Het Gesproken Corpus van de zuidelijk-Nederlandse Dialecten

Anne BREITBARTH, Melissa FARASYN, Anne-Sophie GHYSELEN en Jacques VAN KEYMEULEN[1]

Abstract

In this paper, we report on the construction of a linguistically annotated pilot corpus of the southern Dutch dialects, based on existing tape recordings from the 1960s and 1970s. The corpus provides audio aligned transcriptions in two layers, one closer to the dialect and one closer to Standard Dutch, the latter of which is part-of-speech and syntactically tagged. The corpus is intended to facilitate large-scale research into the syntactic peculiarities of the southern Dutch dialects, which could not be researched systematically on a large scale in an easily reproducible way yet. Two short case studies concerning such peculiarities, i.e. V2 violations and the retention of the old preverbal negation particle, are presented in this paper to support the need for the corpus.

❖ ❖ ❖

1. INLEIDING

De zuidelijk-Nederlandse dialecten[2] kennen heel wat bijzonderheden die in de Nederlandse standaardtaal of in andere Germaanse talen niet of nauwelijks voorkomen. Enkele syntactische voorbeelden zijn het behoud van het oude preverbale negatiepartikel (o.m. Haegeman 1995, Zeijlstra 2004, Neuckermans 2008), dat in het dialect vaak ook nieuwe functies heeft verworven (cf. Neuckermans 2008; Breitbarth & Haegeman 2014, 2015), schendingen van de V2-woordvolgorde (bvb. Vanacker 1967; Lybaert et al. 2019; Haegeman & Greco 2018), subjectverdubbeling na complementeerders (Overdiep 1937, De Vogelaer 2008, Van Craenenbroeck 2010), congruentie met een complementeerder (Zwart 1993, De Vogelaer 2008, Van Koppen 2017), de 'vervoeging' van *ja* en *nee* (Overdiep 1937, Barbiers et al. 2008, Van Craenenbroeck & van Koppen 2008, Haegeman & Weir 2015, 2016), deverbale discours-

[1] Universiteit Gent. Het hier voorgestelde onderzoek werd (en wordt) gefinancierd door het FWO (navorserskrediet 1.5.310.18N aan A. Breitbarth en postdoctoraal mandaat FWO 12P7919N aan M. Farasyn). Naast het FWO bedanken wij Variaties vzw (http://www.variaties.be), vooral Lien Hellebaut en Veronique De Tier, voor hun ondersteuning, en onze transcribenten en vrijwilligers voor hun waardevolle medewerking.

[2] Onder de zuidelijk-Nederlandse dialecten verstaan we in dit artikel en in het corpus alle dialecten gesproken in de Nederlandstalige provincies in België (i.e. West-Vlaanderen, Oost-Vlaanderen, Antwerpen, Vlaams-Brabant en Limburg), het Zeeuws-Vlaams in Zeeland (Nederland) en het Frans-Vlaams in Noord-Frankrijk.

partikels[3] (Haegeman & Hill 2013, Haegeman 2014) en nog heel wat andere taalkundige fenomenen. De studie van die traditionele dialecten is taalkundig interessant aangezien de dialecten niet gestandaardiseerd zijn en dus als 'taal in het wild' gezien kunnen worden. Veel van de unieke eigenschappen van de zuidelijk-Nederlandse dialecten komen enkel voor in heel specifieke discourscontexten die in een geconstrueerde experimentele setting moeilijk geëliciteerd kunnen worden. Bestaande dialectverzamelingen voor syntactisch onderzoek zijn vaak gebaseerd op geëliciteerde gegevens (m.b.v. vragenlijsten). Hoewel elicitatie als grote voordeel heeft dat beïnvloedende factoren/contexten systematischer gecontroleerd kunnen worden (Cornips & Poletto 2005), volstaan vragenlijsten niet altijd om bepaalde fenomenen volledig in kaart te brengen of zelfs op te merken. Dat blijkt ook uit de notities van de veldwerkers van de SAND (Barbiers et al. 2005), waarin opmerkingen zoals de volgende terug te vinden zijn:[4]

(1) H116p Torhout:
veldwerker [v=359]: Met zo n weer je kun nie veel doen.
informant3 [a=n]: Met zo n weer kun je nie veel doen.
Opmerking van de veldwerker: de drie informanten keuren deze zin af; *nochtans komen er nogal wat inversieloze zinnen voor in de spontane spraak.*

Ook het gebruik van introspectieve gegevens is maar een gedeeltelijke oplossing, vooral gezien het gevorderde dialectverlies in Vlaanderen (Vandekerckhove 2009; Ghyselen & Van Keymeulen 2014). Een systematische analyse van spontaan gesproken taalgebruik (hier dialect) kan hier een waardevolle aanvulling betekenen.

Het Gesproken Corpus van de zuidelijk-Nederlandse Dialecten (GCND) is een corpus in opbouw, dat bestaat uit nieuwe transcripties van een voorlopig strategische selectie van 84 van in totaal 783 opnames met spontaan gesproken dialect. De opnames bevatten ongeveer 700 uur aan vrije gesprekken in de lokale dialecten van sprekers in 550 verschillende locaties. De opnames werden in de jaren 60 en 70 gemaakt door dialectologen van de Universiteit Gent onder de supervisie van prof. Willem Pée en prof. Valère Vanacker (cf. Vanacker & De Schutter 1967). Er werden opnames gemaakt van spontaan gesproken

[3] Wij maken hier een verschil tussen *discourspartikels*, die een instelling van de spreker ten opzichte van de uiting uitdrukken (Haegeman & Hill 2013), en partikels die de inhoud van het gezegde zelf modificeren. De laatste worden in de literatuur vooral over het Duits, waar die bijzonder vaak en gevarieerd voorkomen, *Modalpartikeln* of *Abtönungspartikeln* genoemd (zie o.m. Weydt 1979, Thurmair 1989, Coniglio 2011). Zoals uit de literatuur op te maken valt, verschillen discourspartikels en modaalpartikels in syntactisch gedrag en semantisch bereik. De deverbale partikels in de zuidelijk-Nederlandse dialecten zijn vooral discourspartikels.

[4] De notities bij de opgevraagde zinnen zijn te vinden op http://www.meertens.knaw.nl/sand/zoeken/

dialecten in hun natuurlijke vorm, gesproken door 'echte' dialectsprekers, i.e. weinig mobiele sprekers van een zekere leeftijd met een lage geletterdheid die in de plaats van de opname opgroeiden (cf. de NORM-spreker, Chambers & Trudgill 1980). Doordat de sprekers rond 1900 geboren zijn (de oudste in 1871), representeert het corpus bovendien een historisch taalstadium: de opnames leveren een inkijk in de traditionele dialecten die gesproken werden in het zuidelijk-Nederlandse taalgebied in de eerste helft van de twintigste eeuw. Het doel van de opnames was oorspronkelijk de dialectische, en dan vooral syntactische kenmerken van de zuidelijk-Nederlandse dialecten beter te kunnen bestuderen (Vanacker & De Schutter 1967). De opnames werden bij de sprekers ter plaatse vastgelegd op reel-to-reel tape. In 2014 werden de banden gedigitaliseerd en online beschikbaar gemaakt,[5] maar ze zijn slechts gedeeltelijk doorzoekbaar op basis van trefwoorden die aan korte inhouden van de banden zijn toegekend.

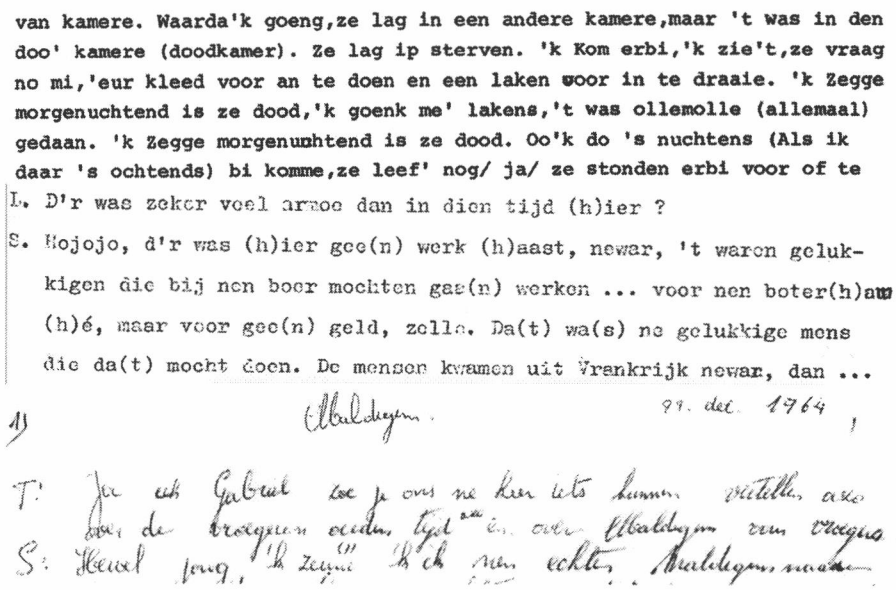
Wij brengen in dit artikel in de eerste plaats verslag uit over de transcriptie en verdere taalkundige verrijking van een aantal van de gedigitaliseerde banden, meer bepaald over de ontwikkeling van een tweelagig transcriptieprotocol en de taalkundige verrijking van de transcripties met lemmatisering, tags voor woordsoorten (part-of-speech) en syntactische informatie (parsing). In vergelijking met andere dialectcorpora zoals de SAND en DynaSAND (Barbiers et al. 2005), SyHD (Fleischer et al. 2015) of SADS (Glaser & Bart 2015) is het GCND uniek doordat het volledig op spontaan gesproken taal gebaseerd is, en niet op bevragingen. In het geval van het Frans-Vlaams bevat het zelfs de laatste getuigenissen van een zo goed als verdwenen taal (Ryckeboer 2013). We bespreken verder twee casestudy's die de noodzaak en het nut van een taalkundig verrijkt corpus van de zuidelijk-Nederlandse dialecten toelichten.

2. De ontwikkeling van een transcriptieprotocol voor de zuidelijk-Nederlandse dialecten

Om een (dialect)opname toegankelijk te maken voor syntactisch, ander taalkundig, maar ook bijvoorbeeld historisch of volkskundig onderzoek, is het noodzakelijk een doorzoekbare tekst te hebben. Toen de dialectopnames in de jaren 60 en 70 gemaakt werden, werden een aantal opnames meteen getranscribeerd (318 in totaal), doorgaans door studenten in de context van een licentiaatsverhandeling. Toch zijn de beschikbare transcripties niet optimaal voor verder onderzoek. Enerzijds zijn de grote kwaliteitsverschillen tussen de transcripties problematisch: een aantal van de tot op vandaag bewaarde transcrip-

[5] www.dialectloket.be/geluid/stemmen-uit-het-verleden/

ties werden getypt en zijn duidelijk leesbaar; bij andere is de inkt vervaagd of staan er heel wat opmerkingen en correcties tussen de regels en in de kantlijn. Elke transcriptie werd immers eerst uitgeschreven, daarna nagekeken en ten slotte uitgetikt. Sommige transcripties hebben die eindfase echter niet gehaald of gingen in de loop der jaren verloren, waardoor een groot aantal van de bewaarde transcripties met de hand geschreven is. Het OCR'en van het grootste deel van de collectie is daardoor vrij arbeidsintensief; heel wat manuele correctie is nodig. Een nog groter probleem is dat de transcripties niet eenvormig zijn. Er was bij het maken van de oude transcripties enkel een zeer summier transcriptieprotocol voorhanden, waardoor er duidelijke variatie is in de manier waarop dialectkenmerken orthografisch weergegeven worden. Afbeelding 1 geeft een indruk van de verschillen tussen de transcripties; bij Torhout is ervoor gekozen om het vocalisme zo getrouw mogelijk af te beelden (met af en toe 'vertalingen' naar de standaardtaal), terwijl bij Wichelen het vocalisme net gestandaardiseerd is, en niet-uitgesproken medeklinkers (tussen haakjes) gereconstrueerd zijn, om de transcriptie ook voor niet-dialectkundigen toegankelijk te maken. De transcriptie voor de opname van Maldegem is met een pen geschreven.


van kamere. Waarda'k goeng,ze lag in een andere kamere,maar 't was in den doo' kamere (doodkamer). Ze lag ip sterven. 'k Kom erbi,'k zie't,ze vraag no mi,'eur kleed voor an te doen en een laken voor in te draaie. 'k Zegge morgenuchtend is ze dood,'k goenk me' lakens,'t was ollemolle (allemaal) gedaan. 'k Zegge morgenuhtend is ze dood. Oo'k do 's nuchtens (Als ik daar 's ochtends) bi komme,ze leef' nog/ ja/ ze stonden erbi voor of te

I. D'r was zeker veel armoe dan in dien tijd (h)ier ?

S. Mojojo, d'r was (h)ier gee(n) werk (h)aast, newar, 't waren gelukkigen die bij nen boer mochten gae(n) werken ... voor nen boter(h)am (h)é, maar voor gee(n) geld, zelle. Da(t) wa(s) ne gelukkige mens die da(t) mocht doen. De mensen kwamen uit Vrankrijk newar, dan ...

Afbeelding 1: *Uittreksels uit de oorspronkelijke transcripties van Torhout (boven), Wichelen (midden), en Maldegem (onder)*

Afbeelding 1 maakt de problemen met de bruikbaarheid van de oude (en onvolledige) transcripties duidelijk. Wij hebben er daarom voor gekozen om zelf een transcriptieprotocol te ontwikkelen, gebaseerd op dat van de DIT-werk-

groep (Dialect in Transcriptie)[6] binnen het SAND-project. Ons aangepaste protocol moest toelaten een breed gamma aan dialecten, van de westelijkste Frans-Vlaamse tot de oostelijkste Limburgse dialecten, zo uniform mogelijk orthografisch weer te geven; er werd ook steeds in het achterhoofd gehouden dat de transcripties toegankelijk en doorzoekbaar zijn moesten zijn voor zowel linguïsten (geïnteresseerd in syntactische, fonologische, fonetische, lexicale en/of morfologische dialectfenomenen), als niet-taalkundigen geïnteresseerd in de inhoud van de opnames. Daarom is ervoor geopteerd met twee transcriptielagen te werken: één dichter bij het dialect en een andere dichter bij het Standaardnederlands.

In de eerste laag (dichtst bij het dialect) worden niet-standaardtalige woordenschat, morfologie en syntaxis bewaard en neergeschreven volgens de Nederlandse spellingsregels.[7] Clitische elementen – zoals *k#weten* ('ik weet') – worden als clusters getranscribeerd, al worden de individuele delen van de cluster van elkaar gescheiden met een # (om de alignering met de tweede laag te faciliteren). In de eerste laag vernederlandsen we enkel de dialectische fonologie. We opteren voor een dergelijke vernederlandsing omdat voor een uniforme en nauwkeurige weergave van de klankvariatie IPA (of een ander fonetisch alfabet) noodzakelijk is, en dat de transcriptie te gecompliceerd, tijdrovend en duur zou maken. De alignering met de audio (zie onder) zorgt er echter voor dat de klankvormen wel toegankelijk blijven. De enige fonetische/fonologische niet-standaardtalige kenmerken die gemarkeerd worden in de eerste laag, zijn deleties en inserties van consonanten in functiewoorden (we schrijven bijvoorbeeld *naa* in plaats van *naar*), aangezien die kenmerken eenduidig getranscribeerd kunnen worden zonder fonetische tekens.

In de tweede laag wordt niet enkel fonologie, maar ook morfologie 'vernederlandst'; *k#weten* wordt bijvoorbeeld *ik weet.* Het voorbeeld van *k#weten* toont ook hoe clitische elementen (die met een # gemarkeerd worden in de eerste laag), als aparte tokens worden weergegeven in laag 2. Niet-standaardtalige syntaxis en woordenschat bewaren we echter, aangezien er vaak discussie mogelijk is over hoe die lexicale en syntactische kenmerken 'vertaald' moeten worden naar het Standaardnederlands. De tweede laag is voor die elementen dus een vernederlandsing, geen vertaling. De tweede laag maakt het gebruikers zonder een uitgebreide kennis van het dialect mogelijk de data te doorzoeken.

6 Aan die werkgroep namen deel: Margreet van der Ham, Tamar Israël, Mathilde Jansen, Willy Jongenburger, Susanne van der Kleij en Gunther De Vogelaer. Guido Vanden Wyngaerd en Sjef Barbiers schreven de definitieve versie.

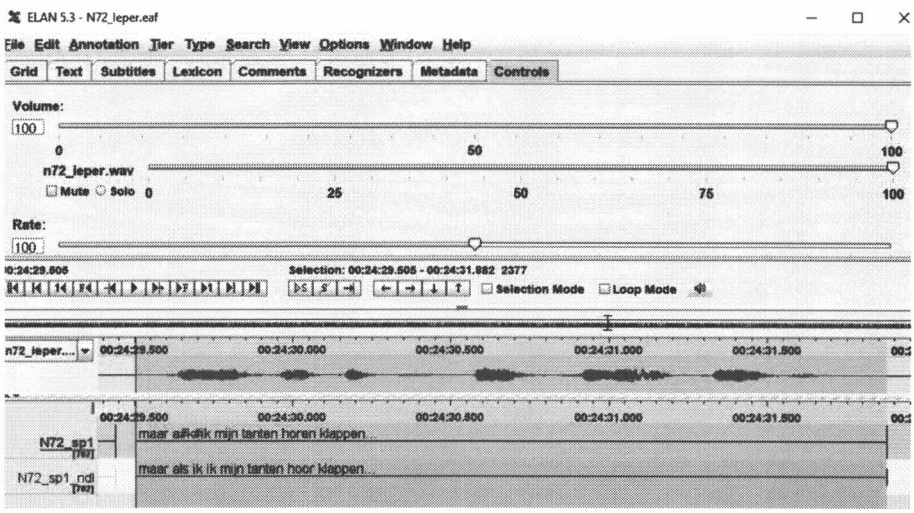
7 Bij twijfel over de spelling van een dialectwoord, wordt in de mate van het mogelijke een beroep gedaan op bestaande bronnen (zoals het Woordenboek van de Brabantse Dialecten, het Woordenboek van de Limburgse Dialecten, het Woordenboek van de Vlaamse Dialecten en het WNT).

Voorbeeld (2) illustreert de werkwijze met een passage uit de transcriptie van Ieper:

(2) laag 1: maar a#k#ik mijn tanten horen klappen… (N72 Ieper)
laag 2: maar als ik ik mijn tanten hoor klappen…
'maar als ik mijn tantes hoor spreken/praten…'

Zoals aan dit voorbeeld te zien is, is de klinker in *mijn* in de eerste laag volgens de standaardspelling getranscribeerd; er wordt niet gepoogd de dialectische uitspraak die in de opname te horen is – [mɪn] – orthografisch te noteren. De dialectische functiewoorden (*a* voor *als*), cliticclusters (*a#k#ik*) en inflectionele morfologie (*horen* voor *hoor*) worden daarentegen wel dialectisch weergegeven. In de tweede laag is de cliticcluster opgelost en zijn de functiewoorden (*a, k*) volgens de standaardconventies geschreven, terwijl de dialectische woordvolgorde behouden blijft (*als ik ik*). Het dialectlexeem *klappen* wordt eveneens niet vertaald, aangezien discussie mogelijk is over welk standaardlexeem dan precies gekozen zou moeten worden (*spreken*, *praten* of *vertellen*?).

In lijn met de CLARIN-filosofie hebben wij er voor het GCND voor gekozen zoveel mogelijk bestaande technologie te hergebruiken. De transcripties worden daarom gemaakt met de bestaande gratis transcriptiesoftware ELAN, waarmee het mogelijk is transcripties te maken die met het geluid gealigneerd zijn. Dat maakt het voor de onderzoeker mogelijk om gevonden syntactische structuren in het corpus ook meteen te beluisteren, en zo bijvoorbeeld de precieze fonetische vorm te achterhalen (zie boven). Afbeelding (2) toont de weergave van voorbeeld (2) in ELAN.



Afbeelding 2: ***Transcriptie van (2) in ELAN***

3. CROWD SOURCING/CITIZEN SCIENCE VOOR KWALITEITS-CONTROLE

De nieuwe transcripties worden aan de Universiteit Gent voor het grootste deel gemaakt door studenten, onder ander in het kader van bachelorproeven. Vaak is de dialect- en wereldkennis van de studenten echter onvoldoende om alle passages te kunnen doorgronden. Soms stellen ongekende dialectwoorden of -constructies de studenten voor problemen; op andere momenten vormt de gespreksmaterie – onderwerpen zoals de vinkenzetting of de vooroorlogse vlasteelt – een uitdaging. Problematische passages markeren studenten met de code '???'.[8] Om die passages toch van een transcriptie te kunnen voorzien, wordt een beroep gedaan op dialectcompetente vrijwilligers, die de accuratesse van de studententranscripties goed kunnen inschatten en door hun levenservaring – het gaat doorgaans om oudere medewerkers – beter met de gespreksmaterie vertrouwd zijn dan de student-transcribenten.[9] Zij zorgen voor aanvullingen bij onduidelijke passages, op papier of in geëxporteerde tekstbestanden uit ELAN. Verder gaan ze ook na of de student-transcribenten de inhoud correct begrepen en getranscribeerd hebben. De correcties en aanvullingen worden daarna geëvalueerd en indien nodig aangevuld in ELAN door een projectmedewerker die het protocol en de software grondig beheerst.

4. PART-OF-SPEECH-TAGGEN, LEMMATISEREN EN PARSEN MET FROG

Om de doorzoekbaarheid van de transcripties te optimaliseren, worden de tekstbestanden taalkundig verrijkt. Die taalkundige verrijking bestaat onder andere uit het (semi-)automatisch toekennen van woordsoorten aan alle tokens in de tweede laag (de sterkere vernederlandsing). Ook hier werd in het licht van de CLARIN-filosofie op zoek gegaan naar al bestaande oplossingen, wat ook de reden is voor het gebruik van de tweede transcriptielaag als input. We evalueerden de performantie van drie bestaande part-of-speech (POS)-taggers voor het Standaardnederlands op onze dialectdata, namelijk TreeTagger (Schmid 1999) met het Nederlandse parameterbestand, de LeTs-toolkit (Van de Kauter et al. 2013) en Frog (Hendrickx et al. 2016). Daartoe hebben we uit elke grote dialectzone en elke overgangszone daartussen één plaats gekozen, behalve uit de overgangszone tussen het Brabants en het Limburgs, aangezien er daar geen opname voorhanden is. Na het automatisch toekennen van de tags door de drie verschillende taggers werden er nadien per tagger 1000 tags per

[8] Deze code wordt enkel gebruikt voor passages die de studenten door een gebrekkige dialect- of wereldkennis niet kunnen doorgronden; passages die om akoestische redenen (slechte opnamekwaliteit, achtergrondlawaai, overlappende spraak) moeilijk te verstaan zijn, worden met de code 'xxx' gemarkeerd.

[9] Gezien de energie en tijd nodig om de ELAN-software en het transcriptieprotocol onder de knie te krijgen, is het moeilijker deze vrijwilligers in te zetten voor het eigenlijke transcriptiewerk.

plaats gecontroleerd. Er werden in totaal 9 plaatsen uitgekozen, waarbij Frog over het algemeen het beste presteerde met een gemiddelde accuratesse van 94.5%, oplopend tot 98.8% voor het dialect van Sint-Niklaas. Tabel 1 toont de Frog-resultaten in detail. TreeTagger en de LeTs-toolkit behaalden een gemiddelde accuratesse van respectievelijk 92.9% en 90.5%.

Tabel 1: Accuratesse van de Frog-tagger voor negen geselecteerde plaatsen uit het GCND

Plaats	dialect	correct	niet correct	% correct	totaal
Oudenburg (H24)	wvl. (kust)	952	48	95.2	1000
Maldegem (I154)	wvl.-ovl.	973	27	97.3	1000
Westdorpe (I166)	zvl.	988	12	98.8	1000
Sint-Niklaas (I175)	ovl.-brab.	946	54	94.6	1000
Gent (I241)	ovl.	929	71	92.9	1000
Ieper (N72)	wvl. (Westhoek)	926	74	92.6	1000
Hardifort (N94)	fvl.	938	62	93.8	1000
Sint-Joris-Weert (P130)	brab.	893	107	89.3	1000
Uikhoven (Q013)	limb.	959	41	95.9	1000
Totaal		8504	496	94.5	9000

Frog bestaat uit een set van *natural language processing* (NLP)-hulpmiddelen gebaseerd op het *memory-based learning* softwarepakket TiMBL (Daelemans et al. 2007), waarvan de meeste onderdelen ontwikkeld werden door de ILK Research Group van Tilburg University en het CliPS Research Centre van de Universiteit Antwerpen (Van den Bosch et al. 2007). Met Frog kunnen tekstbestanden onder andere getokeniseerd, gepart-of-speech-tagged en gelemmatiseerd worden. Bovendien is het mogelijk om syntactische tags te laten toekennen. De POS-tags die door Frog worden toegekend, zijn gebaseerd op de tagset van het Corpus Gesproken Nederlands (CGN; Oostdijk 2000). Er wordt bij het toekennen van de tag ook telkens een score berekend die aangeeft hoe zeker de classificeerder is van een bepaalde tag. De POS-tagger is getraind op een corpus van gesproken Nederlands van 10.975.324 tokens, waarvan er 90% deel uitmaakt van het CGN.

De grote accuratesse van Frog op de GCND-transcripties kan waarschijnlijk verklaard worden door het feit dat het CGN ook veel Belgisch-Nederlandse data bevat. De meeste incorrecte woordsoorttoekenningen in onze testset bleken te wijten aan het feit dat FROG getraind werd op (geïntendeerd) standaardtalige data. Toch kent de tagger ook een groot aantal correcte tags toe aan niet-standaardtalige woorden, zoals *bè* als dialectaal tussenwerpsel, *ne* als dialectaal onbepaald lidwoord en *lijk* als onderschikkend voegwoord. De

grootste groep van tokens die incorrect getagd worden, zijn verschillende andere soorten interjecties, zoals *awel*. In de transcriptie van Sint-Niklaas zijn bijvoorbeeld alle incorrect toegevoegde tags aan zulke interjecties te wijten. Verder wordt het ontkennend partikel *en* doorgaans als conjunctie getagd. Omdat de fouten vrij regelmatig zijn, kan veel geremedieerd worden aan de hand van scripts; manuele correcties kunnen tot een minimum beperkt worden.

Frog geeft naast een POS-tag ook automatisch andere informatie weer die nuttig is voor het pilootcorpus. Elk token wordt bijvoorbeeld gelemmatiseerd en aangevuld met morfologische informatie met behulp van MBLEM en MBMB, respectievelijk een memory-based lemmatiser en een memory-based morphological analyser die specifiek getraind zijn op Nederlandse, en bij uitbreiding ook op Engelse en Duitse data. De woordsoortinformatie is gebaseerd op de lexicale database CELEX, die informatie bevat over lexicografie, fonologie, morfologie, syntaxis en woordfrequentie (Baayen et al. 1995). Ook die NLP-hulpmiddelen presteren goed op de data uit het pilootcorpus. Zo wordt het dialectische *gewrocht* bijvoorbeeld correct als voltooid deelwoord van *werken* herkend. De gelemmatiseerde XML-verrijking zal het mogelijk maken om de niet-standaardtalige lexemen in de tweede laag van de transcripties, zoals *klappen* in voorbeeld (2), in de toekomst te linken aan lemma's uit gedigitaliseerde dialectwoordenboeken. De gebieden vertegenwoordigd in dit corpusproject worden behandeld in het (Elektronische) Woordenboek van de Vlaamse Dialecten ((e-)WVD, wvl., ovl., zvl., fvl.), het (Elektronische) Woordenboek van de Brabantse Dialecten ((e-)WBD, brab.) en het (Elektronische) Woordenboek van de Limburgse Dialecten ((e-)WLD, limb.), die momenteel in één database samengevoegd worden tot de (digitale) *Dictionary of the Southern Dutch Dialects* (DSDD, vgl. Van Keymeulen et al. 2018).

Een noodzakelijke stap tussen POS-tagging en parsing is *chunking*. De getranscribeerde data in het pilootproject worden daarvoor voorzien van *base phrase chunks*, labels die aangeven welke tokens samen een woordgroep vormen. Dat is opnieuw een onderdeel van de NLP-tools van Frog. In een aparte laag van het XML-document wordt aan elk token een Beginning-Inside-Outside-tag (BIO-tag) toegekend. Zo'n tag geeft aan waar een woordgroep begint, welke tokens er tot de woordgroep behoren en wat de functie van de woordgroep is. Zo worden de tokens *horen* en *klappen* uit voorbeeld (2) voorzien van de respectievelijke labels B-VP en I-VP: *horen* is het begin van een werkwoordgroep (B), *klappen* bevindt zich binnenin de woordgroep (I). Samen vormen ze één werkwoordgroep (*verb phrase*, VP).

In de laatste fase, de *parsing*-fase, specificeert Frog de grammaticale relaties tussen de verschillende woordgroepen die de chunker identificeerde. Frog gebruikt hiervoor de *Constraint-satisfaction inference-based dependency parser* (CSI-DP; Canisius et al. 2006), die afhankelijkheidsrelaties tussen twee to-

kens aangeeft, waarbij één token het hoofd (*head*) is en het andere token het afhankelijke element (*dependent*). Tabel 2 illustreert die afhankelijkheidsrelaties aan de hand van een voorbeeldzin uit de Ieperse opname (*Je zat nog in het groene kooltje*).

Tabel 2: *Chunking en parsing informatie in Frog (zin uit N72 Ieper)*

	token	chunk	token number in dependency graph	dependency relation
1	je	B-NP	2	su
2	zat	B-VP	0	ROOT
3	nog	B-ADVP	2	mod
4	in	B-PP	2	mod
5	het	B-NP	7	det
6	groene	I-NP	7	mod
7	kooltje	I-NP	4	obj1
8	.	O	7	punct

De laatste kolom bevat de afhankelijkheidsrelaties. Eén token is de wortel (ROOT), waarvan alle andere tokens direct of indirect afhangen, in dit voorbeeld het finiete werkwoord *zat*. Het onderwerp (su) is *je*; het naamwoord *kooltje* wordt als lijdend voorwerp (obj1) geannoteerd, dat ingebed zit in een voorzetselconstituent (*in het groene kooltje*) met als hoofd het voorzetsel *in* dat net als het onderwerp direct afhangt van de ROOT. De afhankelijkheidsgraad wordt uitgedrukt door de cijfers in de voorlaatste kolom; hoe kleiner het getal, hoe nauwer de relatie met de ROOT. In het voorbeeld in Tabel 2 zijn het onderwerp, het bijwoord *nog* en het hoofd *in* van de PP *in het groene kooltje* dus direct afhankelijk van het werkwoord *zat*, terwijl het naamwoord *kooltje* er een indirectere relatie mee heeft.

Informatie over grammaticale relaties kan helpen om het corpus op basis van syntactische structuren te doorzoeken zonder dat er op concrete tekst gezocht moet worden.

5. Twee gevalstudies naar typologische bijzonderheden van de zuidelijk-Nederlandse dialecten

In de zuidelijk-Nederlandse dialecten komen heel wat syntactische kenmerken voor die het Standaardnederlands of andere Germaanse talen niet kennen. Veel van die kenmerken vallen bovendien enkel op te merken indien ze in spontaan gesproken taal bestudeerd worden. Hieronder bespreken we twee van die kenmerken: schendingen van de V2-woordvolgorde en het negatiepartikel *en*.

5.1. Schendingen van de V2-woordvolgorde

Het Nederlands en alle andere Germaanse standaardtalen hebben, met uitzondering van het Engels, V2-woordvolgorde in een mededelende hoofdzin (3): het vervoegde werkwoord komt op de tweede zinsplaats te staan. Na een voorafgaande constituent, bijvoorbeeld na een bijwoordelijke bepaling, treedt er inversie op (4).

(3) Het spatte in alle richtingen.
(4) Als je erop sloeg met een pikhouweel, spatte het.

De situatie is anders in een aantal zuidelijk-Nederlandse dialecten. Vooral in het West-Vlaams komen ook zinnen met vooropgeplaatste adjuncten zonder inversie voor (Vanacker 1977, Haegeman & Greco 2016, Lybaert et al. 2019). In die zinnen, die de V2-conditie lijken te schenden, volgt het subject meteen op het vooropgeplaatste zinsdeel (5).

(5) a je derop sloeg met een pioche het spetterde
AN: 'als je erop sloeg met een pikhouweel, spatte het'

In de SAND is dat soort inversieloze zinnen echter maar beperkt geattesteerd. De reden daarvoor is dat vele types V2-schendingen maar in heel specifieke contexten gerealiseerd lijken te worden. Haegeman & Greco (2016, 2018) geven bijvoorbeeld aan dat het voorkomen van sommige inversieloze zinnen in het West-Vlaams sterk afhankelijk is van het voorgaande discours. Zo kunnen schendingen van de woordvolgorde onder andere een verrassingseffect aanduiden, waarbij de spreker een referentiekader introduceert ten opzichte waarvan de situatie in de hoofdzin ongebruikelijk of onverwacht lijkt (6).

(6) Oa-me tuskwamen, de voordeure stond open en de lucht was an.
AN: 'Toen we thuiskwamen stond de voordeur open en het licht was aan.'
(Haegeman & Greco 2016, voorbeeld 21)

Het GCND zal van groot belang zijn om de frequentie en specificiteiten van dat fenomeen bloot te leggen. Het onderzoek van Haegeman & Greco (2016) suggereert dat er regionale verschillen zijn in V3-eigenschappen in bepaalde discourscontexten tussen het West-Vlaams in België en het Frans-Vlaams (i.e. het West-Vlaams gesproken in het noordwesten van Frankrijk). Bovendien wijst de studie van Lybaert et al. (2019) erop dat V2-schendingen vaker voorkomen in het Frans-Vlaams dan in het West-Vlaams. Toch is de exacte geografische verspreiding van V2-schendingen met verschillende soorten constituenten tot nog toe niet in detail onderzocht. Studies die het voorkomen van

V2-schendingen op basis van frequentie hebben onderzocht, nemen, hoewel ze gebaseerd zijn op de bandencollectie, slechts een klein aantal plaatsen onder de loep, aangezien gestructureerde zoekacties in het corpus niet mogelijk zijn. Bovendien zijn de resultaten niet of slechts heel moeilijk reproduceerbaar.

5.2. Negatiepartikel en

Het partikel *en*, de oorspronkelijke markeerder van zinsontkenning, die in meerdere Vlaamse variëteiten als overblijfsel van de Jespersencyclus bewaard bleef (cf. Koelmans 1967; Haegeman 1995; Neuckermans 2008), is hierboven al meerdere keren vermeld, enerzijds omdat het door de vormelijke overeenkomst (maar uiteraard niet de syntactische distributie) met het nevenschikkend *en* problemen voor de automatische POS-tagging oplevert, anderzijds omdat het in voorbeeld (3) al voorkwam. Volgens Neuckermans (2008) verschijnt *en* in ontkennende zinnen vaker in bijzinnen dan in hoofdzinnen, normaal gezien meteen voor de persoonsvorm. Op 23 plaatsen in haar corpus (gebaseerd op een combinatie van het SAND-materiaal, de RND-zinnen[10] en een selectie van de bestaande transcripties van dialectbanden (zie boven)) zijn er ook voorbeelden van wat Neuckermans (2008) restrictief en expletief gebruik noemt (7)[11], en erg zeldzaam niet-negatief gebruik (8)[12], dat vooral tot Brabant (en een plaats in Oost-Vlaanderen) beperkt lijkt te zijn, en waar het ook voor een niet-finiet werkwoord kan verschijnen.

(7) Ten i moar een tjootn (N125 Wulvergem)
AN: 'Het en is maar een rare man.'
(Neuckermans 2008: 162)

(8) a. Ik dacht dage op café en zat (O250, Sint-Pieters-Leeuw)
AN: 'Ik dacht dat je op café zat.'
(Neuckermans 2008: 176)

b. IJ zal nog wel en komen (O022 Merelbeke)
AN: 'Hij zal nog wel komen'
(Neuckermans 2008: 175)

In de 45 tot nu toe getranscribeerde opnames in het GNCD zijn er nu al aanwijzingen voor een veel verregaander gebruik van niet-ontkennend *en*, al dan

[10] http://www.dialectzinnen.ugent.be

[11] Bij restrictief gebruik zoals in (7) wordt *en* in niet-ontkennende zinnen met een restrictief bijwoord zoals *maar* of *juist* gebruikt. Expletief is volgens Neuckermans het gebruik van *en* in zwak negatief-polaire contexten, zoals in vergelijkende en voorwaardelijke zinnen.

[12] D.w.z. zonder restrictief of negatief-polair element dat het gebruik van *en* zou motiveren zoals bij restrictief of expletief gebruik.

niet voor infiniete werkwoorden. Het verschijnt bijvoorbeeld in een niet-ontkennende bijzin van een ontkennende hoofdzin (9a), in een (niet-ontkennende) bijzin van een hoofdzin met een restrictief partikel (9b)[13], en als enige markeerder van zinsontkenning in een hoofdzin met een complement-vraagzin (9c), een constructie die vooral voor het Middelnederlands beschreven is (bijvoorbeeld Postma 2002). Bovendien is het ook in een breder verspreidingsgebied in niet-negatieve zinnen en voor niet-finiete werkwoorden geattesteerd (9d).

(9) a. [als het zo maar naar de winkel of ... of naar de patisserie en is of gelijk wat ...] je gaat niet binnen (O80 Waregem)
b. en ze danste maar het waren maar juist haar voeten [die van de grond en gingen] (N42 Pittem)
c. ik en weet [of dat nu nog veel meer gedaan werd] (O265 Ronse)
d. met zijn beste kleren aan ... je had dien een keer moeten en zien (N42 Pittem)

Het GCND biedt nieuwe mogelijkheden om het gebruik en de regionale verspreiding van het partikel *en* te onderzoeken. Het corpus bestaat immers uit spontaan gesproken materiaal en maakt zo onderzoek naar heel specifieke discourscontexten mogelijk. Op basis van de besproken voorbeelden kunnen we verwachten dat het GCND meer inzicht zal bieden in de gebruiksvoorwaarden van *en* dan momenteel mogelijk is op basis van de bestaande literatuur.

6. Besluit

In dit artikel hebben wij gerapporteerd over de constructie van het taalkundig geannoteerde Gesproken Corpus van de zuidelijk-Nederlandse Dialecten, een pilootcorpus dat grootschalig onderzoek naar de gesproken dialecten in Nederlandstalig België, Frans-Vlaanderen en Zeeland mogelijk zal maken. De verrijkte geluidsopnames zullen in de toekomst online beschikbaar gesteld worden in samenwerking met het Instituut voor de Nederlandse Taal. Uiteraard hopen we het aantal banden ook te kunnen uitbreiden via nieuwe projecten en eventueel zelfs om de opnamecollectie verder uit te breiden. De opbouw van het corpus met een output in XML-formaat zal onderzoekers uit andere disciplines verder toelaten om informatielagen aan het corpus toe te voegen, bijvoorbeeld voor *oral history*-onderzoek naar de inhoud van de banden.

[13] (9a) is trouwens ook een voorbeeld van een V>2 constructie na een adverbiale bijzin, zie 5.1.

BIBLIOGRAFIE

Barbiers, S., H. Bennis, G. De Vogelaer, M. Devos, & M. van Der Ham. 2005. *Syntactische atlas van de Nederlandse dialecten: Deel 1: Pronomina, Congruentie en Vooropplaatsing*. Amsterdam: Amsterdam University Press.

Barbiers, S., O. Koeneman & M. Lekakou. 2008. Syntactic doubling and the structure of chains. In C.B. Chang & H.J. Haynie (red.), *Proceedings of the 26th West Coast Conference on Formal Linguistics*, 77-86. Somerville, Massachusetts: Cascadilla Press.

Baayen, R. H., Piepenbrock, R. & Gulikers, L. 1995. CELEX2 LDC96L14. Web Download. Philadelphia: Linguistic Data Consortium.

Breitbarth, A. & L. Haegeman. 2014. The distribution of preverbal *en* in (West) Flemish: syntactic and interpretive properties. *Lingua* 147: 69-86.

Breitbarth, A. & Haegeman, L. 2015. *'En' en is niet wat we dachten*: a Flemish discourse particle. *MIT Working Papers in Linguistics* 75: 85-102.

Canisius, S., T. Bogers, A. van den Bosch, J. Geertzen & E. Tjong Kim Sang. (2006). Dependency parsing by inference over high-recall dependency predictions. In *Proceedings of the Tenth Conference on Computational Natural Language Learning, CoNLL-X '06*, 176-180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chambers, J. & Trudgill, P. (1980). *Dialectology*. Cambridge: Cambridge University Press.

Coniglio, M. 2011. *Die Syntax der deutschen Modalpartikeln*. Berlin: De Gruyter.

Cornips, L., & Poletto, C. 2005. On standardising syntactic elicitation techniques (part 1). *Lingua* 115: 939-957.

Daelemans, W., Zavrel, J., Van der Sloot, K., & Van den Bosch, A. 2007. *Timbl: Tilburg memory-based learner*. Version, 6, 07-03.

De Vogelaer, G. 2008. *De Nederlandse en Friese subjectsmarkeerders: geografie, typologie en diachronie*. Gent: Koninklijke Academie voor Nederlandse Taal- en Letterkunde.

Fleischer, J., A.N. Lenz & H. Weiß. 2015. Syntax hessischer Dialekte (SyHD). In R.Kehrein, A.Lameli & S.Rabanus (red.), *Regionale Variation des Deutschen. Projekte und Perspektiven*, 261-287. Berlin: De Gruyter.

Ghyselen, A.S. & J. Van Keymeulen. 2014. Dialectcompetentie en functionaliteit van het dialect in Vlaanderen anno 2013. *Tijdschrift voor Nederlandse Taal- en Letterkunde* 130, 117-139.

Glaser, E. & G. Bart. 2015. Dialektsyntax des Schweizerdeutschen. In R.Kehrein, A.Lameli & S.Rabanus (red.), *Regionale Variation des Deutschen. Projekte und Perspektiven*, 81-107. Berlin: De Gruyter.

Goeman, A. & J. Taeldeman. 1996. Fonologie en morfologie van de Nederlandse dialecten. Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten. *Taal & Tongval* 48: 38-59.

Haegeman, L. 1995. *The Syntax of Negation*. Cambridge: CUP.

Haegeman, L. 2014. West Flemish verb-based discourse markers and the articulation of the Speech Act layer. *Studia Linguistica* 68(1): 116-139

Haegeman, L. and C. Greco. 2016. V>2 in West Flemish. *Rethinking verb second: assessing the theory and data.* St John's college. University of Cambridge.

Haegeman, L. & C. Greco. 2018. West Flemish V3 and the interaction of syntax and discourse. *Journal of Comparative Germanic Linguistics.*

Haegeman, L. & V. Hill. 2013. The syntactization of discourse. In R.Folli, C.Sevdali & R.Truswell (red.), *Syntax and its limits*, 370-390. Oxford: OUP.

Haegeman, L. & A. Weir. 2015. The cartography of *yes* and *no* in West Flemish. In J. Bayer, R. Hinterhölzl & A. Trotzke (red.), *Discourse-oriented syntax*, 175-210. Amsterdam: Benjamins.

Haegeman, L. & A. Weir. 2016. Finiteness and response particles in West Flemish. In K. Melum Eide (red.), *Finiteness matters: On finiteness-related phenomena in natural languages*, 211-254. Amsterdam: Benjamins.

Hendrickx, I., Van den Bosch, A., Van Gompel, M., Van der Sloot, Ko. 2016. Frog. A Natural Language Processing Suite for Dutch. Reference guide. In *Language and Speech Technology Technical Report Series 2* (June 2016), Nr. 16, S. Draft 0.13.1

Koelmans, L. 1967. Over de verbreiding van het ontkennende *en*. *De Nieuwe Taalgids* 60: 12-18.

Koppen, M. van. 2017. Complementizer agreement. In M. Everaert & H. van Riemsdijk (red.), *The Wiley-Blackwell Companion to Syntax—second edition*, 923-962. Wiley-Blackwell.

Lybaert, C., B. De Clerck, J. Saelens, & L. Decuypere. 2019. A Corpus-Based Analysis of V2 Variation in West Flemish and French Flemish Dialects. *Journal of Germanic Linguistics* 31.1.

Neuckermans, A. 2008. *Negatie in de Vlaamse dialecten volgens de gegevens van de Syntactische Atlas van de Nederlandse Dialecten (SAND)*. PhD dissertation, Ghent University.

Oostdijk, N. 2000. The Spoken Dutch Corpus. Overview and first evaluation. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis & G. Stainhaouer (red.), Proceedings of LREC-2000 (Second International Conference on Language Resources and Evaluation). Vol. II: 887-894. [http://lands.let.ru.nl/cgn/publs/2000_07.ps]

Overdiep, G.S. 1937. *Stilistische grammatica van het moderne Nederlandsch.* Zwolle: Tjeenk Willink.

Ryckeboer, H. 2013. A West Flemish dialect as a minority language in the North of France. In F. Hinskens & J. Taeldeman (red.), *Language and Space: An International Handbook of Linguistic Variation.* Vol. 3, Dutch: 782-800. Berlin: De Gruyter.

Schmid, H. 1999. Improvements in part-of-speech tagging with an application to German. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann & D. Yarowsky (red.), *Natural language processing using very large corpora*, 13-25. Dordrecht: Springer. [http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/]

Thurmair, M. 1989. *Modalpartikeln und ihre Kombinationen.* Tübingen: Niemeyer.

Vanacker, V.F. 1967. Syntaktische Daten aus französisch-flämischen Tonbandaufnahmen. In *Verhandlungen des 2. Internationalen Dialektologenkongresses. Band II*, 844-855. (*Zeitschrift für Mundartforschung*. Beihefte, Neue Folge, Heft 4)

Vanacker, V.F. & G. De Schutter. 1967. Zuidnederlandse dialekten op de band. *Taal en Tongval* 19: 35-51.

Vanacker, V.F. 1977. Syntaktische overeenkomsten tussen Frans-Vlaamse en Westvlaamse dialekten. *De Franse Nederlanden / Les Pays-Bas Français*: 206-216.

Van Craenenbroeck, J.. 2010. *The syntax of ellipsis. Evidence from Dutch dialects.* New York: OUP.

Van Craenenbroeck, J. & M. van Koppen. 2008. Pronominal doubling in Dutch dialects: big DPs and coordinations. In S. Barbiers, O. Koeneman, M. Lekakou & M. van der Ham (red.), *Microvariation in syntactic doubling. Syntax and Semantics* 36, 207-249. Bingley: Emerald.

Van de Kauter, M., G. Coorman, E. Lefever, B. Desmet, L. Macken, & V. Hoste. 2013. LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal* 3: 103-120. [https://biblio.ugent.be/publication/4228576/file/6807307]

Vandekerckhove, R. 2009. Dialect loss and dialect vitality in Flanders. *International Journal of the Sociology of Language* 196/197: 73-97.

Van den Bosch, A., Busser, G.J., Daelemans, W., and Canisius, S. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch, In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (red.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, Leuven, Belgium, 99-114.

Van Keymeulen, J., Chambers, S., De Tier, V., De Does, J., Depuydt, K., Schoonheim, T., Vandenberghe, R. & Hellebaut, L. 2018. Sustaining the Southern Dutch Dialects: the Dictionary of the Southern Dutch Dialects (DSDD) as a case study for CLARIN and DARIAH. In Skadina, I. & Eskevich, M. (eds.), *CLARIN Annual Conference 2018 Proceedings*. Online beschikbaar via https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf

Weydt, H. 1979. *Die Partikeln der deutschen Sprache.* Berlin: De Gruyter.

Zeijlstra, Hedde. 2004. *Sentential negation and negative concord.* PhD dissertation, Universiteit Utrecht.

Zwart, C. Jan-Wouter. 1993. *Dutch syntax: A minimalist approach.* PhD dissertation, Rijksuniversiteit Groningen.