voor originele ambtelijke teksten gekozen die zo goed mogelijk het lokale taalgebruik tussen de burgers uit de stad weerspiegelen. Het gaat om schepenbrieven, statuten en reglementen, lokale rechtspraak, verkoopsovereenkomsten, huurcontracten, schenkingen, huwelijkscontracten en wilsbeschikkingen. Ambtelijke communicatie tussen de lokale overheid en het hogere gezag zoals de graaf, hertog, koning of keizer komt niet in aanmerking omdat de herkomst van de

klerk veel onzekerder is en ook het taalgebruik meer bovengewestelijke kenmerken kan vertonen. Gezien de mogelijke dialectvariatie heb ik ambtelijke teksten gekozen uit drie centrale dialectstreken: enerzijds Vlaanderen en Brabant vanwege de vroege en rijke schriftelijke productie in de volkstaal en anderzijds Holland voor een optimale vergelijking met de latere standaardtaal (cf. *infra*).

Complementair met het corpus ambtelijke teksten, heb ik ook een deelcorpus met teksten verzameld die het meer gestandaardiseerde taalgebruik vanaf het einde van de zestiende eeuw moet weerspiegelen. De invulling van het tweede deelcorpus zorgt voor heel wat minder methodologische moeilijkheden dan het geval was bij het hoger beschreven ambtelijke deelcorpus. Om te beginnen hoeft bij het verzamelen van het tweede deelcorpus geen rekening meer gehouden te worden met mogelijke dialectvariatie in de teksten, gezien de regionale nivellering van het geschreven taalgebruik bij de ontwikkeling van de standaardtaal. Voor een optimale vergelijkbaarheid met de ambtelijke teksten heb ik evenwel uitsluitend teksten uit Holland gekozen, waar sinds het begin van de zeventiende eeuw bovendien een grote tekstproductie in druk ontstaan is. Daarnaast is het aanbod teksten die betrouwbaar in tijd en ruimte te plaatsen zijn niet beperkt tot egodocumenten en ambtelijke documenten, zoals dat in de middeleeuwen het geval was. Met de renaissance treden auteurs immers uit de anonimiteit zodat we een zicht krijgen op de herkomst van de schrijver en ook het moment van schrijven. Concreet zijn voor het tweede deelcorpus teksten geselecteerd waarvan bekend is dat de auteur een geboren en getogen Hollander is.

Bij de uiteindelijke selectie van teksten is ervoor geopteerd om in het tweede deelcorpus enkel prozateksten op te nemen. Ik herinner eraan dat er vanaf de renaissance naast de traditionele berijmde literatuur ook heel wat nieuwe tekstgenres ontstaan in prozavorm. Door enkel prozateksten toe te laten in het tweede deelcorpus sluit het taalgebruik van dat deelcorpus dichter aan bij het corpus ambtelijke teksten dat enkel prozateksten bevat. Op die manier verhoogt de homogeniteit van het gehele compilatiecorpus aanzienlijk en is het taalgebruik van beide deelcorpora tot op zekere hoogte te vergelijken. Om te benadrukken dat de prozateksten in het tweede deelcorpus een meer narratief — veelal zelfs een literair — karakter hebben dan de ambtelijke teksten in prozavorm zal ik in wat volgt naar het tweede deelcorpus verwijzen als het corpus narratieve teksten.

4. Compilatie van een evenwichtige steekproef

Nu de samenstelling van het compilatiecorpus uitgebreid verantwoord is, zal ik verder ingaan op de concrete invulling van het deelcorpus ambtelijke teksten en het deelcorpus narratieve teksten aan de hand van bestaande uitgegeven collecties historische teksten. Na een korte introductie van de geraadpleegde tekstverzamelingen zal ik uitvoerig ingaan op de selectie van de teksten uit die

grote corpora om zo tot een evenwichtige steekproef historische teksten te komen in overeenstemming met de methodologische keuzes uit de vorige twee paragrafen.

4.1 Deelcorpus ambtelijke teksten (1250–1799)

Bij de compilatie van het deelcorpus ambtelijke teksten kon ik voor de dertiende en veertiende eeuw putten uit twee gedigitaliseerde corpora met ambtelijke bronnen, nl. het eerder genoemde Corpus Gysseling en het Corpus Van Reenen-Mulder. Daarnaast bevat het deelcorpus ambtelijke teksten ook een selectie gescande teksten uit papieren tekstedities van de kanselarijbronnen van de grotere steden. In wat volgt, zal ik kort ingaan op de bijzonderheden van die drie broncollecties voor het deelcorpus ambtelijke teksten.

4.1.1 Corpus Gysseling (1250–1299)

Voor de oudste teksten in het ambtelijke deelcorpus heb ik een beroep kunnen doen op het Corpus van Middelnederlandse Teksten, een exhaustieve verzameling van alle overgeleverde literaire en ambtelijke teksten tot en met het jaar 1300. Die enorme tekstcollectie is in de jaren zeventig samengesteld door Maurits Gysseling, en staat daarom ook bekend als het Corpus Gysseling. Meer details over de precieze samenstelling van het corpus en over de manier waarop de teksten getranscribeerd werden, zijn na te lezen in Gysseling (1977). De oorspronkelijke papieren uitgave van het dertiende-eeuwse corpus is later door het Instituut voor het Nederlandse Lexicologie te Leiden gedigitaliseerd. Meer informatie over dat digitaliseringsproject, waarin het corpus ook van lemmatisering en woordsoortinformatie voorzien werd, is terug te vinden bij Pijnenburg en Schoonheim (1998). De digitale editie van het Corpus Gysseling is voor het grote publiek toegankelijk via de gebruiksvriendelijke interface van de al genoemde cd-rom Middelnederlands. Voor de samenstelling van het compilatiecorpus kon ik echter rechtstreeks gebruik maken van de achterliggende tekstbestanden die door het INL verstrekt zijn in het najaar van 2004.

4.1.2 Corpus Van Reenen-Mulder (1300–1399)

De ambtelijke teksten voor de veertiende eeuw zijn afkomstig uit het Corpus Veertiende-eeuwse Middelnederlandse Oorkonden, dat verzameld en gedigitaliseerd werd aan de Vrije Universiteit Amsterdam. Die oorkondeverzameling staat beter bekend als het Corpus Van Reenen-Mulder, genoemd naar de vroegste samenstellers van het corpus (Van Reenen en Mulder 1993). Later is het corpus verder aangevuld met meer zuidelijk tekstmateriaal in samenwerking met de Universiteit Gent. In tegenstelling tot het Corpus Gysseling bevat het Corpus Van Reenen-Mulder slechts een selectie van alle overgeleverde ambtelijke teksten uit de veertiende eeuw, aangezien de productie van ambte-

lijke teksten in de volkstaal tegen die tijd al te omvangrijk geworden is voor een integrale uitgave. We hebben dus te maken met een gecontroleerde steekproef, waarbij de veertiende-eeuwse ambtelijke teksten zoveel mogelijk in tijd en ruimte gespreid zijn. De digitale transcriptie van de ambtelijke teksten is net zoals de digitale editie van het Corpus Gysseling verrijkt met lemmatisering en woordsoortinformatie. In afwachting van de publieke release van het corpus, kon ik voor de samenstelling van het compilatiecorpus gebruik maken van een voorlopige versie van het corpus die in het najaar van 2004 door Piet van Reenen ter beschikking is gesteld.

4.1.3 Geschiedkundige rechtsbronnen (1400–1799)

Voor de periode vanaf de vijftiende eeuw zijn helaas geen digitale corpora met ambtelijke teksten beschikbaar, zodat ik voor het compilatiecorpus een beroep heb moeten doen op papieren tekstedities van de kanselarijbronnen van de grotere steden. Die tekstedities zijn tot stand gekomen omstreeks het einde van de negentiende eeuw in het kader van de grootschalige ontsluiting van de ambtelijke bronnen uit de stedelijke archieven door historici uit Nederland en België. Kenmerkend voor de geschiedkundige tekstedities is dat er slechts zelden commentaar gegeven wordt over hoe het tekstmateriaal verzameld is, over de manier waarop de handschriften precies getranscribeerd zijn en over de werkwijze waarop eventuele onduidelijke passages zijn opgelost. Soms kan men wel een voetnoot in de tekstuitgave aantreffen waarin melding gemaakt worden van een onleesbaar woord door bijvoorbeeld een watervlek, maar in tegenstelling tot het Corpus Gysseling en het Corpus Van Reenen-Mulder kan men bezwaarlijk van een echt diplomatisch verantwoorde uitgave spreken.

Een aantal van de tekstedities documenteren niet alleen de middeleeuwse rechtsgeschiedenis maar bieden ook een overzicht van de stadsadministratie vanaf de middeleeuwen tot aan het einde van het Ancien Régime (in de Nederlanden tot 1795). Ik heb van de gelegenheid gebruik gemaakt om het corpus ambtelijke teksten, dat in principe bedoeld is om de regionale variatie in het middeleeuwse taalgebruik te representeren, waar mogelijk aan te vullen met jongere ambtelijke documenten zoals notariële akten, stadsreglementen en officiële brieven uit de zeventiende en achttiende eeuw. Dergelijk continu aanbod ambtelijke teksten maakt het mogelijk om taalkundige tendensen ononderbroken te volgen binnen een homogeen tekstgenre vanaf de dertiende tot de achttiende eeuw in de dialectregio's Vlaanderen, Brabant en Holland.

4.1.4 Compilatie van de bronnen tot een evenwichtige steekproef

Nu de bijzonderheden van de drie gebruikte bronnen voor het deelcorpus ambtelijke teksten kort zijn voorgesteld, kan in meer detail ingegaan worden op de precieze compilatie van het deelcorpus tot een verantwoorde steekproef van het taalgebruik uit de middeleeuwen en daarna. In wat voorafging, heb ik

geargumenteerd dat het deelcorpus ambtelijke teksten de regionale variatie moet weerspiegelen uit de dialectstreken Vlaanderen, Brabant en Holland. Ik heb dan ook in de gebruikte bronnenverzamelingen enkel teksten geselecteerd uit die drie dialectstreken. Aangezien er in de bronnenverzamelingen een vrij grote spreiding is van de teksten over de verschillende steden heb ik ernaar gestreefd teksten te verzamelen uit een vijftal steden per regio om zo tot een nog betere regionale dekking van het deelcorpus te komen. Concreet zijn de Vlaamse teksten uit de steden Brugge, Ieper, Kortijk, Gent en Oudenaarde afkomstig; de Brabantse teksten komen uit Brussel, Leuven, Mechelen, Antwerpen en Breda; en de Hollandse teksten ten slotte komen uit Dordrecht, Amsterdam, Haarlem, Gouda en Leiden. De keuze voor bovenstaande steden is grotendeels bepaald door het aanbod teksten in de geraadpleegde bronnen. Ik heb ernaar gestreefd voor elk van de vijftien gekozen steden ongeveer tweeduizend vijfhonderd woorden tekstmateriaal per tijdsdoorsnede van vijfentwintig jaar te selecteren. In tabel 1 wordt een overzicht gegeven van het aantal woorden dat verzameld is per tijdsdoorsnede en per stad apart.

Hieruit blijkt dat het aardig gelukt is om een mooie spreiding van de teksten te bereiken op regionaal vlak en zelfs ook op stadsniveau in het deelcorpus ambtelijke teksten. Helaas vertoont de tabel ook nogal wat witte plekken door cellen die leeg gebleven zijn. In wat volgt, zal ik een aantal van die leemtes in de tabel proberen te motiveren.

Om te beginnen blijken de verzamelde teksten uit de dertiende eeuw slechts uit vier steden afkomstig te zijn, nl. Brugge, Gent, Mechelen en Dordrecht. Die tendens kan in verband gebracht worden met de ongelijkmatige verschriftelijking van het Nederlandse taalgebied (Burgers 1995). De productie van ambtelijke teksten in de volkstaal is meer bepaald het vroegst van start gegaan in de Vlaamse steden. Hierdoor was het geen probleem om in het Corpus Gysseling de beoogde hoeveelheid tekstmateriaal te verzamelen voor grote steden als Brugge en Gent. De schrijftraditie in de volkstaal begint later in Brabant en Holland, waardoor er in het Corpus Gysseling enkel vanaf het laatste kwart van de dertiende eeuw voldoende teksten beschikbaar zijn voor de steden Mechelen en Dordrecht. Om de dataschaarste in de dertiende eeuw enigszins op te vangen, heb ik waar mogelijk ook meer tekstmateriaal verzameld in de vier steden dan de vooropgestelde tweeduizend vijfhonderd woorden. Rekening houdend met het ongelijke regionale aanbod in het Corpus Gysseling valt de regionale spreiding van de ambtelijke teksten voor de dertiende eeuw vrij evenwichtig uit in het deelcorpus.

Daarnaast toont de tabel een opvallend tekort aan Brabantse teksten, in het bijzonder vanaf de vijftiende eeuw. Die tekstschaarste kan toegeschreven worden aan een gebrek aan beschikbare papieren tekstedities voor de Brabantse steden. De enkele beschikbare tekstcollecties blijken bovendien hoofdzakelijk de oudste oorkonden te herbergen, waardoor de geschiedkundige tekstverzamelingen voor Brabant weinig nieuwe teksten aandragen naast wat al beschikbaar was via het Corpus Gysseling en het Corpus Van Reenen-Mulder.

Tabel 1: Invulling van het deelcorpus ambtelijke teksten (n = 393 332 woorden)

	Vlaanderen					Brabant					Holland					Totaal
	Brug.	Ieper	Gent	Kortr.	Oud.	Antw.	Breda	Bruss.	Leuv.	Mech.	Amst.	Dordr.	Goud.	Haarl.	Leiden	
1250–74	22829	89	3770							2863						29551
1275–99	20255		16319		699			601		15242		11479			130	64725
1300–24	2153	234	531		1387	2428	650	1932	383			3894	239	425	686	14942
1325–49	1790	1734	1002		2575		1172	1948			360	1377	3230	662	2960	18810
1350–74	1910	1377	1717	286	3070	400	1769	1464			1879	1921	2858	1340	2384	22375
1375–99	1801	2442	1054	511	2677	1415	2507	1519			2786	2662	2909	2891	2653	27827
1400–24	1389	1528	2836	1760	233	82	1328	1420			1543	2735	2975	2797	628	21254
1425–49	2944	1008	1977		602		2860	1664			2767	2346	444	3312	2596	22520
1450–74	3159	309	2368	368			3083				2301	2730	2063	1991	2138	20510
1475–99	2740	423	3002	426	102		2104				2547	1558	1682	2304	2803	19691
1500–24	2045	3205	2864	517	1325	761					2845	2334	2268	2805	1786	22755
1525–49	1601	1618	1280	2311	434	997						2792	1919	2428	1624	17004
1550–74	1624	592	1385	2082	1241	1349		309				2885	1981	2704	2424	18576
1575–99	2422		2497	178	185	890						2197			2168	10537
1600–24	228	1270	1456		192						2890				1728	7764
1625–49	1233	122	2033		659			998			2886				2324	10255
1650–74		2763	2334					3607			2557				2254	13515
1675–99	790								1669						2228	4687
1700–24	825				442										2414	3681
1725–49	1265	2400	351						4812						1498	10326
1750–74	2877	1197	215												2880	7169
1775–99					380										2313	2693
Totaal	75880	22311	51156	8439	16203	8322	15473	15462	6864	18105	25361	40910	22568	23659	42619	393332

Een vergelijkbaar probleem van dataschaarste treedt ook vanaf de zeventiende eeuw voor Holland op. Het merendeel van de papieren tekstedities voor de Hollandse steden loopt maar tot 1600, zodat slechts voor Amsterdam en Leiden ook jongere teksten voorhanden zijn.

Al bij al biedt het deelcorpus ambtelijke teksten een steekproef van ambtelijke teksten verspreid over vijftien steden uit drie verschillende dialectregio's vanaf de tweede helft van de dertiende eeuw tot het einde van de achttiende eeuw. Hoewel zo veel mogelijk getracht is om evenveel tekstmateriaal uit de vijftien steden te verzamelen voor elke tijdsdoorsnede, lieten het huidige aanbod uitgegeven ambtelijke teksten niet toe om een steekproef samen te stellen die dialectologisch onderzoek toe moeten laten tot op stadsniveau. Hiervoor zal bijkomend excerpeerwerk nodig zijn van ambtelijke documenten die dateren van na de vijftiende eeuw in de stadsarchieven.

4.2 Deelcorpus narratieve teksten (1575–2000)

Terwijl voor het deelcorpus ambtelijke teksten gebruik gemaakt moest worden van verschillende bronnencollecties is het deelcorpus met narratieve teksten voornamelijk afkomstig van de Digitale Bibliotheek voor de Nederlandse Letteren (zie www.dbnl.org). Die online bibliotheek bevat een schat aan gedigitaliseerde literaire teksten vanaf de vroegste bronnen in het Nederlands tot vandaag, aangevuld met secundaire literatuur en biografische informatie. Uit dat ruime aanbod heb ik enkel narratieve teksten verzameld waarvan bekend is dat de auteur een geboren en getogen Hollander is. Dat houdt in dat de oudste teksten van het deelcorpus narratieve teksten pas vanaf het einde van de zestiende eeuw dateren, aangezien de oudere narratieve teksten in de dbnl doorgaans anoniem overgeleverd zijn.

De precieze keuze van het type narratieve teksten hangt in sterke mate af van het aanbod in de Digitale Bibliotheek voor de Nederlandse Letteren. In de zeventiende eeuw zijn de narratieve teksten in de dbnl vooral beperkt tot religieuze, filosofische, geschiedkundige en politieke traktaten. Meer literaire werken zoals kluchten of tragedies zijn nog in hoofdzaak berijmd en komen dus niet in aanmerking voor het deelcorpus narratieve teksten. Om de relatieve schaarste van narratieve werken in de zeventiende eeuw op te vangen heb ik ervoor geopteerd ook de inleidingen bij berijmde werken te excerperen. Die inleidingen van de hand van de auteur of de uitgever zijn doorgaans slechts korte narratieve teksten, maar ze dragen bij tot een verdere diversificatie van de narratieve teksten in de zeventiende eeuw. Met de verdere ontwikkeling van de Nederlandse literatuur worden meer tekstgenres geschikt voor het deelcorpus narratieve teksten. In de latere tijdsdoorsneden zijn naast traktaten en inleidingen ook meer literaire genres als de schelmenroman, het reisverslag, de historische roman, de novelle en het kortverhaal opgenomen in het deelcorpus.

Voor elke tijdsdoorsnede van vijfentwintig jaar heb ik ernaar gestreefd om een vijftal narratieve teksten van verschillende auteurs te verzamelen van elk tweeduizend vijfhonderd woorden lang. In tegenstelling tot de meeste ambtelijke teksten overschrijden de narratieve teksten doorgaans dat vooropgestelde streefcijfer ruimschoots. Met het oog op een evenwichtige vertegenwoordiging van de verschillende auteurs in het corpus, heb ik ervoor gekozen om toch vast te houden aan de invulling van elke cel door tweeduizend vijfhonderd woorden in plaats van al het beschikbare materiaal op te nemen. Enkel wanneer er binnen één tijdsdoorsnede een schaarste aan bronnen dreigde, is van dit principe afgeweken door meer materiaal voor enkele auteurs te verzamelen. In de appendix is een overzicht gegeven van de gekozen auteurs, de titel van de narratieve teksten, het tekstgenre van die teksten en ten slotte de grootte van de fragmenten die in het deelcorpus geëxcerpeerd zijn.

5. Uniforme vormgeving van het compilatiecorpus

Nu de precieze invulling van het compilatiecorpus grondig uit de doeken is gedaan, moet nog de nodige aandacht besteed worden aan de manier waarop de verzamelde teksten tot een uniform corpus omgevormd zijn. Ik zal achtereenvolgens bespreken hoe de uiteenlopende opmaak op het vlak van taalkundige annotaties en beschikbare metatalige informatie is geüniformeerd in de verzamelde steekproef.

Om te beginnen zijn de verzamelde teksten niet op een consistente manier van taalkundige annotaties voorzien in de verschillende bronnen. Zo bevatten de ambtelijke teksten uit het Corpus Gysseling en het Corpus Van Reenen-Mulder woordsoortinformatie en lemmatisering terwijl de gescande ambtelijke teksten en narratieve teksten van de dbnl helemaal niet taalkundig verrijkt zijn. Oorspronkelijk was het de bedoeling om de taalkundige annotatie van het Corpus Gysseling en het Corpus Van Reenen-Mulder te behouden in het nieuw samengestelde compilatiecorpus. Dat betekent dat er gestreefd zou moeten worden naar een uniforme opmaak van de taalkundige informatie in de tekstbestanden voor een optimale doorzoekbaarheid in het hele compilatiecorpus. Het maken van uniform vormgegeven annotaties bleek al heel gauw een erg delicate klus. Zo zijn de codes voor de taalkundige verrijking in beide bronbestanden op een afwijkende manier aan de woorden gehecht (in het Corpus Gysseling door zogenaamde vishaken rond het woord en in het Corpus Van Reenen-Mulder door een liggend streepje achter het woord of in de meest recente bestanden in de complexe XML-opmaak). Daarnaast wijken de gebruikte taalkundige codes in beide bronbestanden soms op minieme punten van elkaar af. Beide verschillen in opmaak van de taalkundige annotaties zijn niet onoverkomelijk, maar het lijkt wenselijker dat de samenstellers van beide corpora samen een standaard proberen te bereiken, eventueel in het kader van een overkoepelend corpusproject dat beide corpora tot één verzameling ambtelijke teksten samenbrengt en aanvult met jongere ambtelijke documenten. Bo-

vendien zou het behouden van de taalkundige annotaties voor de dertiende- en veertiende-eeuwse ambtelijke teksten voor de bijkomende complicatie gezorgd hebben dat het vroegste deel van het compilatiecorpus niet hetzelfde formaat had als de rest van het corpus (dat het leeuwendeel van het tekstmateriaal uitmaakt). Ik heb dus uiteindelijk het vroegste deel van het compilatiecorpus van zijn annotaties ontdaan in overeenstemming met de latere niet-verrijkte delen van het corpus. Het verwijderen van het kluwen van codes en haakjes uit het corpus heeft ook als voordeel dat de leesbaarheid van de teksten aanzienlijk stijgt.

Niet alleen op het vlak van taalkundige annotaties, maar ook wat de beschikbare metatalige informatie betreft, wijken de teksten uit de verschillende bronnen aanzienlijk van elkaar af. Zo bevatten de digitale tekstbestanden afkomstig uit het Corpus Gysseling en het Corpus Van Reenen-Mulder bijzonder veel details over het getranscribeerde manuscript zelf (het gebruik van speciale tekens of afkortingen in de tekst, het voorkomen van onleesbare passages, het begin van een nieuwe regel of bladzijde, afwijkingen in het handschrift die wijzen op verschillende kopiisten). In de gescande oorkonden echter vinden we slechts hier en daar een sporadische voetnoot met wat beknopte informatie over een onleesbare passage als bijvoorbeeld het gevolg van een watervlek. In de Digitale Bibliotheek voor de Nederlandse Letteren is ten slotte veel zorg gehecht aan het adequaat weergeven van titels, paragrafen en paginanummers uit gedrukte publicaties.

Uit dat bijzonder heterogene aanbod metatalige informatie heb ik de gegevens gedestilleerd die beschikbaar waren voor alle teksten in het compilatiecorpus. Om te beginnen heb ik informatie over de herkomst van een tekst en het jaar waarin die tekst geschreven is uit de tekstbestanden gehaald en op een uniforme manier verwerkt in de bestandsnaam van elke tekst. Die documentnaam bestaat achtereenvolgens uit de stad waar de tekst neergeschreven is, het jaartal van compilatie en een volgnummer die teksten van elkaar onderscheidt die op dezelfde tijd en plaats geschreven zijn (bv. amsterdam_1349_1.txt). Daarnaast zijn woorden of passages die door de uitgever van een teksteditie op één of andere manier als onleesbaar gemarkeerd zijn systematisch tussen rechte haakjes geplaatst (bv. in [orconde] desen brieue). De aanduiding van een nieuwe paragraaf is systematisch weergegeven door middel van een tabteken zodat de tekst netjes visueel gestructureerd wordt. Alle andere metatalige informatie is uit de tekstbestanden verwijderd zodat uiteindelijk een platte tekst overblijft. In die grote opschoonoperatie is ook alle interpunctie systematisch uit de tekstbestanden verwijderd. De oorspronkelijk motivatie hiervoor was dat leestekens in het Middelnederlands op een heel andere manier in de lopende tekst zijn aangebracht dan vandaag de dag het geval is. Bovendien vergemakkelijkt de afwezigheid van interpunctie het zoeken naar aaneengesloten reeksen van woorden zonder hierbij rekening te hoeven houden met doorbrekende leestekens. De keuze is achteraf gezien nogal ongelukkig, aangezien leestekens in een digitale tekst vrij eenvoudig zelf door de gebruiker ver-

wijderd kunnen worden door middel een *tokenizer* programma. Bovendien heeft het verwijderen van interpunctie ook de algemene leesbaarheid van de tekstbestanden tot op zekere hoogte aangetast, hoewel het gebruik van hoofdletters dat nadeel toch grotendeels weer goedmaakt.

Ten slotte zijn er in de platte tekst van ambtelijke teksten annotaties toegevoegd die het clichématige begin en einde van een oorkonde onderscheiden van het meer verhalende middendeel. Door de verschillende onderdelen van een oorkonde te markeren, kan een onderzoeker ervoor kiezen om de meest stereotiepe stukken uit van de ambtelijke tekstproductie uit het compilatiecorpus te weren (cf. Coussé 2008).

In de volgende figuur is de uiteindelijke opmaak van de tekstbestanden geïllustreerd met een veertiende-eeuwse oorkonde uit Amsterdam (amsterdam_1349_1.txt).

```
<begin>
Wi jacob gherits soen ende jacob ghisetgijns soen scepene in
Aemstelredamme orconden ende kennen
</begin>
<main>
dat peter hilmers soen gheliede voer ons dat [hi] scoudich is heynric
iacobs soen tien scellinghe hollants an comans ghelt iarlic ser rente
staende tot eweliken daghen op siin huus ende op erue daer hi nv
inwoent gheleghen an die steghe diemen ter [olen] gaet tusschen claes
hauics erue op die ene zide ende peter leyts erue op die ander zide
tebetalen dese rente alle iare op den meye dach
</main>
<end>
in [orconde] desen brieue bezeghelt mit onsen zeghelen Ghegheuen inden
jare ons heren mccc neghen ende viertich des woensdaghes na sente
iacobs dach
</end>
```

6. Troeven en beperkingen van het compilatiecorpus

Om de bespreking van de samenstelling en vormgeving van het compilatiecorpus af te sluiten, wil ik kort de troeven en — helaas ook onvermijdelijk — de beperkingen van het corpus evalueren voor historisch onderzoek van het Nederlands.

De grootste troef van het compilatiecorpus ligt mijns inziens in de zorgvuldige selectie van de teksten volgens een aantal welgedefinieerde criteria zoals de regionale herkomst en de datering van de teksten. Zo is er in het deelcorpus ambtelijke teksten systematisch naar gestreefd een evenwichtige steekproef samen te stellen van originele, lokale ambtelijke teksten uit de drie dialectstreken Vlaanderen, Brabant en Holland. Binnen elke tijdsdoorsnede van vijfentwintig jaar is getracht om steeds ongeveer tweeduizend vijfhonderd woorden aan tekstmateriaal te verzamelen in vijftien verschillende steden voor

de tijdspanne van 1250 tot 1800. Het deelcorpus narratieve teksten beoogt dan weer het geschreven taalgebruik in Holland evenwichtig te representeren vanaf het einde van de zestiende eeuw tot vandaag. Per tijdsdoorsnede van vijfentwintig jaar is hier gestreefd naar een gediversifieerde steekproef van tekstmateriaal van een vijftal verschillende schrijvers die in Holland geboren en getogen zijn. De evenwichtige samenstelling van het compilatiecorpus maakt het mogelijk om taalveranderingen ononderbroken in de tijd en ruimte te volgen vanaf het vroegste Nederlands tot vandaag. Om een zicht te geven op de precieze mogelijkheden van het compilatiecorpus zal ik in wat volgt kort enkele al bestaande casestudies aanhalen die met behulp van het compilatiecorpus zijn uitgevoerd.

Om te beginnen blijkt het compilatiecorpus bijzonder geschikt voor syntactisch onderzoek van constructies die een vrij hoge frequentie hebben in geschreven taalgebruik. Het compilatiecorpus is oorspronkelijk samengesteld in het kader van mijn doctoraatsonderzoek naar woordvolgordepatronen in tweeledige werkwoordclusters (Coussé 2008). Tweeledige werkwoordclusters hebben een vrij hoge incidentie in het compilatiecorpus, aangezien ze in het Nederlands in toenemende mate gebruikt worden voor het uitdrukken van de voltooide tijd (bv. *heeft geschreven, is gekomen*), het passief (bv. *wordt verkocht*) of de toekomende tijd (bv. *zal geschieden*). Concreet konden 4327 attestaties van tweeledige clusters geëxcerpeerd worden in het deelcorpus ambtelijke teksten en 1681 attestaties in het deelcorpus narratieve teksten. Op basis van die overvloedige attestaties was het mogelijk om de evolutie van woordvolgordepatronen in tweeledige werkwoordclusters ononderbroken te traceren vanaf het vroege Middelnederlands tot het hedendaagse Nederlands tot op vijfentwintig jaar nauwkeurig. Het compilatiecorpus bevatte bovendien ook voldoende attestaties om naast de werkwoordsvolgorde ook andere volgordepatronen in de zin te onderzoeken in tijdsdoorsneden van vijftig jaar (zie ook Coussé 2009) en daarnaast ook een zicht te geven op de grammaticalisatie van de tweeledige werkwoordclusters (zie ook Coussé 2006). Het valt te verwachten dat het compilatiecorpus even bruikbaar zal zijn voor syntactisch onderzoek van constructies met een vergelijkbare of zelfs hogere frequentie.

Daarnaast heeft het compilatiecorpus in zijn korte bestaan ook al zijn nut bewezen bij morfologisch onderzoek van historisch Nederlands. Zo legden De Vogelaer en Coussé (2008) op basis van het deelcorpus ambtelijke teksten de diachrone ontwikkeling van complexe persoonlijke meervoudspronomina (bv. *jullie, haarlieden*) naast de oorspronkelijke simplexvormen (bv. *wij, haar*) bloot voor het Middelnederlands. Met een relatief hoge frequentie van 6819 geattesteerde voornaamwoorden kon de verhouding tussen de complexe meervoudsvormen en de simplexvormen tot op vijftig jaar nauwkeurig onderzocht worden voor de drie dialectregio's Vlaanderen, Brabant en Holland afzonderlijk. Het ligt zeker binnen de mogelijkheden om de ontwikkeling van andere morfologische verschijnselen van een vergelijkbare frequentie te onderzoeken met behulp van het compilatiecorpus. Ik wil er evenwel aan herinneren dat het cor-

pus niet voorzien is van enige woordsoortinformatie of lemmatisering, iets wat morfologisch onderzoek aanzienlijk vergemakkelijkt. Daarnaast is ook enkel het oudste deel van het compilatiecorpus afkomstig van diplomatisch getranscribeerde bronnen, waardoor het corpus niet meteen geschikt is voor morfologisch onderzoek naar bijvoorbeeld casusverlies, waar een accurate transcriptie van de eindletters van woorden van het grootste belang is.

Zo ben ik geleidelijk tot de beperkingen van het compilatiecorpus gekomen. Aangezien het corpus samengesteld is met het oog op syntactisch onderzoek van de werkwoordgroep, is het daar uiteraard het beste op toegerust qua omvang en transcriptiedetail. Het spreekt voor zich dat het compilatiecorpus al gauw tegen zijn limieten aanloopt bij onderzoek naar minder frequente constructies, zoals de ditransitieve constructie. Bij dergelijk weinig voorkomende verschijnselen is het vooral van belang om een voldoende groot corpus teksten bij elkaar te verzamelen, eventueel ten koste van de strikte criteria die voor het compilatiecorpus zijn gehanteerd. Daarnaast is het compilatiecorpus niet bijzonder geschikt voor klassiek dialectologisch onderzoek dat de geografische distributie van taalverschijnselen tot op het niveau van de stad wil blootleggen. Bij de samenstelling van het ambtelijke deelcorpus bleek immers niet voor elke stad voldoende materiaal ter beschikking te zijn voor de hele tijdspanne tot het einde van de achttiende eeuw.

7. Aanbevelingen voor de toekomst

Een opsomming van de beperkingen van het compilatiecorpus hoeft niet noodzakelijk tot pessimisme te leiden, maar kan evengoed gelezen worden als een aanbeveling voor toekomstige corpusprojecten. Het spreekt voor zich dat de bestaande ambtelijke corpora van de dertiende en veertiende eeuw in de toekomst aangevuld zullen moeten worden met even geografisch gediversifieerd en kwalitatief getranscribeerd en geannoteerd materiaal uit latere periodes. Daarnaast zou het ook interessant zijn om de Digitale Bibliotheek voor de Nederlandse Letteren beter toegankelijk te maken voor taalkundig onderzoek. Op dit moment moet de historisch taalkundige het erg diverse tekstmateriaal op een nogal omslachtige manier doorzoeken op geschikte corpusteksten, zelf bibliografische gegevens van de auteurs zoeken en ten slotte de gekozen teksten één voor één downloaden. Ten slotte is het de ultieme wens van iedere historisch taalkundige om in de toekomst gebruik te kunnen maken van een gediversifieerd corpus met historische teksten van de vroegste bronnen tot vandaag dat uniform is vormgegeven en op verschillende taalkundige niveaus is geannoteerd. Alleen valt bij een dergelijk onderneming te verwachten dat de methodologische uitdagingen die in paragraaf twee geschetst zijn voor moeilijkheden zullen blijven zorgen: er is en blijft een discrepantie in het aanbod geschreven taal in de geschiedenis van het Nederlands en ook het concept van de geschreven taal zelf is veranderd door de tijd heen.

Bibliografie

Boonen, U. 2005. De begin- en slotformules in Utrechtse oorkonden uit de dertiende en veertiende eeuw: een vergelijking van Middelnederlandse en Latijnse formuleringen. *Neerlandistiek.nl* 05(06): 1-55.

Burgers, J. 1995. Over de lokalisering van Middelnederlandse ambtelijke bescheiden. *Taal en Tongval* themanummer 8 *Historische dialectologie*: 139-164.

Coussé, E. 2006. De historische wortels van volgordevariatie in het *hebben*-perfectum. *Taal en Tongval* 58: 250-277.

Coussé, E. 2007. Digitale bronnen voor taalhistorisch onderzoek van het Nederlands. *Nederlandse Taalkunde* 12: 275-279.

Coussé, E. 2008. *Motivaties voor volgordevariatie. Een diachrone studie van werkwoordsvolgorde in het Nederlands*. Proefschrift. Gent: Universiteit Gent.

Coussé, E. 2009. Focus, complexiteit en extrapositie. Over de veranderende woordvolgorde in het Nederlands. *Neerlandistiek.nl* 09(04): 1-31.

Coussé, E. 2010. Grammaticalisatie in de ambtelijke formule 'zoals voorzeid is'. *Tijdschrift voor Nederlandse Taal- en Letterkunde* 126: 20-33.

De Vogelaer, G. en E. Coussé. 2008. De kracht van disambiguering: nieuwe meervoudspronomina van het Middelnederlands tot nu. *Taal en Tongval* themanummer 21 *Dialectgeografie en interne factoren*: 13-35.

Gysseling, M. 1977. *Corpus van Middelnederlandse teksten (tot en met het jaar 1300)*. 's-Gravenhage: Nijhoff.

Hendriks, J.B. 1998. *Immigration and Linguistic Change. A Socio-cultural Linguistic Study of the Effect of German and Southern Dutch Immigration on the Development of the Northern Dutch Vernacular in the 16th/17th-Century Holland*. Proefschrift. Madison: University of Wisconsin.

Marynissen, A. 1999. ... allen dengenen die dese letteren sien selen / selen sien ende horen lezen ... Over volgordevariatie in de werkwoordelijke eindgroep in de Middelnederlandse bijzin. *Taal en Tongval* themanummer 12 *De verschriftelijking van het Nederlands*: 136-158.

Pijnenburg, W.J.J. en T.H. Schoonheim. 1998. De geschiedenis van een project. Schoonheim, T.H. en Th.P.F. Wortel (Reds.). 1998. *Een samenleving verwoord. Het Vroegmiddelnederlands Woordenboek*: 9-27. Den Haag: Sdu Uitgevers.

Van der Horst, J.M. 1997. Over en naar aanleiding van Zuid-Nederlandse doorbrekingen. Van Santen, A. en M.J. Van der Wal (Reds.). 1997. *Taal in tijd en ruimte*: 299-307. Leiden: SNL.

Van der Sijs, N. 2004. *Taal als mensenwerk, de geschiedenis van het ontstaan van het ABN*. Den Haag: Sdu Uitgevers.

Van der Wal, M.J. 1992. *Geschiedenis van het Nederlands*. Utrecht: Het Spectrum.

Van der Wal, M.J. 1995. *De moedertaal centraal. Standaardisatie-aspecten in de Nederlanden omstreeks 1650*. Den Haag: Sdu Uitgevers.

Van Loon, J. 2002. Op zoek naar de ideale tekst. Vanhoutte, E. (Red.). 2002. *Talig erfgoed. De zuidelijke Nederlanden in de 14de eeuw*: 73-90. Gent: KANTL.

Van Reenen, P.Th. en M. Mulder. 1993. Een gegevensbank van 14de-eeuwse Middelnederlandse dialecten op de computer. *Lexikos* 3: 259-281.

Appendix: Invulling van het deelcorpus narratieve teksten (n = 214 338 woorden)

Periode	Auteur	Jaar	Titel	Genre	Tot.
1572–99	Jacobsz., Wouter	1572	*Dagboek Broeder Wouter Jacobsz.*	dagboek	4334
	Coornhert, Dirk	1585	*Zedenkunst dat is wellevenskunste*	traktaat	5652
	Van Hout, Jan	1596	*Loterijspel*	voorwoord	2261
1600–24	De Groot, Hugo	1613	*Der heeren Staten van Hollandt ende West-Vrieslandt godts-diensticheyt*	traktaat	5229
	Orlers, Jan J.	1614	*Beschrijvinge der stad Leyden*	traktaat	5607
	Bredero, Gerbrand A.	1617	*Den Spaanschen Brabander*	voorwoord	2300
	Van Hogendorp, Gijsbrecht	1617	*Truer-spel van de moordt, begaen aen Wilhelm, prince van Oraengien*	voorwoord	1125
	Vander Plasse, Cornelis L.	1619	*De klucht van den molenaar (van Bredero)*	voorwoord	1084
	Bredero, Gerbrand A.	1622	*Groot lied-boeck*	voorwoord	1378
	Van de Venne, Adriaen	1622	*Tafereel van sinne-mal*	voorwoord	248
	Vander Plasse, Cornelis L.	1622	*Groot lied-boeck (van Bredero)*	voorwoord	285
1625–49	Vos, Jan	1641	*Aran of Titus*	voorwoord	1308
	Hooft, Pieter C.	1642	*Nederlandsche Historien. Het Leids beleg en ontzet, 1574*	kroniek	7444
	Coster, Samuel	1648	*De ses eerste vertoningen op de eeuwige vrede*	voorwoord	774
	Six, Jan	1648	*Medea*	voorwoord	679
1650–74	Huygens, Constantijn	1667	*Zee-straet*	traktaat	3298
	Vos, Jan	1667	*Medea*	voorwoord	3604
	Meyer, Lodewijk	1668	*Verloofde koninksbruidt*	voorwoord	3738
1675–99	Brandt, Geeraardt	1682	*Het leven van Joost van den Vondel*	biografie	6046
	Heinsius, Nicolaas	1695	*Den vermakelyken avanturier*	roman	4004
	Rotgans, Lukas	1698	*Wilhelm de Derde*	voorwoord	743
1700–24	Van Hoogstraten, David	1700	*Aenmerkingen over de geslachten der zelfstandige naemwoorden*	voorwoord	3124
	Sewel, Willem	1708	*Nederduytsche spraakkonst*	voorwoord	1568
	Alewyn, Abraham	1714	*Beslikte Swaantje en drooge Fobert*	voorwoord	395

	Langendijk, Pieter	1714	*Het wederzyds huwelyksbedrog*	voorwoord	599
	Alewyn, Abraham	1719	*De Puiterveense helleveeg*	voorwoord	512
	Bidloo, Lambert	1720	*Panpoëticon Batavum*	voorwoord	787
	Ten Kate, Lambert	1723	*Aenleiding tot het verhevene deel der Nederduitsche sprake*	voorwoord	6338
1725–49	Hoogvliet, Arnold	1728	*Abraham, de aartsvader*	voorwoord	2015
	Poot, Hubert	1728	*Gedichten*	voorwoord	2264
	Huydecooper, Balthazar	1730	*Proeve van taal- en dichtkunde*	voorwoord	2174
1750–74	Kersteman, Franciscus L.	1756	*Zeldzaame levens-gevallen van J.C. Wyerman*	biografie	9006
	Van Winter, Nicolaas	1769	*De jaargetyden*	voorwoord	233
	Van Alphen, Daniel	1772	*Levenbericht van Jan Wagenaar*	grafrede	4167
1775–99	Emmery, Willem	1782	*Onderwijs voor kinderen*	voorwoord	3855
	Corver, Marten	1786	*Tooneel-aantekeningen*	traktaat	1257
	Kinker, Johannes	1788	*De post van den Helicon*	spectator	2129
	Paape, Gerrit	1789	*Het land der willekeurigen*	reisverhaal	2570
	Fokke, Arend S.	1792	*De moderne Helicon, een droom*	satire	2140
1800–24	Bilderdijk, Willem	1807	*De ziekte der geleerden*	voorwoord	2742
	Klijn, Hendrik	1814	*Jan Frederik Helmers, in eene redevoering uitgesproken*	grafrede	6129
	Da Costa, Isaac	1823	*Bezwaren tegen den geest der eeuw*	pamflet	3568
	Loosjes, Adriaan	1823	*Het leven van Mauritz Lijnslager*	roman	3000
1825–49	Drost, Aarnout	1832	*Hemingard van de Eikenterpen*	roman	4060
	Geel, Jacob	1835	*Gesprek op den Drachenfels*	traktaat	2486
	Bosboom-Toussaint, Anna	1840	*Het huis Lauernesse*	roman	1671
	Kneppelhout, Johannes	1841	*Studenten-Typen*	schets	2741
	Van Koetsveld, Cornelis E.	1843	*Schetsen uit de pastorij te Mastland*	schets	2464
1850–74	Multatuli	1860	*Max Havelaar*	roman	2009
	De Génestet, Petrus A.	1861	*Over kinderpoëzy*	traktaat	1633
	Wolbers, J.	1861	*Geschiedenis van Suriname*	traktaat	1355
	Van Schaick, Cornelis	1866	*De Manja. Familie-tafereel uit het Surinaamsche volksleven*	schets	2946
	Beets, Nicolaas	1867	*Over kinderboeken. Gesprek met Crito*	traktaat	1976

1875–99	Busken Huet, Coenraad	1879	*Het land van Rubens*	reisverhaal	3024
	Emants, Marcellus	1879	*Een drietal novellen*	novelle	2675
	Vosmaer, Carel	1880	*Amazone*	roman	4052
	Van Eeden, Frederik	1885	*De kleine Johannes*	novelle	3627
	Netscher, Frans	1886	*Studie's naar het naakt model*	schets	4071
1900–24	Couperus, Louis	1901	*De boeken der kleine zielen*	roman	2007
	Van Booven, Henri	1904	*Tropenwee*	reisverhaal	3037
	Heijermans, Herman	1908	*Een wereldstad. Berlijnsche impressies en schetsen*	schets	2947
	Van der Leeuw, Aart	1908	*Sint-Veit*	kortverhaal	4048
	Van Looy, Jacobus	1917	*Jaapje*	roman	1559
1925–49	Roland Holst, Adriaan	1935	*Nederland. Oorlogstuig*	kortverhaal	1650
	Timmerman, Aegidius W.	1938	*Tim's herinneringen*	memoires	3236
	Terborgh, F.C.	1940	*De condottiere*	kortverhaal	3127
	Reve, Gerhard	1949	*De avonden*	roman	5215
	Roland Holst, Adriaan	1949	*Borrelpraat*	kortverhaal	802
1950–74	Wolkers, Jan	1963	*Wespen*	kortverhaal	3776
	Presser, Jacob	1965	*Ondergang. De vervolging en verdelging van het Nederlandse jodendom*	traktaat	2732
	Vinkenoog, Simon	1965	*Liefde. Zeventig dagen op ooghoogte*	dagboek	1322
	Morriën, Adriaan	1968	*Cryptogram*	anekdote	2396
	Arends, Jan	1974	*Ik had een strohoed en een wandelstok*	kortverhaal	1345
1975–99	Mulisch, Harry	1975	*Twee vrouwen*	roman	1800
	Hermans, Willem F.	1976	*De raadselachtige Multatuli*	biografie	2345
	Biesheuvel, Jacob M.A.	1983	*Reis door mijn kamer*	novelle	1858
	Brakman, Willem	1983	*Een wak in het kroos*	essay	4066
	Nooteboom, Cees	1991	*Het volgende verhaal*	roman	2567
Totaal					**214 338**