

# Over het GCND

Anne Breitbarth, Melissa Farasyn, Anne-Sophie Ghyselen, Lien Hellebaut

Versie 21 oktober 2024

## Inhoudsopgave

<b>1</b>	<b>Inleiding</b>	<b>3</b>
<b>2</b>	<b>Achtergrond</b>	<b>5</b>
2.1	‘Nederlandse Dialecten’? . . . . .	5
2.2	Voorgeschiedenis . . . . .	5
2.2.1	Dialectvragenlijsten . . . . .	5
2.2.2	Stemmen uit het Verleden . . . . .	7
2.3	Doelstellingen . . . . .	9
2.4	Financiering . . . . .	10
2.5	Partners . . . . .	10
<b>3</b>	<b>Inhoud van het corpus</b>	<b>12</b>
3.1	Spraakopnames . . . . .	12
3.2	Transcripties . . . . .	13
3.3	Annotaties . . . . .	15
3.3.1	Syntactische annotatie . . . . .	15
3.3.2	Lemma’s en POS-tags . . . . .	19
3.4	Metadata . . . . .	20
3.4.1	Opnamemetadata . . . . .	20
3.4.2	Locatiemetadata . . . . .	21
3.4.3	Sprekermetadata . . . . .	28
<b>4</b>	<b>Corpusstatistieken</b>	<b>30</b>
<b>5</b>	<b>Toegang en gebruiksrechten</b>	<b>30</b>
<b>6</b>	<b>Zoekinterface</b>	<b>31</b>
<b>7</b>	<b>Transcriptieprotocol</b>	<b>31</b>

<b>8</b>	<b>Protocol syntactische annotatie</b>	<b>32</b>
8.1	Subjectsverschijnselen . . . . .	32
8.1.1	Subjectsver(drie)dubbeling . . . . .	32
8.1.2	Subject in objectsvorm . . . . .	33
8.1.3	Presentatief <i>het</i> . . . . .	33
8.2	Uitbreidingen van de zin: TAG en SAT . . . . .	33
8.2.1	Linker zinsperiferie . . . . .	33
8.2.2	V2-bijzinnen – pseudodirecte rede . . . . .	37
8.2.3	Intercalaties/parentheses/interpositio's . . . . .	38
8.3	Complementizer-fenomenen . . . . .	39
8.3.1	Afwijkende comparatieve voegwoorden . . . . .	39
8.3.2	Directe rede ingeleid door <i>van</i> . . . . .	40
8.3.3	Expletief <i>dat</i> . . . . .	40
8.3.4	Beknopte bijzinnen ingeleid door <i>voor</i> of <i>van</i> i.p.v. <i>om</i> . . . . .	41
8.3.5	Afhankelijke <i>ja/nee</i> -vragen ingeleid door <i>als</i> i.p.v. <i>of</i> . . . . .	42
8.3.6	Ingeleide bijzin met hoofdzinsvolgorde (V2-bijzin) . . . . .	43
8.4	Negatieverschijnselen . . . . .	43
8.4.1	Meervoudige negatie(markeerders) . . . . .	43
8.4.2	Adjectieven die met <i>geen</i> gecombineerd worden . . . . .	44
8.4.3	<i>Doen</i> -replieken (met en zonder negatie). . . . .	45
8.5	Andere . . . . .	45
8.5.1	<i>en zo/ of zo / en al / en alles / maar ja / en si en la</i> . . . . .	45
8.5.2	Woordherhaling voor klemtoon . . . . .	46
8.5.3	Spiegelconstructies . . . . .	46
8.5.4	Apokoinouconstructies . . . . .	47
8.5.5	Opsomming van cijfers met betekenis 'ongeveer' . . . . .	47
8.5.6	Discontinue telwoorden <i>drie jaar en half, drie kilo een half</i> . . . . .	48
8.5.7	Tijdsbepalingen met ( <i>van</i> ) <i>te: te zessen, van te zessen</i> . . . . .	48
8.5.8	<i>De die</i> . . . . .	48
8.5.9	Code-switches naar het Frans/Engels/... . . . . .	49
8.5.10	Geluiden en klanknabootsingen . . . . .	50
8.5.11	Exclamatieve infinitiefzinnen . . . . .	50
8.5.12	Circumpositie . . . . .	51

# 1 Inleiding

Het *Gesproken Corpus van de (zuidelijk-)Nederlandse Dialecten* (GCND) is het eerste grammaticaal verrijkte corpus van spontaan gesproken dialecten in het Europees-Nederlandse taalgebied.

Een van de belangrijkste doelen van de wetenschap is om onderliggende structuren te ontrafelen in de schijnbaar – aan de oppervlakte – chaotische en eindeloze variatie in observeerbare fenomenen. Dat is ook het doel van de studie van de grammatica van menselijke talen. Binnen het microcomparatieve of microvariationistische onderzoeksprogramma (Kayne 1996; Brandner 2012) worden nauw verwante talen, vaak variëteiten van eenzelfde taal, met elkaar vergeleken, met als doel om de grenzen van de mogelijke variatie in menselijke talen beter te begrijpen. Net als in de traditionele dialectologie, wordt er daarom binnen dit onderzoekskader met systematische vragenlijsten gewerkt. Veel van de recente dialectatlassen (o.a. SAND/FAND/MAND, ScanDiaSyn, ASIt, SynAlm, SyHD, ...) zijn uit deze traditie ontstaan.<sup>1</sup> Een probleem met dit soort elicitatie is echter dat men alleen vindt waarnaar men zoekt, vaak gebaseerd op bestaande literatuur, en mogelijk interessante fenomenen over het hoofd ziet. Een tweede probleem is *priming*: taalstructuren (bv. woordvolgorde) in de vraag hebben vaak een invloed op de taalstructuren in het antwoord (Van Craenenbroeck, van Koppen en van den Bosch 2019). Een derde probleem is dat het niet altijd eenvoudig is om grammaticale constructies te eliciteren. Zeker op het vlak van syntaxis is elicitatie vaak problematisch. Een voorbeeld is te vinden in de notities van de veldwerkers van het SAND-project<sup>2</sup> bij de antwoorden op SAND-zin 359 (1). We lezen er dat informanten uit meerdere locaties de zin afkeuren en aangeven in hun eigen dialect de standaardtalige woordvolgorde te hanteren met inversie van onderwerp en finiet werkwoord (dus, *met zulk weer kun je niet veel doen*), ondanks het feit dat de veldwerker(s) aangeven de V3-volgorde (zie ook 8.2.1, (d)) geregeld in de spontane spraak van de informanten op te merken.

(1) Met zulk weer je kunt niet veel doen. (SAND-zin 359)

Dit laat zien dat bepaalde grammaticale opties pas in specifieke discoursituaties naar boven komen, en suggereert dat het voor de studie van bepaalde grammaticale kenmerken noodzakelijk is om de systematische bevragingen met vragenlijsten aan te vullen met observaties over spontaan gesproken taal. Dit

<sup>1</sup>SAND = Syntactische Atlas van de Nederlandse Dialecten: <http://www.meertens.knaw.nl/sand/>; (Barbiers, Bennis, Vogelaer, Devos e.a. 2005; Barbiers, Bennis, Vogelaer, Auwera e.a. 2008)

ScanDiaSyn/The Nordic Syntax Database: <https://tekstlab.uio.no/nsd/>; (Lindstad e.a. 2009)

ASIt: <http://asit.maldura.unipd.it/>; (Benincà en Poletto 2005)

SynAlm: <https://www.ling.uni-stuttgart.de/institut/ilg/forschung/projekte/synalm/html/>; (Brandner 2015)

SyHD: <http://www.syhd.info/>; (Fleischer, Lenz en Weiß 2015).

<sup>2</sup>Cf. het commentaargeedeelte bij de data van de bevragingen op [http://www.meertens.knaw.nl/sand/zoeken/lijst\\_met\\_plaatsen.php](http://www.meertens.knaw.nl/sand/zoeken/lijst_met_plaatsen.php).

is geen nieuw inzicht; Blancquaert merkte het al op in 1948 in zijn terugblik op 25 jaar werk aan de *Reeks Nederlandse Dialectatlassen*:

Theoretisch krijgt men betere dialect-gegevens wanneer men er zich toe beperkt, spontane gesprekken af te luisteren en daaruit “in de vlucht” te noteren wat men kan; men zou deze gesprekken, indien men met de nodige materiële middelen was voorzien, ook op band kunnen registreren en ze daarna in fonetische teksten omzetten. En ik wil graag erkennen dat een dergelijke documentatie zeer kostbaar zou zijn ter aanvulling van het systematisch onderzoek met vaste vragenlijst.” (Blancquaert 1948, p. 12)

Om die reden zijn er aan de UGent in de jaren zestig en zeventig meer dan 750 opnames gemaakt van spontane gesprekken met dialectsprekers van hoge leeftijd uit het hele zuidelijk-Nederlandse taalgebied (zie verder 2.2 beneden). Destijds (en ook nog later) werd een deel van die opnames weliswaar getranscribeerd, maar er werd niet met een uniform transcriptieprotocol gewerkt. Vanuit hedendaags perspectief kunnen de opnames als historisch taalmateriaal beschouwd worden, met het voordeel daadwerkelijk gesproken taal te bevatten. Dit is in tegenstelling tot historische corpora van oudere taalstadia, waarbij de overgeleverde tekst(soort)en, zoals poëzie of rechtsteksten, vaak een taal vertegenwoordigen die ver afstaat van spontaan taalgebruik.

Binnen de historische syntaxis heeft de intrede van geparseerde corpora – te beginnen met de Penn Parsed Corpora of Historical English (Kroch 2020) – een ware revolutie betekend. Door teksten te verrijken met woordsoortinformatie (POS-tags) en informatie over de syntactische functie van woorden en woordgroepen (parsing) kan de onderzoeker immers efficiënter op zoek naar taalstructuren.

Binnen die context ontstond het idee om de Gentse audiocollectie via systematische transcriptie en grammaticale annotatie om te smeden tot een toegankelijk grammaticaal geannoteerd corpus met gebruiksvriendelijke zoekmodule. Pas zo kan het bestaande systematische onderzoek naar de grammatica van dialecten op basis van geëliciteerde data immers aangevuld worden met data uit spontaan taalgebruik, die niet alleen toelaten het bestaan van een fenomeen op een bepaalde plaats te attesteren, maar ook de kwantitatieve distributie ervan te documenteren.

## 2 Achtergrond

### 2.1 ‘Nederlandse Dialecten’?

In het GCND-project definiëren we Nederlandse dialecten als dialecten die, historisch gezien, door de Nederlandse cultuurtaal worden overkoepeld. Dat betekent dat ook het Frans-Vlaams, dat vandaag overkoepeld wordt door de Franse cultuurtaal, tot de Nederlandse dialecten wordt gerekend. De classificatie van dialecten is (al lang) een onderwerp van academisch debat. Voor het Nederlands werden die discussies vooral begin de jaren 70 gevoerd door o.a. taalkundigen Jan Goossens en Félicien De Tollenaere. Goossens stelde voor om Nederlandse dialecten te definiëren als dialecten die gesproken worden in gebieden waar Nederlands de cultuurtaal is, waardoor de Frans-Vlaamse dialecten buiten zijn definitie vielen. De Tollenaere bekritiseerde die aanpak, aangezien politieke grenzen niet altijd samenvallen met taalkundige realiteiten, en hij wees erop dat sociologische factoren vaak een grotere rol spelen in de dialectindeling. Hoewel deze discussies in de loop der jaren zijn gaan liggen, zijn ze nog steeds niet opgelost. We kiezen bij het GCND voor een ruime definitie.

De variëteiten die we tot de zuidelijk-Nederlandse dialecten rekenen, zijn de dialecten van het Nederlands die gesproken worden in Vlaanderen, het zuiden van Nederland (Limburg, Noord-Brabant en Zeeland), en enkele gebieden in Noord-Frankrijk. Deze dialecten maken deel uit van een groter dialectcontinuüm, wat betekent dat ze geleidelijk in elkaar (en andere dialecten) overlopen zonder scherpe scheidslijnen. Hoewel de dialecten stapsgewijs meestal onderling verstaanbaar zijn, worden de verschillen duidelijker naarmate men verder reist van het oorspronkelijke dialectgebied.

Binnen de zuidelijk-Nederlandse dialecten zijn er vier hoofdgroepen te onderscheiden: de Vlaamse, Brabantse, Zeeuwse en Limburgse dialecten. Limburgse dialecten worden (*grosso modo*) gesproken in Belgisch en Nederlands Limburg en hebben in Nederland een speciale status als erkende streektaal.<sup>3</sup>

In 3.4.1 geven we meer info over de dialectgroepen die we bij het GCND onderscheiden.

## 2.2 Voorgeschiedenis

### 2.2.1 Dialectvragenlijsten

In het Nederlandse taalgebied kwamen vanaf de jaren 20 van de vorige eeuw een aantal monumentale dialectverzamelingen tot stand op basis van systematische enquêtering:

---

<sup>3</sup>De historische invloed van politieke grenzen heeft ervoor gezorgd dat sommige dialecten anders worden geclassificeerd in de verschillende landen waarin ze worden gesproken. Zo wordt het Limburgs in Nederland erkend als een regionale taal naast het Nederlands volgens het Europees handvest voor regionale talen of talen van minderheden. In België daarentegen wordt het Limburgs beschouwd als een dialect van het Nederlands. Ook in Frankrijk wordt het Frans-Vlaams (dat daar gewoon Vlaams genoemd wordt) als streektaal erkend. Deze erkenningen veranderen niets aan de taalkundige achtergrond van deze dialecten. De naamgeving van het corpus moet gezinszins als een politieke statement opgevat worden.

- Voor de *Reeks Nederlandse Dialectatlassen* werden er tussen 1925 en 1982 meer dan een kwart miljoen dialectzinnen in verschillende plaatsen in Nederland en Vlaanderen verzameld. De atlas bestaat uit 16 volumes met fonetische transcripties uit 1956 plaatsen in Nederland en Vlaanderen en bijhorende taalkaarten. De RND stimuleerde vooral dialectgeografisch onderzoek op fonetisch en fonologisch gebied, maar was ook voor andere taaldomeinen zoals morfologie, syntaxis en lexicon belangrijk.
- Om in de RND ontbrekende klankverschijnselen op te vragen, werden in het Goeman-Taeldeman-Van Reenenproject (GTRP) 1976 woorden en korte zinnen opgevraagd in 622 plaatsen in het Nederlandse taalgebied met als resultaat de *Fonologische Atlas van de Nederlandse Dialecten* (FAND, 1998–2005) en de *Morfologische Atlas van de Nederlandse Dialecten* (MAND, 2005–2008).
- Tegelijk met het verschijnen van het eerste deel van de MAND en het laatste deel van de FAND werd ook het eerste deel van de *Syntactische Atlas van de Nederlandse Dialecten* (SAND) gepubliceerd. Die atlas brengt de uitgebreide variatie in zinsbouw in kaart in de dialecten van het Nederlands aan het begin van de 21ste eeuw (Barbiers, Van der Auwera e.a. 2005, p. 5). De data voor de SAND zijn op gelijkaardige wijze verzameld als voor de FAND en de MAND. In 267 meetpunten, gelijkmatig verdeeld over België, Nederland en Frans-Vlaanderen, werd aan honkvaste, laagopgeleide, oudere dialectsprekers gevraagd – schriftelijk, mondeling of telefonisch – om zinnen te vertalen, aan te vullen en te beoordelen. De FAND, de MAND en de SAND vormen samen een drieluik dat uniek is in de wereld. Geen enkel ander taalgebied beschikt over een dergelijke uitgebreide en gedetailleerde kartering en beschrijving van de grammaticale taalvariatie (Barbiers, Van der Auwera e.a. 2005, p. 5).
- Naast grootschalige dialectprojecten op vlak van fonologie, morfologie en syntaxis bestaan er ook enkele indrukwekkende dialectlexicografische projecten. Het *Woordenboek van de Brabantse Dialecten* (WBD), het *Woordenboek van de Limburgse Dialecten* (WLD) en het *Woordenboek van de Vlaamse Dialecten* (WVD) zijn drie zusterprojecten en documenteren samen de dialectwoordenschat van de zuidelijke Nederlanden (cf. figuur 4), “het gehele gebied ten zuiden van de rivieren” (Goossens 1977: 309) zonder de Zeeuwse eilanden en Goeree-Overflakkee<sup>4</sup>. Sinds 2020 worden de WLD-, WBD- en WVD-data geïntegreerd in de *Database van de Zuidelijk-Nederlandse Dialecten* (<https://dsdd.ivdnt.org/DSDD/search>), een applicatie waarmee je de gegevens aan de hand van een cartografische tool ook op een kaart kan weergeven.

---

<sup>4</sup>De dialectwoordenschat uit die gebieden is namelijk opgenomen in het *Woordenboek der Zeeuwse Dialecten* (WZD), een regionaal woordenboek met een andere opzet.

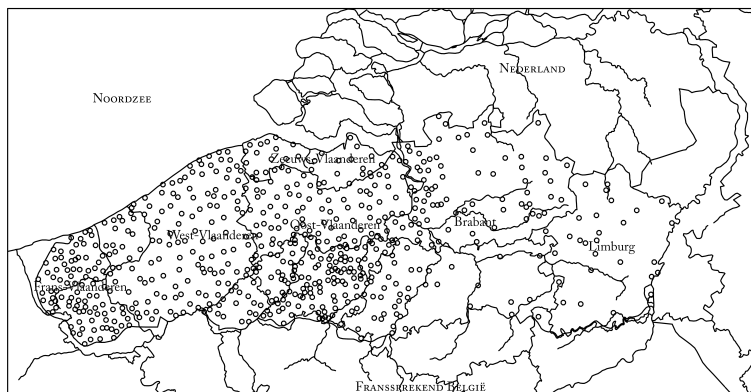
### 2.2.2 Stemmen uit het Verleden

De boven besproken projecten bouwen alle op dialectdata die *geëliciteerd* werden via systematische vragenlijsten. Zoals in de inleiding al werd aangegeven, kennen dergelijke data een aantal beperkingen. Om die reden namen Valeer Vanacker en Willem Pée in de jaren 60 en 70 van de vorige eeuw het initiatief om het *spontaan gesproken dialect* in Vlaanderen, Zeeuws-Vlaanderen en Frans-Vlaanderen op band vast te leggen (V. F. Vanacker en de Schutter 1967). Tussen 1963 en 1976 werden 783 dialectopnames gemaakt van zogenaamde ‘vrije gesprekken’ op 550 verschillende locaties, goed voor meer dan 700 uur aan gesproken dialect.

Bij de selectie van de zegspersonen werden vrij strenge criteria gehanteerd om ervoor te zorgen dat de onderzoekers werkten met authentieke en onvervalste dialecten. De ideale zegspersoon was lokaal geboren en getogen, van een zekere leeftijd, laaggeschoold en honkvast. Daarnaast waren zowel de ouders als de partner van de spreker idealiter uit dezelfde plaats afkomstig en sprekers van hetzelfde dialect. Dit profiel leidde vaak naar boeren en arbeiders, die het minst waren beïnvloed door de standaardtaal en andere registers. Veel van de informanten die werden geïnterviewd, waren rond 1900 geboren, lang vóór de invoering van de leerplicht in België in 1914. De mensen spraken vrijwel uitsluitend hun lokale dialect en hadden vaak geen toegang gehad tot andere vormen van onderwijs of media die de taal konden beïnvloeden. Sommige van deze sprekers kwamen uit dorpen die vandaag de dag niet meer bestaan, zoals de verdwenen polderdorpen Lillo, Oorderen en Wilmarsdonk, die in de jaren 60 moesten wijken voor de uitbreiding van de Antwerpse haven. Door de toenemende scholing en mobiliteit in de 20e eeuw begon dit type dialectspreker echter zeldzamer te worden, wat het werk van de onderzoekers extra dringend maakte.

De opnames werden zoveel mogelijk bij de zegspersoon thuis gemaakt om de natuurlijke omgeving te behouden en de spontaniteit van het gesprek te bevorderen. Om de invloed van de standaardtaal en de onderzoeker (vaak studenten taalkunde in het kader van hun licentiaatsverhandeling) zelf te minimaliseren, werd het gesprek vaak geleid door een tussenpersoon, die meestal ook hetzelfde dialect sprak. Deze persoon kende de zegspersoon goed en wist hem op zijn gemak te stellen, terwijl de onderzoeker zich op de techniek richtte in een andere kamer. Dit hielp om de zogenaamde “observer’s paradox” te vermijden, waarbij de aanwezigheid van een onderzoeker de natuurlijke taalproductie van de informant zou kunnen verstoren. Toch hadden deze huiselijke opnames ook hun nadelen: er waren vaak achtergrondgeluiden, zoals tikkende klokken, kletterende kookpotten of een blaffende hond, die de opnames verstoorden.

Wat de inhoud van de gesprekken betreft, werd de zegspersoon de vrije hand gegeven om over een onderwerp naar keuze te spreken. In de praktijk werden de gesprekken echter vaak gestuurd naar onderwerpen die een rijke bron van verhalen opleverden, zoals oude landbouwtechnieken, traditionele ambachten (zoals mandenmaken of kuipen), of de ervaringen van veteranen uit de Eerste en Tweede Wereldoorlog. Veel van de informanten hadden de introductie van moderne technologieën zoals de fiets, auto, radio, elektriciteit en televisie



Figuur 1: Oorspronkelijke opnames binnen het project Stemmen uit het Verleden

meegemaakt, en rapporteren hierover in de opnames.

Als je de verspreiding van de opnames op kaart bekijkt, valt op dat Oost- en West-Vlaanderen het sterkst vertegenwoordigd zijn. Dit komt deels doordat de meeste studenten Nederlands aan de Universiteit Gent uit deze provincies afkomstig waren en vaak optraden als tussenpersoon bij het zoeken naar zegspersonen in hun thuisdorpen. Vooral in deze provincies werd een groot aantal opnames gemaakt, soms zelfs meerdere per dorp. In Frans-Vlaanderen werd er een extra inspanning geleverd om opnames te maken in elk dorp waar nog Nederlands werd gesproken, met de overtuiging dat het Vlaamse dialect in deze regio op het punt stond te verdwijnen (V.-F. Vanacker 1967). Deze opnames zijn nu vaak de enige bewaarde getuigen van het Vlaams dat er eeuwenlang gesproken werd. Voor Zeeuws-Vlaanderen speelde professor Johan Taeldeman een belangrijke rol in het maken van zoveel mogelijk opnames in deze streek.

De opnames – die later de naam 'Stemmen uit het Verleden' kregen – vormen een ongeziene registratie van de traditionele dialecten en zijn een belangrijke aanvulling bij het materiaal in de Reeks Nederlandse Dialectatlassen en de MAND/FAND/SAND. Bovendien complementeren ze een verzameling dialectopnames uit verschillende plaatsen in Nederland, die al in de jaren 50 op initiatief van Jo Daan van het Dialectenbureau in Amsterdam gemaakt werd (Rensink 1962). De Stemmen uit het Verleden vormen de grootste verzameling levensverhalen van laaggeletterde Vlamingen met unieke getuigenissen over het leven van omstreeks 1900. De collectie is dus niet alleen taalkundig van belang,



maar is ook volkskundig heel waardevol.<sup>5</sup>

Helaas bleef de collectie lang onderbenut voor zowel taal- als volkskundig onderzoek. Het materiaal werd weliswaar in de loop der jaren gedigitaliseerd (dankzij o.a. Johan Taeldeman) en in 2014 online raadpleegbaar gemaakt via [www.dialectoket.be](http://www.dialectoket.be) (dankzij o.a. Jacques Van Keymeulen), maar de bruikbaarheid voor onderzoek bleek beperkt doordat:

- maar een fractie van het materiaal getranscribeerd was;
- de weinige beschikbare transcripties niet vertrokken vanuit een uniform transcriptieprotocol en dus erg heterogeen waren;
- de meeste beschikbare transcripties enkel in de vorm van scans van de oorspronkelijke getikte kopij of zelfs handgeschreven tekst beschikbaar waren;
- het materiaal daardoor niet digitaal doorzocht kon worden op woordvormen (en al zeker niet op syntactische patronen of constructies).

### 2.3 Doelstellingen

Om de collectie dialectopnames te kunnen laten dienen waarvoor ze geconcipeerd werd — taalkundig onderzoek – werd vanaf 2019 financiering gezocht en gevonden met als doel:

- de opnames uit de collectie Stemmen uit het Verleden (aangevuld met nieuwe opnames en opnames uit de collectie van het Meertens Instituut, zie hierover 3.1) te transcriberen op basis van een uniform transcriptieprotocol;
- de gemaakte transcripties te verrijken met informatie over de woordsoort van individuele woorden ('pos-tags') en over de syntactische functies van de woordgroepen ('parsing');
- een online applicatie met zoekmodule te ontwikkelen zodat het corpus efficiënt raadpleegbaar en doorzoekbaar is voor iedere onderzoeker die met de data aan de slag wil.

Sinds 2024 loopt een vervolgproject (GCND+) om:

- nog meer opnames (ook uit de collectie van het Meertens Instituut) te transcriberen, zodat het GCND niet langer een corpus is van enkel *zuidelijk-Nederlandse* dialecten, maar er een van *Nederlandse* dialecten wordt;
- de grammaticale verrijking verder manueel te corrigeren;
- een parser te hertrainen op de al geverifieerde GCND-annotaties zodat die in de toekomst beter met dialectische spraak omkan en er op die manier minder manueel verificatiewerk nodig is.

---

<sup>5</sup>Om de collectie ook inhoudelijk toegankelijk te maken, werd er een apart project opgezet om de banden ook inhoudelijk te annoteren aan de hand van een nieuwe, hiërarchische trefwoordenthesaurus. De thematische zoekmodule is beschikbaar via <https://www.dialectloket.be/geluid/stemmen-uit-het-verleden/zoeken/>

## 2.4 Financiering

Het project werd mogelijk gemaakt dankzij verschillende kredieten:

- 2024-2028: FWO Middelzware onderzoeksinfrastructuur I.0.021.24N (GCND+)
- 2021-2024: FWO Postdoctoraal mandaat senior 1.2.P79.22N aan M. Farasyn (Frans-Vlaamse opnames)
- 2020-2024: FWO Middelzware onderzoeksinfrastructuur I.0.101.20N (GCND)
- 2018-2020: FWO Navorserskrediet 1.5.310.18N aan A. Breitbarth (pilootproject)
- 2018-2021: FWO Postdoctoraal mandaat junior 1.2.P79.19N aan M. Farasyn (Frans-Vlaamse opnames)
- 2019-2021: Subsidies provincies Zeeland, West-Vlaanderen en Oost-Vlaanderen (pilootproject)

## 2.5 Partners

Aan het GCND-infrastructuurproject werkten verschillende partners: taalkundigen van de UGent zorgden voor de projectaanvragen die de opbouw van de infrastructuur financierden, de protocollen voor transcriptie en grammaticale verrijking, nieuwe dialectopnames en de database van transcripties en annotaties (dat laatste in samenwerking met het Ghent Centre for Digital Humanities). Het Instituut voor de Nederlandse Taal (INT) ontwikkelde de applicatie waarmee de taalkundige annotaties doorzocht kunnen worden en staat in voor de hosting van de collectie. Het Meertens Instituut KNAW leverde geluidsopnames aan die in Nederland gemaakt werden. Variaties vzw hielp met het zoeken van vrijwilligers. De Rijksuniversiteit Groningen ontwikkelde de Alpino-parser voor het Nederlands en hielp om die in te zetten voor de dialecten in het GCND-corpus.

### 1. UGent

- Prof. dr. Anne Breitbarth (promotor-woordvoerder, Vakgroep Taalkunde)
- Prof. dr. Anne-Sophie Ghyselen (Vakgroep Taalkunde)
- Dr. Melissa Farasyn (Vakgroep Taalkunde)
- Lien Hellebaut (Vakgroep Taalkunde)
- Prof. dr. Timothy Coleman (Vakgroep Taalkunde)
- Prof. dr. Els Lefever (Ghent Centre for Digital Humanities)
- Frederic Lamsens (Ghent Centre for Digital Humanities)
- Sally Chambers (Ghent Centre for Digital Humanities)

- Pieterjan De Potter (Ghent Centre for Digital Humanities)
2. Instituut voor de Nederlandse Taal (INT)
    - Katrien Depuydt
    - Dr. Jesse de Does
    - Koen Mertens
  3. Meertens Instituut KNAW
    - Prof. dr. Antal van den Bosch
    - Prof. dr. Marjo van Koppen
    - Douwe Zeldenrust
  4. Variaties vzw
    - Prof. em. dr. Jacques Van Keymeulen
    - Veronique De Tier
  5. Rijksuniversiteit Groningen
    - Prof. dr. Gertjan van Noord

Naast deze partners hebben veel jobstudenten uit België en Nederland bijgedragen aan het GCND. Zij werkten namelijk mee aan de transcripties en annotaties. Bovendien hebben talloze vrijwilligers geholpen met het corrigeren van de transcripties. Allen verdienen zonder uitzondering onze oprechte dank.

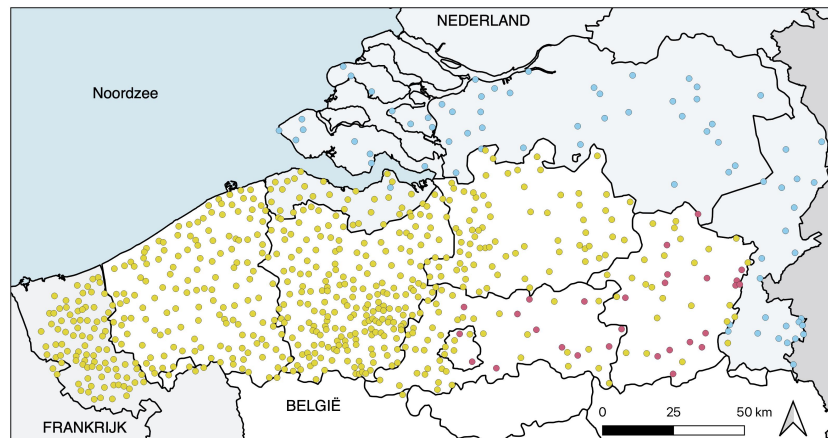
Bij het GCND+-project (2024-2028) is een nieuwe partner betrokken: de vakgroep LT3 van de Universiteit Gent (met prof. dr. Veronique Hoste als promotor), die verantwoordelijk is voor het hertrainen van de parser.

## 3 Inhoud van het corpus

### 3.1 Spraakopnames

Het GCND bouwt in eerste instantie op **opnames uit de collectie 'Stemmen uit het Verleden'** (zie hierover 2.2.2), maar daarnaast werden ook nog andere opnames toegevoegd (cf. Figuur 2):

1. In 2022 werden voor het GCND 30 nieuwe opnames gemaakt om de leemtes in het oosten van het gebied (Brussel, Vlaams-Brabant en Belgisch-Limburg) op te vullen. Daarbij werd geprobeerd dezelfde selectiecriteria als uit de jaren 60 aan te houden. Dat bleek evenwel niet altijd haalbaar – de sprekers in de nieuwe opnames zijn vaak mobieler en beter opgeleid dan die in de oudere opnames – maar alle sprekers hebben wel gemeen dat ze hun taalsocialisatie vóór de jaren 60 meegemaakt hebben (voor het grootschalige dialectverlies) en dat ze dialect als moedertaal hebben.
2. Verder werd de collectie ook aangevuld met 73 opnames uit de Dialectenbank van het Meertens Instituut, meer specifiek uit Zeeland, Noord-Brabant, en Nederlands Limburg. Die opnames werden opgenomen om dialectologisch 'logische' gehelen te kunnen bestrijken (Vlaamse, Limburgse, Brabantse, Zeeuwse en Limburgse dialecten), los van de staatsgrens die vanuit dialectologisch perspectief van secundair belang is.



Figuur 2: Spreiding van de opnames in het GCND (groen: *Stemmen uit het Verleden*, blauw: Dialectenbank KNAW Meertens Instituut, rood: nieuwe GCND-opnames))

## 3.2 Transcripties

Als eerste stap in het GCND-project werden alle opnames manueel (en dus niet via spraakherkenning, cf. Ghyselen, Breitbarth e.a. 2020) getranscribeerd volgens een gestandaardiseerd transcriptieprotocol. Dat protocol werd gepubliceerd en gemotiveerd in Ghyselen, Van Keymeulen e.a. 2020. De meest recente versie, met extra voorbeeldmateriaal, is te vinden op de GCND-website via onderstaande link: <https://www.gcnd.ugent.be/corpusannotatie/>

De transcripties werden gemaakt met het softwarepakket ELAN, dat het mogelijk maakt de transcripties te aligneren met de audio. De alignering gebeurde op segmentniveau (tijdscode voor begin en einde van een syntactische eenheid die meestal overeenstemt met een zin), niet op woordniveau. Om redenen van haalbaarheid en consistentie werd ervoor geopteerd om orthografisch, niet fonetisch te transcriberen. We schrijven de dialecten dus neer vertrekkend vanuit de Nederlandse spelling, alsof de woorden, vormen en constructies ook in het Standaardnederlands voorkomen, en maken abstractie van fonetische variatie bij inhoudswoorden. Realisaties als [kønɪŋk] or [kynɪŋ] worden dus neergeschreven als *koning*. Doordat de orthografische transcriptie echter gealigneerd is met de audio, kunnen onderzoekers, die geïnteresseerd zijn in de fonetiek of fonologie van de zuidelijk-Nederlandse dialecten, eenvoudig relevante woorden of passages in de audio raadplegen en zelf fonetische annotaties toevoegen (volgens eigen conventies en standaarden).

De orthografische transcriptie gebeurde in twee lagen: in de ene laag (de zogenaamde ‘lichte vernederlandsing’) blijft de transcriptie dicht bij het dialect, terwijl de andere laag (de ‘zware vernederlandsing’) dicht bij het Standaardnederlands staat.

- (2) a. m#e wijder de tweeëntwintigste oktobere ‘s morgens om acht ure vertrokken.  
b. we hebben wij de tweeëntwintigste oktober ‘s morgens om acht uur vertrokken.

In de lichte vernederlandsing 2a worden niet-standaardtalige woordenschat, functiewoorden, morfologie en syntaxis bewaard en neergeschreven volgens de Nederlandse spellingsregels. Clitische elementen – zoals m#e (‘wij hebben’) – worden als clusters getranscribeerd, waarbij de afzonderlijke delen van de cluster van elkaar worden gescheiden met een #. Dat wordt gedaan om de koppeling met de tweede laag te vergemakkelijken. In beide lagen staan precies even veel elementen, in geval van 2 dus twaalf.

Bij de zware vernederlandsing 2b worden ook morfologische elementen en dialectale functiewoorden vernederlandsd. De vorm *oktobere* wordt *oktober*, *wijder* wordt *wij*, en clitische elementen worden als afzonderlijke onderdelen weergegeven; m#e wordt bijvoorbeeld *we hebben*. Niet-standaardtalige woordenschat blijft wel behouden omdat het vaak onduidelijk is hoe die moet worden ‘vertaald’ in het Standaardnederlands. De syntaxis in de zware vernederlandsing blijft getrouw aan het dialect, zoals aan de verdubbeling van het onderwerp te

zien is (*m#e wijder* wordt *we hebben wij* in plaats van *we hebben*). De zware vernederlandsing maakt het ook voor gebruikers zonder een uitgebreide dialectkennis mogelijk de data efficiënt te doorzoeken, en dient bovendien als input voor de op standaardspelling getrainde tools voor de verdere semi-automatische taalkundige verrijking.

Van de spraak van de interviewer(s) werd er alleen een lichte vernederlandsing gemaakt, om de context van de tekst van de spreker(s) toegankelijk te houden.

Alle dialectopnames zijn getranscribeerd door studenten Taal- en Letterkunde, die na een korte opleiding over ELAN en het transcriptieprotocol op basis van de kwaliteit van een testtranscriptie geselecteerd werden. De dialectkennis van de studenten – die opgroeiden in een tijd van doorgedreven dialectverlies (cf. Ghyselen and Van Keymeulen 2014) – bleek echter soms te beperkt om de (historische) opnames volledig te doorgronden. Een bijkomend probleem was dat ze vaak ook weinig vertrouwd waren met de gespreksonderwerpen die in de banden worden aangesneden, zoals traditionele landbouw of verdwenen ambachten, waardoor hun transcripties onvermijdelijk een aantal fouten of onnauwkeurigheden bevatten. Woorden als *vaars* (‘jonge koe die nog niet of voor de eerste maal gekalfd heeft’), *hofstede* (‘boerderij’) of *betterapen* (‘bieten’) bleken bijvoorbeeld soms verkeerd begrepen te worden en getranscribeerd als *vijs*, *hoofdsteden* of *boter rapen*. Vele woorden of passages kregen ook het label ‘???’ , waarmee de transcribent aangeeft die niet te verstaan. Om dit soort problemen op te lossen, werden vrijwilligers ingeschakeld die wel nog goed vertrouwd zijn met zuidelijk-Nederlandse dialecten en ze dus beter begrijpen dan de (meestal jongere) transcribenten, maar in veel gevallen niet over de taalkundige expertise en/of digitale vaardigheden beschikken om zelf in ELAN te transcriberen. Zij corrigeerden de transcripties. Een foutenanalyse bij een steekproef van 13 afgewerkte transcripties – die voor die analyse geverifieerd werden door variatilinguïsten met moedertaalkennis van de dialecten op de banden in kwestie – leverde een Word Error Rate van 1.3 procent op, wat wijst op een degelijke transcriptiekwaliteit (Ghyselen, Deklerck e.a. 2024).

### 3.3 Annotaties

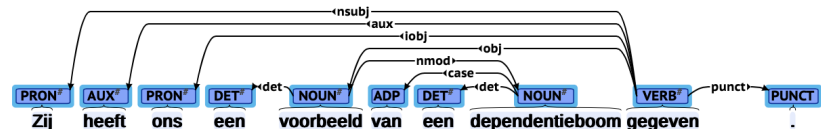
De GCND-transcripties zijn **syntactisch geannoteerd**, of *geparset*, op basis van **dependentiestructuren**. Bij dependentie-parsing wordt de grammaticale structuur van een zin geanalyseerd als relaties tussen woorden/domeinen. Die informatie wordt in de vorm van labels toegevoegd aan de transcripties. De output van het annotatie-proces is bij het GCND een XML-formaat, waarin ook de tijdscode van de opnames staan, zodat elke zin, inclusief alle annotaties met de bijbehorende opname gealigneerd kan worden.

#### 3.3.1 Syntactische annotatie

Voor de syntactische annotatie van het (eerste deel van het) GCND is de tool **Alpino** gebruikt (Van Noord 2006), een tool die voor de syntactische annotatie (parsing) van het Nederlands, meer specifiek het LASSY-corpus van geschreven Nederlands (Noord e.a. 2013) ontwikkeld en getraind is (Farasyn e.a. 2022). Omdat het GCND een corpus van gesproken dialect is, moeten er soms aanpassingen in de LASSY-parsingrichtlijnen gebeuren om spreektaalige en dialectische fenomenen te kunnen annoteren. Hiervoor valt het GCND terug op de syntactische annotatie van het Corpus Gesproken Nederlands (CGN) (Hoekstra e.a. 2002).

Alpino produceert **dependentiestructuren**, dwz. een (gerichte acyclische) graaf met **gelabelde knopen** en **bogen** tussen de knopen, waarbij de labels informatie over de categoriale eigenschappen (woordsoort/frasale informatie zoals constituentensoort/zinssoort) (zie 3.3.2) en dependenties tussen de knopen bevatten (zie 3.3.1). Die laatste soort labels geeft aan in welke relatie een zinsdeel tot andere zinsdelen in de zin staat. Er wordt telkens een onderscheid gemaakt tussen het hoofd en een complement of modificeerder, waarbij een complement door het hoofd geselecteerd wordt (in termen van thematische structuur/subcategorisatie) en een modificeerder het hoofd modificeert.

Fig. 3 demonstreert hoe zo'n graaf van gelabelde knopen en gelabelde bogen tussen de knopen eruit ziet: De knopen – hier elk woord plus de punctuatie – bevatten de woordsoortinformatie (blauwe rechthoeken). De bogen zijn met de grammaticale functies (dependenties) gelabeld.

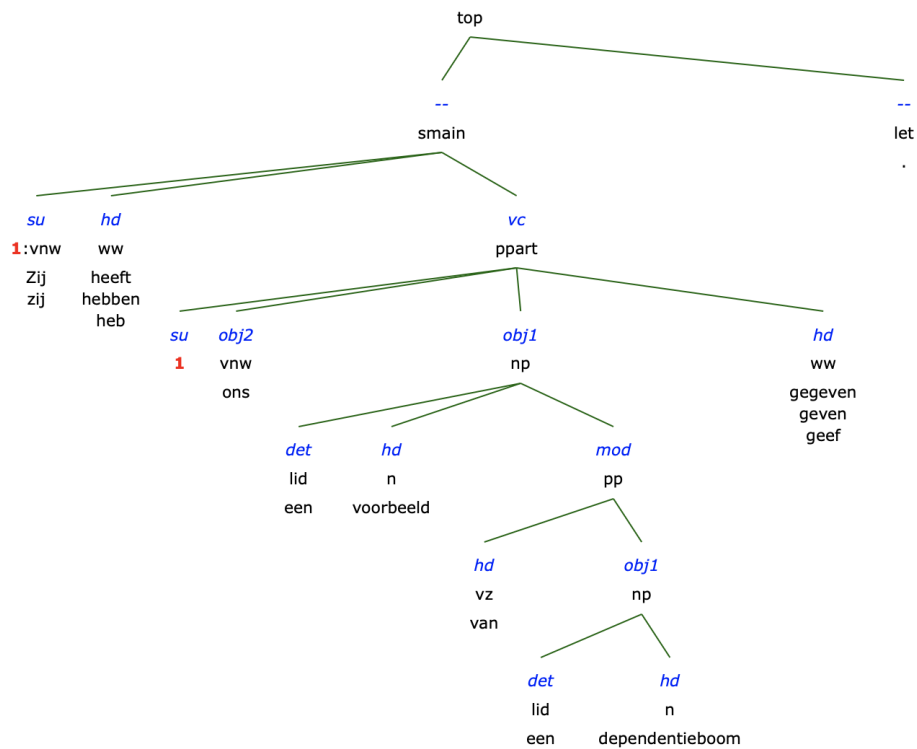


Figuur 3: Dependentiestructuur van de zin *Zij heeft ons een voorbeeld van een dependentieboom gegeven.*

Elke zin heeft het lexicale werkwoord als hoofd (in ons voorbeeld *gegeven*), waarvan alle andere elementen in de zin afhankelijk zijn, zoals het onderwerp (*zij*), lijdend voorwerp (*een voorbeeld van een dependentieboom*), modificatoren enz.,

en in het voorliggende voorbeeld ook het hulpwerkwoord (*heeft*), aangezien de zin in de voltooid tegenwoordige tijd staat. Intern complexe zinsdelen (constituenten) worden als **lokale dependentedomeinen** beschouwd en hebben daarom zelf ook een hoofd, waarvan de rest van dat zinsdeel afhangt. In ons voorbeeld is het lijdend voorwerp intern complex: het naamwoord *voorbeeld* heeft een artikel (*een*) en een complement (*van een dependentieboom*). Alleen dat hoofd heeft in Fig. 3 een directe dependenterelatie met het hoofd van de zin (*gegeven*).

Zij heeft ons een voorbeeld van een dependentieboom gegeven .



Figuur 4: Dependentieboom van de zin *Zij heeft ons een voorbeeld van een dependentieboom gegeven.*

Alpino neemt echter ook de hiërarchische structuur van de zin in acht (echter in functie van de lexicale subcategorisatie-/selectie-eisen van de hoofden, niet 100% vergelijkbaar met constituentenstructuur), en maakt een onderscheid tus-



sen **gelede** (vertakkende) en **ongelede** (terminale) **knopen**. Ongelede knopen zijn de woorden van de zin met hun woordsoort-label (POS-tag) en het lemma (zie 3.3.2). Gelede knopen zijn de lokale dependentiedomeinen, die ipv. een POS-tag een (frasaal) categorie-label hebben. Fig. 4 laat zien hoe dit eruit ziet.

De POS-tags en lemma's bij de ongelede knopen (dus die boven de woorden van de zin) worden in 3.3.2 besproken. In Fig. 4 zijn er zes gelede knopen:

- top,
- [- -: smain],
- [vc: ppart],
- [obj1: np],
- [mod: pp], en binnen dit laatste nog eens
- [obj1: np]

Daarbij is het blauwe en in de lijst hierboven links staande label het **dependentielabel** en het zwarte/rechtse label het **categorielabel**. De gelede knopen zijn dus de moeders van telkens een lokaal dependentiedomein. Elke lokale dependentiestructuur heeft ook haar eigen hoofd (**hd**). Enkel leestekens, “.” in fig. 4, staan buiten de eigenlijke zin en vormen samen met de smain de gelede knoop “top”. Hier is er geen dependentierelatie omdat de ene de andere dochter niet selecteert/modificeert.

In de CGN/Alpino-annotatie hebben complexe werkwoordsvormen, zoals de voltooid tegenwoordige tijd in ons voorbeeld, een geneste dependentiestructuur. Dit komt omdat de samenstellende hoofden verschillende subcategorisatie-eisen hebben. Op zinsniveau is dus het finiete hulpwerkwoord *heeft* het hoofd en dat hoofd neemt een onderwerp (**su**) en een verbaal complex (**vc**) als dependents. Van dat verbaal complex is dan het lexicale werkwoord *gegeven* het hoofd, met drie complementen:

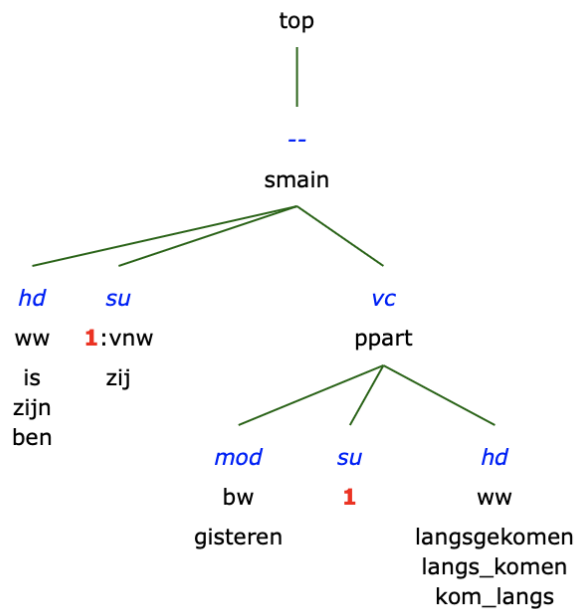
- het onderwerp, dat via de index **1** als identiek aan het onderwerp van *heeft* gemarkeerd wordt,
- het meewerkend voorwerp (**obj2**) en
- het lijdend voorwerp (**obj1**)

Van het lijdend voorwerp [**obj1**: np] is het naamwoord (n) *voorbeeld* het hoofd. Van de modificeerder [**mod**: pp] is het voorzetsel (**vz**) *van* het hoofd, en van het complement [**obj1**: np] van dat voorzetsel is het naamwoord (n) *dependentiedomeinen* het hoofd.<sup>6</sup> Het hoofd produceert dus een lokaal dependentiedomein dat categoriaal overeenkomt met de categorie van het hoofd:  $n \rightarrow np$ ,  $vz \rightarrow pp$ , het deelwoord *gegeven*  $\rightarrow ppart$  (participiaal vc).

<sup>6</sup>Het feit dat ook het complement van een voorzetsel als **obj1** gelabeld wordt komt omdat het direct door het voorzetsel geselecteerd wordt, precies zoals een lijdend voorwerp door een werkwoord.

Het is belangrijk te onthouden dat in dependentiebomen vooral de selectie-/dependentie-/modificatie-relatie tussen de knopen gerepresenteerd wordt. Tokens die qua selectie/dependentie dicht bij elkaar horen worden ook naast elkaar gezet. Daardoor is de lineaire volgorde van de tokens in de boom niet altijd dezelfde als in de transcriptie, zie fig. 5.

gisteren is zij langsgesproken



Figuur 5: Dependentieboom van de zin *Gisteren is zij langsgesproken*.

In het XML-bestand (zie fig. 6) staat echter bij elk token wat zijn lineaire positie in de zin is (in fig. 5 niet zichtbaar). Zo staat er bij de annotatie van het bijwoord *gisteren*: `node begin="0" end="1"`, wat betekent dat het het eerste token in de zin is. Bij het hulpwerkwoord *is* staat er dan `node begin="1" end="2"`, en bij het onderwerp *zij* `node begin="2" end="3"`.

```

▼<alpino_ds version="1.15">
  <parser build="Alpino-git20240510.104003-glibc2.35-SICStus3.12.11-x86_64-linux-
  glibc2.5" date="2024-08-10T15:48" cats="1" skips="0"/>
  ▼<node begin="0" cat="top" end="4" id="0" rel="top">
    ▼<node begin="0" cat="smain" end="4" id="1" rel="--">
      <node begin="1" end="2" frame="verb(unacc,sg_is,aux_psp_zijn)" his="normal"
      his_l="normal" id="2" infl="sg_is" lcat="smain" lemma="zijn" pos="verb"
      postag="WW(pv,tgw,ev)" pt="ww" pvagr="ev" pvtijd="tgw" rel="hd" root="ben"
      sc="aux_psp_zijn" sense="ben" stype="declarative" tense="present" v_per="thi"
      word="is" wvorm="pv"/>
      <node begin="2" case="nom" def="def" end="3"
      frame="pronoun(nwh,thi,both,de,nom,def)" gen="de" genus="fem" getal="ev"
      his="normal" his_l="normal" id="3" index="1" lcat="np" lemma="zij"
      naamval="nomin" num="both" pdtype="pron" per="thi" persoon="3v" pos="pron"
      postag="VNW(pers,pron,nomin,vol,3v,ev,fem)" pt="vnw" rel="su" rnum="sg"
      root="zij" sense="zij" status="vol" vwtype="pers" wh="nwh" word="zij"/>
      ▼<node begin="0" cat="ppart" end="4" id="4" rel="vc">
        <node begin="0" end="1" frame="tmp_adverb" his="normal" his_l="normal" id="5"
        lcat="advp" lemma="gisteren" pos="adv" postag="BW()" pt="bw" rel="mod"
        root="gisteren" sense="gisteren" special="tmp" word="gisteren"/>
        <node begin="2" end="3" id="6" index="1" rel="su"/>
        <node begin="3" buiging="zonder" end="4"
        frame="verb(unacc,psp,ninv(intransitive,part_intransitive(langs)))"
        his="normal" his_l="part-V" id="7" infl="psp" lcat="ppart" lemma="langs_komen"
        pos="verb" positie="vrij" postag="WW(vd,vrij,zonder)" pt="ww" rel="hd"
        root="kom_langs" sc="part_intransitive(langs)" sense="kom_langs"
        word="langsgekomen" wvorm="vd"/>
      </node>
    </node>
  </node>
  <sentence sentid="129.125.14.125">gisteren is zij langsgekomen</sentence>
</alpino_ds>

```

Figuur 6: Alpino-XML-bestand van de zin *Gisteren is zij langsgekomen*.

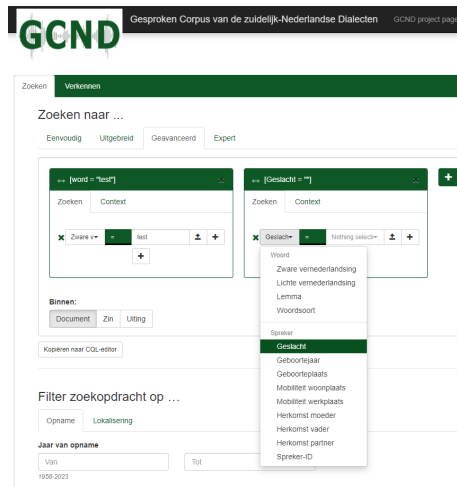
Het GCND volgt in zijn syntactische annotatie in eerste instantie de LASSY-richtlijnen,<sup>7</sup> gecombineerd met de CGN-richtlijnen voor spreektaalige fenomenen.<sup>8</sup> Daarnaast zijn er een aantal dialectische constructies waarvoor het GCND eigen parsingbeslissingen moest nemen. Die worden in hs. 8 extra toegelicht.

### 3.3.2 Lemma's en POS-tags

Onderaan in de boom – in de ongelede knopen – staan de individuele tokens (doorgaans woorden). Hun categorielabel komt overeen met hun woordsoort (**POS-tag**). Bij werkwoorden wordt eerst de aangetroffen vorm aangegeven zoals die in de transcripties staat (in fig. 4: *heeft, gegeven, ...*), dan het **lemma** (in dit geval de infinitief *hebben, geven*). Verder staat er ook nog de “betekenis” van het woord bij, dus *heb, geef* (= de stam van het werkwoord).

<sup>7</sup>[https://www.let.rug.nl/~vannoord/Lassy/sa-man\\_lassy.pdf](https://www.let.rug.nl/~vannoord/Lassy/sa-man_lassy.pdf).

<sup>8</sup><http://www.ccl.kuleuven.be/Papers/NT2002-vdw-et-al.pdf>,  
[https://www.mpi.nl/tools/synpathy/syn\\_prot.pdf](https://www.mpi.nl/tools/synpathy/syn_prot.pdf).



Figuur 7: Sprekermetadata specificeren in de geavanceerde zoekmodus

### 3.4 Metadata

In de corpusapplicatie kan in de uitgebreide, geavanceerde en expertmodus gefilterd worden op verschillende types metadata, die in drie categorieën ondergebracht zijn: opname-, locatie- en sprekermetadata. De filter voor de eerste twee categorieën vind je onder 'Filter zoekopdrachten op'. Om sprekermetadata te specificeren:

- ga je in de *uitgebreide zoekmodus* naar het tabblad 'Spreker' (naast 'Woord') of
- voeg je in de *geavanceerde zoekmodus* een zoekblok toe waarin je via het dropdownmenu sprekergegevens specificeert (cf. afbeelding fig. 7)

In wat volgt, lichten we de metadata uitgebreider toe.

#### 3.4.1 Opnamemetadata

##### Jaar van opname

In welk jaar werd de opname gemaakt?

##### Verificatiestatus

De GCND-transcripties werden in eerste instantie gemaakt door studenten taalkunde, en vervolgens nagekeken door vrijwilligers met een sterke dialectcompetentie. De transcripties werden in een volgende stap via Alpino automatisch grammaticaal verrijkt met informatie over lemma's, woordsoorten en de relationele functie van woorden en woordgroepen in de zin. Aangezien Alpino getraind is op standaardtalige (vaak schriftelijke) data, sluipen er in de automatische analyse van spontaan gesproken dialect dikwijls fouten. Om die reden

werd de Alpino output manueel gecorrigeerd door daarvoor opgeleide studenten taalkunde.

Op 24 oktober 2024, wanneer de eerste versie van het GCND werd gelanceerd, was het verificatiewerk (van zowel de transcripties als de Alpino-output) nog niet voor alle opnames rond. Omdat in sommige contexten ook niet-gecontroleerde data relevant zijn, werden die ook in de GCND-applicatie opgenomen. De gebruiker kan zelf specificeren in welke mate het doorzochte materiaal geverifieerd moet zijn:

- Transcriptie gecontroleerd, taalkundige verrijking gecontroleerd (✓✓)
- Transcriptie gecontroleerd, taalkundige verrijking niet gecontroleerd (✓✓)
- Transcriptie gecontroleerd, geen taalkundige verrijking > door technische issues zijn deze bestanden nog niet grammaticaal verrijkt (✓×)
- Transcriptie niet gecontroleerd, taalkundige verrijking niet gecontroleerd. (× ×)

Veel hangt bij de specificatie van de verificatiestatus af van de aard van de zoekopdracht. Wanneer je bijvoorbeeld op zoek bent naar specifieke woorden of woordcombinaties, is verificatie van de grammaticale verrijking irrelevant.

In de GrETEL Treebank zijn enkel volledig geverifieerde data opgenomen; de redenering is immers dat de gebruiker daar gericht op zoek gaat naar abstractere syntactische patronen, en de zoekresultaten daarvan zijn te weinig betrouwbaar voor niet-geverifieerde data.

Bij de weergave van de resultaten (ook wanneer je die exporteert) wordt de verificatiestatus telkens weergegeven aan de hand van vinkjes (vet vinkje = gecontroleerd, vinkje 1=transcriptie, vinkje 2= taalkundige verrijking). Zo kun je makkelijker bepalen voor welke zoekresultaten je graag nog even ter controle de audio beluistert.

### **Bron**

Is de opname afkomstig uit de collectie van de Universiteit Gent of van het Meertens Instituut?

### **3.4.2 Locatiemetadata**

#### **Plaats van opname**

Hier kun je filteren door namen van gemeentes in te geven (bv. 'Langemark' of 'Gent'). Je kunt ook meerdere namen opgeven, gescheiden door een puntkomma ('Langemark; Gent').

#### **Provincie – Land**

Je kunt ook filteren op

- provincie (West-Vlaanderen, Oost-Vlaanderen, Antwerpen, Vlaams-Brabant, Be-Limburg, Nl-Limburg, Noord-Brabant, Zeeland, Frans-Vlaanderen\*) of
- land (België/Nederland/Frankrijk).

\* Dit is geen officiële provincie; we verwijzen ermee naar het deel van het Franse Noorderdepartement dat vroeger deel uitmaakte van het Graafschap Vlaanderen.

### **Dialectindeling 'ruw' en 'fijn'**

Provincies en landen zijn administratieve eenheden, die niet noodzakelijk ook dialectologische eenheden vormen. Zo is vrij algemeen aanvaard onder taalkundigen dat het Oost-Vlaamse dialectgebied en de provincie Oost-Vlaanderen slechts gedeeltelijk samenvallen. Om de gebruiker ook toe te laten op dialectgebied te zoeken, kenden we iedere opnamelocatie een dialectregio (bvb. Limburgs / Vlaams / ...) en -subregio (bvb. Oost-Limburgs / West-Zeeuws-Vlaams / ...) toe.

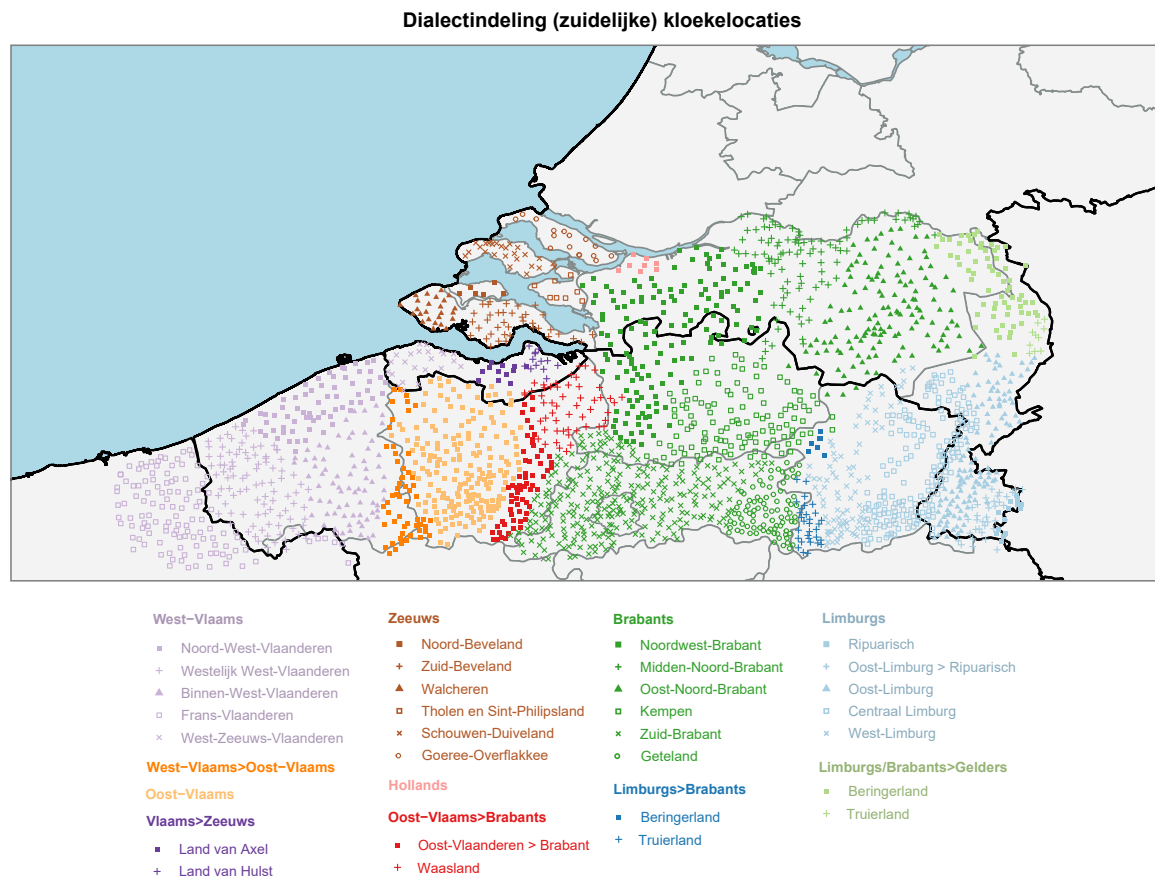
### ***Gebiedsindeling***

1. Limburgs
  - 1.1 Ripuarisch gebied (WLD 10)
  - 1.2 Overgangsgebied Oost-Limburg-Ripuarisch (WLD 20)
  - 1.3 Oost-Limburg (WLD 30)
  - 1.4 Centraal Limburg (WLD 40)
  - 1.5 West-Limburg (WLD 51, 52, 54) – inclusief Budels (WBD 90)
2. Overgang Limburgs > Brabants
  - 2.1 Beringerland (WLD 53) > overgang Limburgs-Brabants
  - 2.2 Truierland (Zuid-Brabants-West-Limburgs overgangsgebied) – (WLD 60 + Brabantse plaatsen Rummen (P107a) en Grazen (P108))
3. Overgang Brabants/Limburgs > Gelders (voorlopig enkel plaatsen in provincies Limburg en Brabant)
  - 3.1 Kleverland/Cuijk (WBD 80 Cuijk + WLD 100 Kleverland)
  - 3.2 Zuid-Gelders Limburg (WLD 90)
4. Brabants
  - 4.1 Noordwest-Brabant (WBD 10)
  - 4.2 Midden-Noord-Brabant (WBD 20)
  - 4.3 Oost-Noord-Brabant (WBD 30, inclusief Lommels: WLD 80)

- 4.4 Kempen (WBD 40, incl. WLD 72)
- 4.5 Zuid-Brabant (WBD 50), incl. dialecten van de Denderstreek (provincie Oost-Vlaanderen) waar ook Brabantse dialecten worden gesproken (cf. Taeldeman 2005 & Belemans e.a. 2000, p. 31)
- 4.6 Geteland (WBD 60 > incl. WLD73)
- 5. Hollands (voorlopig enkel plaatsen in provincie Brabant)
  - 5.1 Westhoek (WBD 70)
- 6. Overgang Brabants/Oost-Vlaams
  - 6.1 Oost-Vlaanderen > Brabant
  - 6.2 Waasland (cf. Taeldeman 2005) incl. Zeeuwse plekken Koewacht t/m Nieuw-Namen (cf. Taeldeman 1979) + Antwerpse dorpen Zwijndrecht, Burcht en St-Anneke Belemans e.a. 2000, p. 31)
- 7. Oost-Vlaams
  - 7.1 Oost-Vlaanderen (cf. Taeldeman 2005) > incl. Zeeuwse plekken Zandstraat, Overslag, Zuiddorpe en Westdorpe (cf. Taeldeman 1979), alsook Sas van Gent (ondanks eilandkarakter) > ‘Echte’ Oost-Vlaams
- 8. Overgang West-Vlaams–Oost-Vlaams
  - 8.1 West-Vlaanderen > Oost-Vlaanderen (cf. Taeldeman 2005)
- 9. West-Vlaams
  - 9.1 Noord-West-Vlaanderen (cf. Devos en Vandekerckhove 2005)
  - 9.2 Westelijk West-Vlaanderen (cf. Devos en Vandekerckhove 2005)
  - 9.3 Binnen-West-Vlaanderen (‘continentaal West-Vlaams’ bij Devos en Vandekerckhove 2005)
  - 9.4 Frans-Vlaanderen
  - 9.5 West-Zeeuws-Vlaanderen (cf. Taeldeman 1979), incl. Eede
- 10. Overgang Vlaams naar Zeeuws
  - 10.1 Land van Axel (sterkste afwijkingen van Oost-Vlaamse en Wase dialecten, grootste overeenkomsten met die van de Zeeuwse eilanden en West-Vlaanderen, cf. Taeldeman 1979) + Philippine (ondanks eilandkarakter)
  - 10.2 Land van Hulst (vrij sterk verbonden met de oostelijk Wase dialecten, cf. Taeldeman 1979, incl. Hulst zelf, ondanks eilandkarakter)
- 11. Zeeuws (Van Driel 2004)
  - 11.1 Noord-Beveland

- 11.2 Zuid-Beveland
- 11.3 Walcheren
- 11.4 Tholen en Sint-Philipsland
- 11.5 Schouwen-Duiveland
- 11.6 Goeree-Overflakkee





Figuur 8: Dialectindeling GCND

## *Verantwoording*

### *Limburgse en Brabantse subgebieden*

Voor de afbakening van de Limburgse en Brabantse subgebieden namen we de hoofdregio's van het Woordenboek van de Limburgse Dialecten en het Woordenboek van de Brabantse Dialecten als uitgangspunt. Die gebiedsindelingen bouwen op klankgrenzen (isoglossen), en worden beschreven en gemotiveerd in (Belemans e.a. 2000, p. 27–70); *Inleiding WBD deel III*) en (Kruijssen 2001, p. xviii–xli); *Inleiding WLD deel III*). De beschrijvingen bevatten ook oplijstingen van alle locaties/Klokecodes die tot een specifiek gebied worden gerekend; op die manier kon onze database voor de Limburgse en Brabantse dialecten relatief makkelijk met dialectregio-informatie worden verrijkt. Alleen wat een aantal grensdialecten betreft, moesten we nog een aantal knopen doorhakken, omdat die vaak door hun provinciale ligging in resp. het WBD/WLD (of zusterproject WVD) werden behandeld, maar eigenlijk bij een andere dialectfamilie aansluiten (cf. Belemans e.a. 2000, p. 31):

- *Plaatsen in de provincie Brabant waar Limburgs wordt gesproken*
  - De Brabantse plaatsen Rummen en Grazen behoren strikt genomen tot het Limburgse dialectgebied, meer specifiek het Truierlands, maar werden om praktische redenen in het WBD behandeld (WBD 60). Hier delen we deze plaatsen in bij het Limburgse (Truierlandse) dialectgebied.
  - De plaatsen Budel, Groot Schoot, Maarheeze en Soerendonk (WBD 90: Budels) vormen een vergelijkbaar geval: officieel behoren ze tot de provincie Noord-Brabant (vandaar ook behandeld in WBD), maar het dialect is er eigenlijk Limburgs, meer specifiek vertoont het gelijkenissen met de West-Limburgse dialecten. Om die reden rekenen we het Budels bij de West-Limburgse dialecten.
- *Plaatsen in de provincie Limburg waar Brabants wordt gesproken*
  - De Limburgse plaatsen Halen, Linkhout, Loksbergen, Meldert, Zelem en Zelk worden in het WLD afzonderlijk behandeld (WLD 73), omdat ze eigenlijk aansluiten bij de Brabantse ‘Getelandse’ dialecten; hier voegen we ze dan ook samen met de Brabantse Getelandse dialecten (WBD 60).
  - De WLD-gebieden 71 en 72 (resp. Noorderkempens en Zuiderkempens) sluiten aan bij de Kempense dialecten; we behandelen ze in onze database dan ook als delen van het Kempens dialectgebied.

- Het Lommels wordt in het WLD als een aparte regio behandeld (WLD 80) omdat het volgens de WBD-auteurs een Brabants dialect is (Kempenlands); hier rekenen we het dan ook bij de Brabantse dialecten (Oost-Noord-Brabant).
- *Plaatsen in de provincies Limburg en Noord-Brabant waar geen Limburgs of Brabants wordt gesproken*
  - Het Kleverlands wordt in het WLD als aparte regio behandeld (WLD 100) en het Cuijks in het WBD als afzonderlijke regio (WBD 80) omdat beide au fond Gelderse (meer specifiek Kleverlandse) dialecten zijn. Hier behandelen we die twee samen als ‘Kleverland/Cuijk’ en beschouwen we ze als overgangsdialecten van het Limburgs/Brabants naar het Gelders.
  - In de Brabantse Westhoek (WBD70) worden volgens de WBD-redacteuren Hollandse dialecten gesproken; die classificeren we hier dan ook als ‘Hollands’.
- *Plaatsen in de provincie Antwerpen waar geen Antwerps wordt gesproken*
  - Volgens Taeldeman (2005) sluiten de drie Antwerpse dorpen op de linkeroever van de Schelde – Zwijndrecht, Burcht en St.-Anneke (WBD13, Noordwest-Brabants) – aan bij de Waaslandse dialecten (die overgangsdialecten vormen tussen Oost-Vlaams en Brabants). Die plaatsen worden in het GCND dan ook tot de Waaslandse dialecten gerekend.
- *Plaatsen in de provincie Oost-Vlaanderen waar Brabants wordt gesproken*
  - In het oosten van de provincie Oost-Vlaanderen sluiten de dialecten uit de streek van Dendermonde volgens Taeldeman (2005) aan bij het Kleinbrabants (WBD51) en de dialecten uit de Denderstreek bij het Brabantse Pajottenlands (WBD52). We rekenen de dialecten uit de streek van Dendermonde en de Denderstreek dan ook tot de Zuid-Brabantse dialecten in het GCND.

*West-, Oost- en Zeeuws-Vlaamse dialecten*

Voor de afbakening van de West-, Oost- en Zeeuws-Vlaamse dialecten vertrokken we niet vanuit het Woordenboek van de Vlaamse Dialecten en dat van de Zeeuwse dialecten, omdat de insteek bij de gebiedsindelingen daar provincie- en streekaanduidingen waren (zoals ‘Veurne-Ambacht’, ‘Meetjesland’, ‘de streek van Zottegem’) – voor de herkenbaarheid bij de niet-taalkundig geschoolde gebruiker – en niet linguïstische grenzen. De WVD-indeling is daardoor weinig

compatibel met die van het WBD en WLD. Een op fonologische leest gestoelde indeling vinden we wel bij Devos en Vandekerckhove (2005, p. 29) voor de West-Vlaamse dialecten, Taeldeman (2005, p. 19) voor de Oost-Vlaamse dialecten, Taeldeman (1979, p. 172) voor Zeeuws-Vlaanderen en Driel (2004) voor de Zeeuwse dialecten. Voor onze dialectindeling baseren we ons dan ook op de indelingskaarten in die werken.

Voor het Frans-Vlaams (dat we als een subfamilie van het West-Vlaams behandelen) dient de landsgrens België-Frankrijk als uitgangspunt. De dialecten aan Franse zijde van die grens vertonen weliswaar heel veel overeenkomsten met de West-Vlaamse, maar doordat de dialecten aan Franse zijde veel minder beïnvloed werden door het Standaardnederlands (taalisolement vanaf 1713), vind je vandaag in de buurt van de staatsgrens ook verschillende taalgrenzen (Ryckeboer 2013). Bovendien onderscheiden de dialecten ten zuiden van de staatsgrens zich van de West-Vlaamse door intenser taalcontact met het Frans (Ryckeboer 2013; Ryckeboer 2004).

### ***Waarschuwing***

Dialectgebieden afbakenen is een moeilijke onderneming. Zoals de WBD-redactie destijds ook schreef: "Een taallandschap is, bepaald als het wordt door factoren van de meest uiteenlopende aard, een dusdanig gecompliceerde grootheid dat er geen enkel criterium te vinden is aan de hand waarvan men tot een volkomen bevredigende afbakening zou kunnen geraken" (Weijnen en Van Bakel 1967). De GCND-dialectafbakening moet dan ook heel uitdrukkelijk als een rudimentair hulpmiddel bij het zoeken en analyseren gezien worden, niet als de ultieme waarheid over de structuur van het zuidelijk-Nederlandse dialectlandschap. Iedere gebruiker kan indien gewenst andere indelingen maken via de locatie-zoekmodule.

### **3.4.3 Sprekermetadata**

Via de zoekmodule kun je op zoek naar de spraak van specifieke sprekers.

#### **Geslacht**

In de oorspronkelijke fiches over de dialectopnames wordt een onderscheid gemaakt tussen mannelijke (m) en vrouwelijke (v) sprekers. Het GCND neemt die indeling over (ook al zijn we er ons van bewust dat sprekers zich ook als non-binair kunnen identificeren).

#### **Geboortejaar**

In welk jaar is de spreker geboren?

#### **Geboorteplaats**

Waar is de spreker geboren? Wees je ervan bewust dat de geboorteplaats niet per se overeenstemt met de plaats waar de spreker is opgegroeid. De opgroeiplaats stemt doorgaans overeen met de 'plaats van opname'.

### **Mobiliteit woonplaats en werkplaats**

Traditioneel wordt er in de dialectologie van uitgegaan dat de beste dialectsprekers niet-mobiel zijn (cf. NORM-profiel, 'Non-mobile old rural male'). Ze hebben idealiter steeds op dezelfde plaats gewoond en gewerkt. In de GCND-collectie zijn er echter ook een aantal sprekers die niet aan dat profiel beantwoorden (maar die door de opnameleider desondanks toch als competente dialectsprekers werden beschouwd). Wil je die niet meenemen in je analyses, dan kun je filteren op:

- Mobiliteit woonplaats
  - nee = spreker woonde steeds op dezelfde plaats
  - ja, zelfde provincie = spreker is in de loop van zijn/haar leven verhuisd, maar wel binnen dezelfde provincie
  - ja, andere provincie = spreker is in de loop van zijn/haar leven verhuisd, over provinciegrenzen heen
- Mobiliteit werkplaats
  - nee = spreker werkte op zelfde plaats als woonplaats
  - ja, zelfde provincie = spreker werkte op andere plaats dan woonplaats, maar wel binnen dezelfde provincie
  - ja, andere provincie = spreker werkte in een andere provincie dan die waarin hij/zij woonde

Zodra op de infociche een extra woonplaats genoteerd werd, werd de spreker als 'mobiel' gecategoriseerd. Het kan echter ook om mobiliteit in de context van de legerdienst of de wereldoorlogen (vluchten/mobilisatie) gaan. Aangezien die informatie niet consistent werd genoteerd, hanteerden we bij de categorisatie het vrije ruwe criterium 'andere plaats genoteerd = mobiel'.

### **Herkomst moeder/vader**

Bij dialectverwerving speelt de thuistaal uiteraard ook een belangrijke rol. Spreekt je moeder of vader een ander dialect dan dat van de gemeente waarin je bent opgegroeid, dan kan dat een impact hebben op je eigen taalgebruik. Vandaar dat je in het GCND ook kunt filteren op herkomst van vader of moeder:

- zelfde plaats = zelfde gemeente als spreker
- zelfde provincie = andere gemeente dan spreker, wel in zelfde provincie
- andere provincie = gemeente in andere provincie

## 4 Corpusstatistieken

Bij de lancering van de eerste versie van het corpus op 24 oktober 2024 telde het GCND:

- 430 uur 52 minuten 13.799 seconden aan audio
- 650 opnames
- 639 verschillende plaatsen
- 1206 sprekers
- 306 niet-gecontroleerde transcripties met niet-gecontroleerde taalkundige verrijking
- 219 gecontroleerde transcripties met niet-gecontroleerde taalkundige verrijking
- 70 gecontroleerde transcripties met gecontroleerde taalkundige verrijking
- 55 gecontroleerde transcripties zonder taalkundige verrijking
- 4.771.074 tokens (omvang hele corpus)
- 50.111 geverifieerde zinnen (in GrETEL Treebank)
- 452.459 geverifieerde tokens (in GrETEL Treebank)

Deze cijfers zijn echter een momentopname. Dankzij vervolfinanciering (GCND+, zie 2.4) worden momenteel extra banden getranscribeerd (ook meer noordelijke dialecten uit de collectie van het Meertens Instituut) en worden meer banden grammaticaal geannoteerd en geverifieerd. De recentste tokenaantallen zijn in de corpusapplicatie terug te vinden.

## 5 Toegang en gebruiksrechten

Het Gesproken Corpus van de zuidelijk-Nederlandse Dialecten is alleen toegankelijk met een gebruikersnaam en wachtwoord. Gebruikers die studeren of werkzaam zijn aan een (Europese) universiteit, een hogeschool of een onderzoeksinstelling kunnen inloggen met de gebruikersnaam en het wachtwoord van hun organisatie. Gebruikers die niet verbonden zijn aan een academische instelling kunnen het corpus ook raadplegen, maar zij moeten daarvoor eerst een account aanvragen via [www.clarin.eu](http://www.clarin.eu).

Het GCND is beschikbaar gesteld voor niet-commercieel gebruik, uitsluitend voor wetenschappelijk onderzoek en educatieve doeleinden. Het is niet toegestaan het corpus, of delen daarvan, te gebruiken voor commerciële toepassingen zonder uitdrukkelijke schriftelijke toestemming van de beheerders van

het corpus. Onder commercieel gebruik wordt verstaan elke vorm van exploitatie waarbij de gebruiker direct of indirect financiële of commerciële voordelen behaalt. Dit omvat, maar is niet beperkt tot, het integreren van het corpus in commerciële producten of diensten, en het gebruik ervan voor commerciële data-analyse. Door toegang te krijgen tot dit corpus, gaat de gebruiker akkoord met de voorwaarden en verplicht zich deze niet te schenden. Overtreding van deze voorwaarden kan juridische consequenties hebben.

## **6 Zoekinterface**

Alle informatie over zoeken in het GCND vind je in de applicatie onder 'help'.

## **7 Transcriptieprotocol**

Het transcriptieprotocol kan via onderstaande link gedownload worden:

<https://www.gcnd.ugent.be/corpusannotatie/>

## 8 Protocol syntactische annotatie

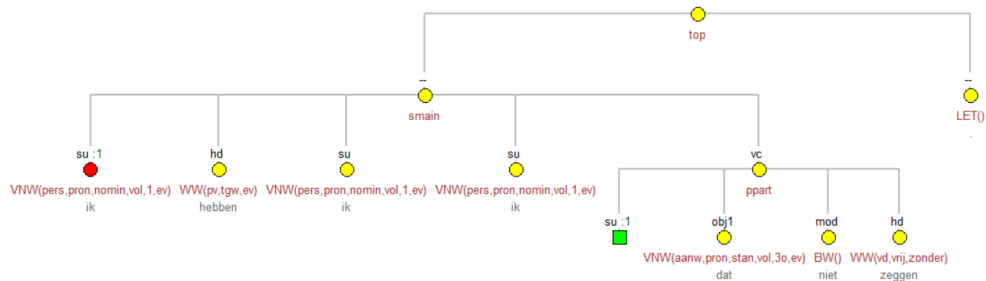
### 8.1 Subjectsverschijnselen

#### 8.1.1 Subjectsver(drie)dubbeling

(De Vogelaer en Devos 2008) onderscheiden verschillende types subjectsverdubbeling:

1. **Clitic doubling:** clitisch element + optioneel sterk pronomen (*werkt zij in Brussel, da zij in Brussel werkt*)
2. **Topic doubling:** pronominaal onderwerp op de eerste plaats, verdubbeling na de persoonsvorm (*Zij/ze werkt zij in Brussel*)
3. **Topic marking:** combinatie van een lexicaal element en een sterk pronomen (*Marie werkt zij in Brussel, dat zij Marie in Brussel werkt, An heeft zij dat niet gedaan*)
4. **Expletieven:** Er is nog een speciaal geval van subjectverdubbeling, waarbij een pronominaal element op de positie van de verdubbeling verschijnt dat niet verwijzend is; het verwijst immers niet naar dezelfde persoon als het onderwerp verwijst (*Marie werkt tet/tjij/hij ook*). (Haegeman, Greco en Phan 2017) noemen dit soort elementen ‘expletives’.

De verschillende types worden in het GCND op dezelfde manier behandeld, namelijk als aparte knopen op hetzelfde niveau in de dependentieboom met het dependentielabel *su* (‘subject’). Het eerste subject krijgt een index; de andere niet, 9.



Figuur 9: Subjectsverdubbeling

Morfologisch taggen (zie 3.3.2) we alle types als ‘nominatief’, ook al gaat het om een vorm die in het Standaardnederlands enkel als objectsvorm wordt gebruikt (zie verder).



**Opmerking:** Heel af en toe zijn er zinnen waarbij de verdubbelde subjecten als eenheid fungeren:

(3) Als **we wij** en mijn vader in de kelder zaten ...

In het GCND wordt *we wij* in dit geval als een **du** (discourse unit) geanalyseerd die nevenschikt is aan de nominale constituent mijn vader.

### 8.1.2 Subject in objectsvorm

Bij gevallen zoals de volgende krijgen de vormen *hem* en *ons* het dependentielabel **su** ondanks de (in de standaardtaal) objectsvorm.

- (4) a. omdat **hem** peinsde dat dat zijn kindje was.  
b. waar zitten wij **ons** hier nu weer

### 8.1.3 Presentatief *het*.

Deze constructie wordt noch in de Lassy-handleiding noch in de CGN-beschrijving besproken. In het GCND wordt dit soort *het* zoals presentatief *er* behandeld, en krijgt dus het dependentielabel **mod**.

(5) **Het** staat daar geen stoel.

## 8.2 Uitbreidingen van de zin: TAG en SAT

### 8.2.1 Linker zinsperiferie

In het Standaardnederlands staat de persoonsvorm in hoofdzinnen doorgaans op de tweede zinsplaats. Zetten we een bepaling op de eerste zinsplaats, dan keren we de elementaire volgorde ‘onderwerp – persoonsvorm’ in de regel om – dan spreken we van inversie – en blijft de pv op de tweede zinsplaats. Soms wordt van dit principe echter afgeweken, en dan spreken we van V3-hoofdzinnen.

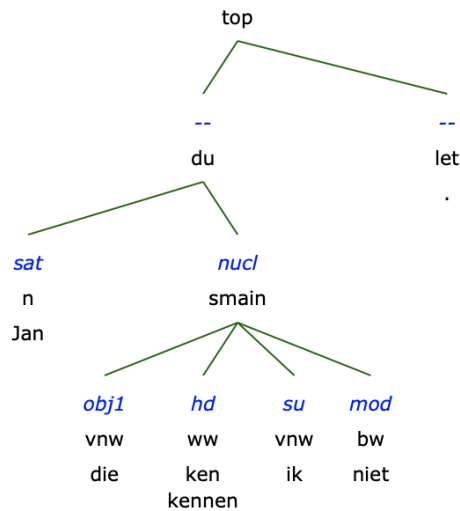
Het is belangrijk om verschillende types V3-zinnen uit elkaar te houden. De verschillende types worden namelijk anders in Alpino geannoteerd.

1. Aanloopconstructie (analyse: **sat**)
2. Hangende topics (analyse: **tag**)
3. Tussenwerpsels, aansporingen (analyse: **tag**)
4. Inversieloze V3-zinnen (analyse: **tag**)

(a) **Aanloopconstructie (=Left Dislocation / Linksdislocatie).** Bij aanloopconstructies staat er een naamwoordgroep in de eerste positie (de aanloop) en aanwijzend voornaamwoord/voegwoord (*die, dat, daar*) – een zog. resumptief – in de tweede positie. Daarop volgt de persoonsvorm.

De naamwoordgroep op de eerste plaats zien we als aanloopzinsdeel (dependentielabel **sat**) dat samen met de kernzin een discourse unit vormt. De kernzin krijgt het dependentielabel **nucl**, Fig. 10.

Jan die ken ik niet .



Figuur 10: Aanloopconstructie/linksdislocatie

(b) **Hanging Topic / Hangend Topic / Nominativus Pendens.** Op het eerste gezicht lijken deze constructies met hanging topics misschien op een linksdislocatie/aanloopconstructie (zie boven), maar er zijn cruciale verschillen:

- De intonatorische pauze is hier niet optioneel.
- Er staat steeds een naamwoordgroep in de eerste positie, die later in de zin door een persoonlijk voornaamwoord (hij, het, zij, hem, haar) wordt opgenomen, dus niet door een aanwijzend voornaamwoord/voegwoord (die, dat, daar) zoals bij Left Dislocation. Dit voornaamwoord hoeft niet op de tweede positie te staan, het kan op eender welke plaats in de zin staan waar nominale constituenten kunnen staan.

- Syntactisch gezien staan hanging topics buiten de kernzin. Dat kun je bijvoorbeeld zien aan het feit dat in een zin als *Iedere boer, hij verwent zijn vrouw* er geen coreferentie tussen *iedere boer* en *zijn* bestaat (de interpretatie is veeleer dat ‘zijn’ op een andere referent verwijst die eerder in het discours geïntroduceerd werd.<sup>9</sup>)

De hanging topics worden in het GCND als **tag** geanalyseerd, de kernzin als **nucl**.

(c) **Tussenwerpsels en aansporingen.** Een werkwoord belandt ook vaak op de derde zinsplaats wanneer een spreker een tussenwerpsel, aansporing of een gereduceerde zin voorop zet. Die elementen beschouwen we net als linksdislocaties niet als deel van de kernzin. Ze staan wel op een andere manier dan die linksdislocaties buiten de zin. Net als hanging topics krijgt dit element het dependentielabel **tag** (cf. ook inleidende zin bij directe redes, zie verder in dit overzicht).

- (6) a. **zo**, dat was plezant.  
 b. **natuurlijk**, moeilijk is het niet.  
 c. **kom**, ik ga er maar vandoor.  
 d. **jongens**, ik vertrek nu.

(d) **Inversieloos V-later-dan-2 / V>2 / Noninverted V3.** De volgende voorbeelden vormen een bijzondere vorm van hanging topics, en komen in de regel niet in de standaardtaal voor. Je hoort ze wel vaak in de dialecten en soms ook in andere gesproken registers.

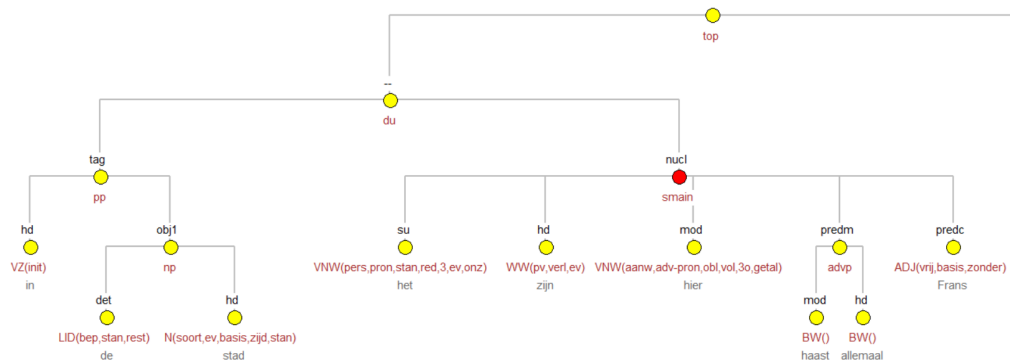
- (7) a. Toen ik thuiskwam, de elektriek was uitgevallen  
 b. In de stad, het was hier haast allemaal Frans  
 c. De boer, als hij gepresseerd is, hè, hij gaat voort in het veld

In dit soort zinnen staat er meestal een bijwoordelijke bepaling voorop, soms meerdere, of een nominale constituent. Soms vind je er ook nog eens een hanging topic tussen (cf. voorbeeldzin over ‘de boer’). Daarna volgt een hoofdzin met het onderwerp voorafgaand aan de persoonsvorm.

Net als hanging topics en tussenwerpsels in de aanloop worden hanging topics in het GCND als **tag** geanalyseerd, en de kernzin als **nucl**.

---

<sup>9</sup>Dit is ook een belangrijk verschil met de aanloopconstructie/linksdislocatie, waar coreferentie wél mogelijk is: *Iedere boer, die verwent zijn vrouw* wordt zo geïnterpreteerd dat het om de vrouw van iedere boer gaat, niet van iemand anders.



Figuur 11: Inversieloos V>2

Voor de laatste drie types constructies, dus

- Hanging topics
- Tussenwerpsels
- Inversieloze V-later-zinnen

wordt er in het GCND dus voor eenzelfde analyse gekozen: een discourse unit **du** met als constituenten **tag** en **nucl**.

Wij baseren ons bij deze analyse op de Lassy-handleiding:

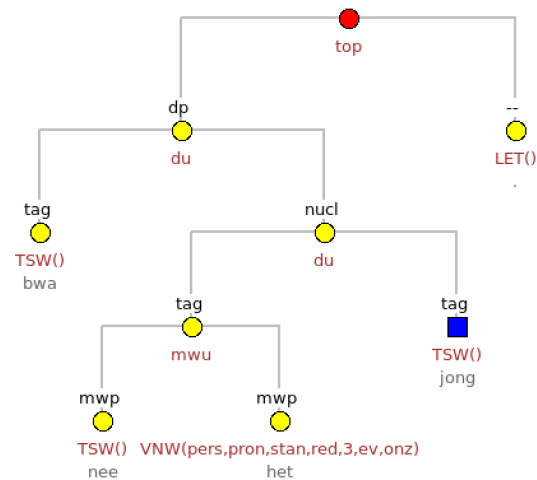
“Het functielabel **tag** maakt in die gevallen duidelijk dat de elementen niet hun ‘normale’ rol [in de zin] vervullen [zoals] bijwoord (**mod** binnen de zin) ...” (p. 188)

Dit geldt voor alle drie types constructie:

- Een hanging topic is geen zinsdeel van de hoofdzin (vgl. [*Iedere boer*]<sub>i</sub>, *hij verwent zijn*<sub>\*i/j</sub> *vrouw*).
- Bij tussenwerpsels e.d. zoals *natuurlijk* vervult *natuurlijk* niet de functie van bijvoeglijk naamwoord of bijwoord in de kernzin.
- Een vooropgespelde bepaling in een inversieloze V3-constructie is geen zinsdeel van de hoofdzin (bijv. een zin als *Toen ik thuis kwam, de elektriciteit was uitgevallen* kan alleen betekenen dat de elektriciteit al voor aankomst uitgevallen was, de temporale bijzin modificeert dus niet (het tijdstip van) *uitgevallen*. Dit is anders bij een gewone V2-zin met inversie: *Toen ik thuis kwam, was de elektriciteit uitgevallen* kan zowel betekenen dat de elektriciteit bij aankomst al uitgevallen was, maar ook dat ze net op het moment van aankomst uitviel. Voor meer argumenten zie Greco & Haegeman (2020)).

(e) **Ingebedde dislocaties.** In navolging van het CGN (p. 70–71) analyseert het GCND zinnen zoals *Wat vindt u der eigenlijk van **dat zulke zinnen** dat die zo geanalyseerd worden?* als ingebedde links-dislocaties. Het deel *dat zulke zinnen* is dan een aanloop (dependentielabel **sat**) op de voegwoordzin *dat die zo geanalyseerd worden*.

(f) **ja/nee het/ik/...** In veel dialecten is het elliptische antwoord op een *ja/nee*-vraag “ja het”/“nee het” of “ja ik/wij/zij” etc. In het GCND wordt de combinatie van *ja/nee* + vnw als een multi-word-unit (categorielabel **mwu**) geanalyseerd met als dependentielabel **tag**. De hele constructie wordt een **dp** (discourse part).



Figuur 12: Dependentieboom van *Bwa nee het jong.*

### 8.2.2 V2-bijzinnen – pseudodirecte rede

Na werkwoorden zoals *denken* volgt in spreektaal vaak een V2-bijzin (een bijzin met het werkwoord op de tweede zinsplaats) i.p.v. een bijzin met een onderschikkend voegwoord en bijzinsvolgorde. Dit soort zinnen – die ook wel eens pseudodirecte redes worden genoemd – worden geannoteerd zoals directe redes (zie p.190 in de Lassy-manual), namelijk als een combinatie van een **tag** en een **nucl**. Bij een zin zoals *Hij zei hij weet het niet* krijgt de inleidende matrixzin (*hij zei*) het categorielabel **smain** en het dependentielabel **tag**. De V2-bijzin (*hij weet het niet*) krijgt als dependentielabel **nucl** en als categorielabel **smain** of – bij werkwoordsinitiële zinnen (*Hij zei, is dat zo?*) – **sv1**.

### 8.2.3 Intercalaties/parentheses/interpositio's

Intercalaties zijn tussenzinnen die syntactisch buiten de kernzin lijken te staan. In de schrijftaal worden dergelijke tussenzinnen vaak met gedachtestreepjes aangeduid. Over dergelijke tussenzinnen wordt noch in de LASSY-manual noch in het CGN-manual iets gezegd. De LASSY-manual heeft het wel over ‘materiaal tussen haakjes’ (of gedachtestreepjes), typerend voor schrijftaal (p. 156), en stelt daarover:

In geval het materiaal tussen haakjes een woordgroep vormt, die ook weggelaten kan worden zonder de syntactische structuur van de hele zin te veranderen, dan wordt het tussen haakjes geplaatste zinsdeel meestal als MOD geannoteerd.

In het GCND krijgen parentheses het dependentielabel **tag**, en zijn op hetzelfde niveau als de hoofdzin ondergebracht. Het niveau waarop intercalaties in de zin ingebed zijn en de functie die ze hebben zijn zeker voor discussie vatbaar: Sommige lijken een soort nabepaling bij een NC (cf. voorbeelden b en c); bij andere staat de intercalatie precies wat lossier (cf. voorbeeld a).

- (8) a. hij bestelde toen – je weet hoe hij is – een Duvel.
- b. en hier in de hof – ik heb hier een grote hof – stonden bonen en prei.
- c. mijn kat – ze is zwart met witte strepen – had toen een muis gevangen.

Het is hierbij belangrijk te onthouden dat een syntactische corpusannotatie geen linguïstische analyse is, maar de functie heeft het vinden van data te vergemakkelijken. De precieze analyse van dit soort zinnen is dus aan de GCND-gebruiker overgelaten. Het dependentielabel **tag** is bedoeld om aan te geven dat ze de syntactische structuur van de zin waarin ze voorkomen niet beïnvloeden.

**Let op!!!** Verwar een intercalatie niet met een bijstelling. Een bijstelling of appositie is “elke naamwoordelijke constituent die als aanvullende informatie achter een andere naamwoordelijke constituent geplaatst wordt” (ANS). Die nominale constituent kan van verbale origine zijn en kan allerlei zinsdelen incorporeren.

- (9) a. we werden gezegend met een kindje een zoon de uh achttiende oktober van zeventwintig .
- b. hoofdstuk vier
- c. de maand maart
- d. de stad Groningen
- e. de afstand Mechelen-Brussel

Deze bijstellingen krijgen geheel volgens de Lassy-aanpak (p. 132) het dependentielabel **app**.

## 8.3 Complementizer-fenomenen

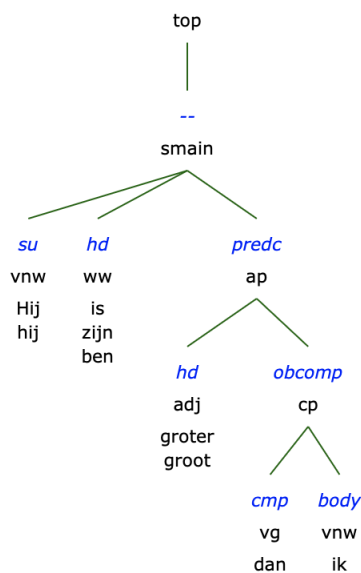
### 8.3.1 Afwijkende comparatieve voegwoorden

In de GCND-opnames vinden we volgende afwijkende comparatieve voegwoorden:

- (10) a. Hij is groter of/als ik.  
b. Dat is niet meer gelijk als vroeger.  
c. om te zeggen gelijk of dat het is had ik er niet moeten bijkomen het was buiten.

Indien voegwoorden *of* en/of *als* gebruikt worden in plaats van het standaardtalige *dan*, wordt de annotatie van voegwoord *dan toegepast*, namelijk dependentielabel **cmp** (complementizer) en categorielabel **vg** (voegwoord). Het voegwoord vormt samen met het element waarmee vergeleken wordt het vergelijkingscomplement (**obcomp**) – hier gedemonstreerd voor standaardtalig *dan*.

Hij is groter dan ik



Figuur 13: Comparatieve voegwoorden – dialectaal *of* / *als* worden geanalyseerd zoals standaardtalig *dan*.

*Gelijk als/gelijk of dat* worden behandeld als multi-word unit, waarbij dit in zijn

geheel als complementizer fungeert, ze krijgen dus samen het categorielabel **mwu** en het dependentielabel **cmp**. De aparte onderdelen (*gelijk | als | of*) hebben als dependentielabel **mwp** (part of multi-word unit).

Dit maakt het mogelijk om andere instanties van *gelijk* (als adjectief e.d.) van het voegwoordgebruik te onderscheiden. Bij gebruik van enkel *gelijk* (zonder *als*) doet Alpino dit sowieso (bvb. in *dat is niet meer gelijk vroeger*.)

### 8.3.2 Directe rede ingeleid door *van*

Directe redes analyseert Alpino doorgaans als een combinatie van een **tag** (=de matrixzin) en een **nucl** (= de directe rede). Een speciaal geval – typerend voor spreektaal – heb je wanneer de directe rede ingeleid wordt door *van*:

(11) Die vroeg aan mij van is die dan getrouwd?

**Let op:** De aanpak is hier dus wat anders dan bij ‘gewone’ directe redes. Directe redes met *van* worden dus als argument bij een matrixwerkwoord beschouwd, terwijl directe rede zonder *van* als **nucl** los van de inleidende zin (tag) worden geanalyseerd. Het GCND volgt in deze beslissing de richtlijnen van Lassy (p. 78) en het CGN (p. 26)<sup>10</sup>

### 8.3.3 Expletief *dat*

In gesproken Nederlands, en in de dialecten, kunnen vier types expletief *dat* onderscheiden worden.

Type 1: na onderschikkend voegwoord

- (12) a. Ik weet niet of **dat** hij komt.  
b. Om het te zeggen gelijk of **dat** het is: ...  
c. ik was getrouwd sinds **dat** hij nog bij het leger was

Type 2: na vraagwoord

- (13) a. Ik weet niet wie **dat** er komt.  
b. we gaan weer moeten de tijd afwachten wat **dat** er allemaal gaat voorvallen

---

<sup>10</sup>“We beschouwen *van* in het voorbeeld (...) als complementeerder ter inleiding van een complementszin in ‘directe rede’ (categorielabel **sv1** of **smain**). De bijzin inclusief *van* krijgt de categorie **svan**.”



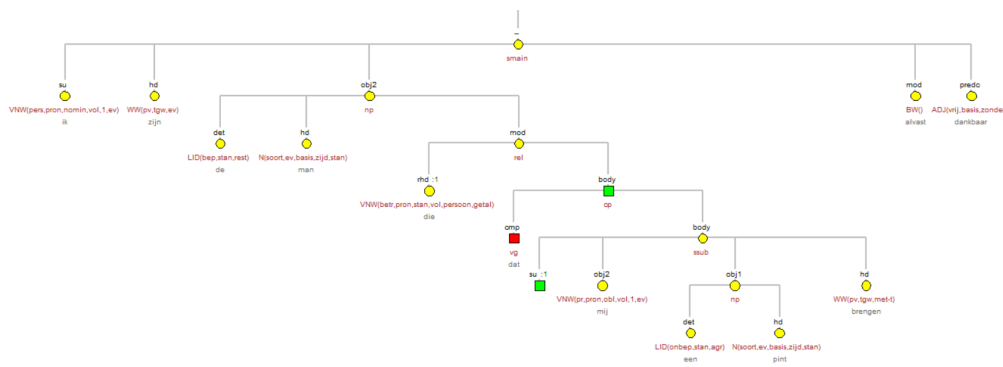
Type 3: na betrekkelijk voornaamwoord

- (14) De mens die dat jou moet helpen, zal vloeken.

Type 4: na vraagwoord + *of* (zeldzaam in Vlaanderen, cf. Lassy-handleiding)

- (15) Zijn er meer mogelijkheden dan wat of **dat** je nu hebt?

Types 1, 2 en 4 worden behandeld in de Lassy-handleiding bij ‘complexe voegwoorden’ (p. 86 en verder). Voor type 3 wordt er in het GCND de analyse voor de vraagwoorden (type 2) toegepast (gezien de formele en interpretatieve overeenkomsten tussen vraagvorming en relativisatie): Het betrekkelijk voornaamwoord krijgt het dependentielabel **rhd** (hoofd van de relatiefzin), de rest van de betrekkelijke bijzin is de **body**, waarbij de body het categorielabel **cp** heeft (= frase ingeleid door een onderschikkend voegwoord), en zelf bestaat uit een complementizer (**cmp|vg**) en een body (**body|ssub**).



Figuur 14: Dependentieboom van *De mens die dat jou moet helpen, zal vloeken.*

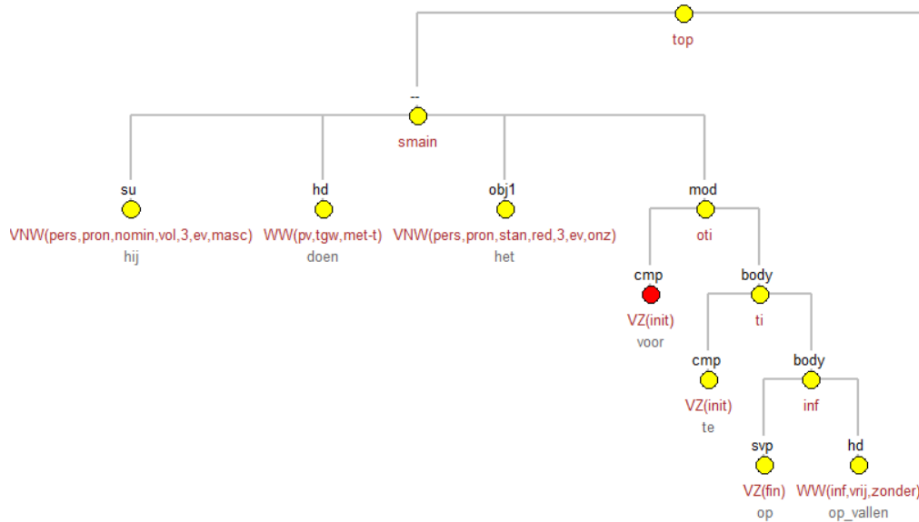
### 8.3.4 Beknopte bijzinnen ingeleid door *voor* of *van* i.p.v. *om*

In het Standaardnederlands worden beknopte bijzinnen ofwel ingeleid door het voorzetsel *om*, ofwel door een nulelement. *Om* is volgens Vandeweghe (1971) verplicht in finale bijzinnen, in resultatieve bijzinnen en in conditionele bijzinnen. Wanneer de beknopte bijzin fungeert als eigenlijk subject, als direct object of als nabepaling bij een substantief, is *om* facultatief. In niet-standaardtalige registers in Vlaanderen leiden de voorzetsels *voor* en *van* echter ook vaak beknopte bijzinnen in.

- (16) a. en als ze kwamen **voor** een papiertje te kiezen  
 b. Jan verdient niet **van** gestraft te worden.

Veralgemeend kunnen we stellen dat *voor* te horen is in contexten waar we in de standaardtaal het verplichte *om* verwachten (cfr. 16a), terwijl *van* varieert met het facultatieve *om* (cfr. 16b). In het GCND worden *van* en *voor* in dit soort zinnen op exact dezelfde manier als *om* geanalyseerd (cf. CGN p. 30), namelijk met het dependentielabel **cmp** en POS-tag **vz(init)**. In de boom maakt het deel uit van een **oti** (*om te*-infinitief-groep).

Wanneer het optionele *om* (of *van*) niet wordt uitgesproken, wordt de *te*-infinitief volgens Lassy-conventies als een **ti** (*te*-infinitief-groep) geannoteerd.



Figuur 15: Dependentieboom van *Hij doet het voor op te vallen*.

### 8.3.5 Afhankelijke *ja/nee*-vragen ingeleid door *als* i.p.v. *of*

In het Nederlands en Belgisch Standaardnederlands worden afhankelijke *ja/nee*-vraagzinnen ingeleid door het voegwoord *of*. In verschillende dialecten – vooral in West- en Oost-Vlaanderen, in Friesland, en ook in het Surinaams-Nederlands – is echter het voegwoord *als* gangbaar:

(17) Ik weet niet **als** hij komt.

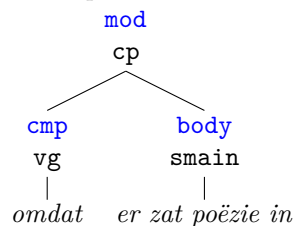
Deze *als* heeft in het GCND dezelfde annotatie als het voegwoord *of* (dependentielabel = **cmp** (complementizer), categorielabel = **vg** (onderschikkend voegwoord)).

### 8.3.6 Ingeleide bijzin met hoofdzinsvolgorde (V2-bijzin)

Af en toe wordt er na een onderschikkend voegwoord toch hoofdzinsvolgorde gebruikt.

(18) Die rol heb ik heel graag gespeeld omdat er zat poëzie in.

De V2-bijzin wordt in het GCND *niet* zoals een hoofdzin met  $V > 2$ ) behandeld, maar als een zin **smain** die het complement is van *omdat*.



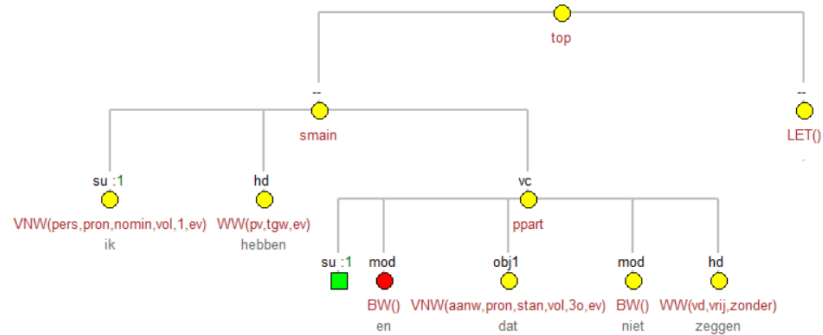
## 8.4 Negatieverschijnselen

### 8.4.1 Meervoudige negatie(markeerders)

In zinnen zoals die in (19) hebben de verschillende negatoren elk afzonderlijk een dependentielabel en plaats in de dependentiestructuur van de zin in het GCND.

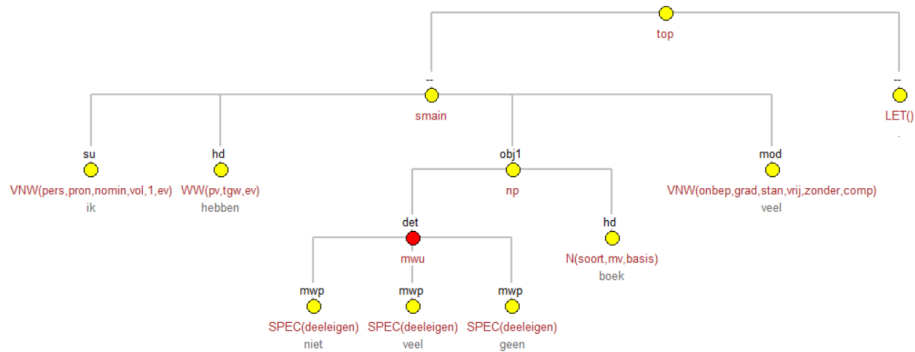
- (19) a. Ik **en** heb dat **niet** gezegd.  
b. Ik heb **niemand niet** gezien.  
c. Ik heb **niets niet** gedaan.  
d. Ik heb dat **nooit niet** gedaan.  
e. Daar zijn **nooit geen** rozen.  
f. Ik heb **geen** boeken **niet meer**.  
g. Er zijn er **niet** veel **niet meer**.  
h. Ik heb **niet veel geen** boeken **meer**.

Het (gewoonlijk) pre-finite element *en* krijgt, net als standaardtalig *niet* het dependentielabel **mod** (categorielabel/POS-tag = **bw()**).



Figuur 16: Dependentieboom van *Ik en heb dat niet gezegd*.

In zin (19h) wordt *niet veel geen* als een multi-word unit beschouwd, die in zijn geheel als determinator functioneert (cf. boom hieronder).



Figuur 17: Dependentieboom van *Ik heb niet veel geen boeken meer*.

#### 8.4.2 Adjectieven die met *geen* gecombineerd worden

In voorbeelden zoals (20) worden *waar* en *erg* in het GCND als hoofd van een substantief-frase (**np**) geïnterpreteerd omdat ze geïntroduceerd worden door een determineerder (**det|vnw**) *geen*.

- (20) a. Het is toch **geen waar** zeker?  
 b. Dat is **geen erg**.

Het GCND beschouwt ze dus als genominaliseerde adjectieven, het hoofd heeft daarom als POS-tag niet *nw*, maar *adj*: (*hd|adj(nom,basis,zonder,zonder-n)*).

### 8.4.3 *Doen*-replieken (met en zonder negatie).

In West- en Zuid-Oost-Vlaanderen hoor je vaak korte repliekzinnnetjes samengesteld uit het plaatsvervangende werkwoord *doen* en een persoonlijk voornaamwoord. Vaak komt er het oude ontkenningsspartikel *en* bij, vooral als een voorgaande bevestigende mededeling ontkent wordt (Ryckeboer 1986; Van Craenenbroeck 2004). Soms zijn replieken ook vragend – om verwondering uit te drukken – en dan hebben ze dezelfde vorm (bevestigend of ontkenkend) als de bewering die eraan vooraf gaat.

- (21) a. A: Jij doet dat toch hé?  
 B: ik **en doe!**
- b. A: Hij heeft dat toch niet gedaan? | het regent toch niet hé?  
 B: ja hij **doet**. | het **doet** het **doet** (klinkt als ‘toetoet’)
- c. A: Ik heb gisteren de brandweer gebeld  
 B: **Doe** je? (om verwondering uit te drukken, betekent zoveel als ‘echt?’)
- d. A: Die kinders mogen niets thuis.  
 B: En **doen** ze? (om verwondering uit te drukken, betekent zoveel als ‘echt?’)

Deze repliekzinnnetjes worden in het GCND als aparte zinnen geparset. Intern onderscheiden we *subj|vnw* + eventueel *mod|bw* (negatiepartikel) + *hd|ww*.

## 8.5 Andere

### 8.5.1 *en zo / of zo / en al / en alles / maar ja / en si en la*

*En zo* e.d. verschijnen soms op zinsniveau, NP-niveau, op het niveau van een infinitief of op conjunctieniveau:

- (22) a. Ik ga m’n eigen bedrijfje beginnen **of zo**.  
 b. Je moet dan wel wat regen **en zo** op de koop toe nemen.  
 c. Die zouden trouwen **of zo**.  
 d. Die was pas getrouwd of die zou gaan trouwen **of zo**.

In het GCND worden *en zo* en consoorten op dezelfde manier geanalyseerd als in het CGN (cf. (Hoekstra e.a. 2002, p. 64)): als een multi-word unit (**mwu**) die als modificeerder fungeert (**mod**).

### 8.5.2 Woordherhaling voor klemtoon

Naar analogie met o.a. subjectverdubbeling verschijnt het herhaalde woord twee keer in de boom, met hetzelfde dependentielabel.

(23) we hebben **niets** niets te kort op geen gebied.

### 8.5.3 Spiegelconstructies

We spreken van een spiegelconstructie (cf. CGN p. 69–70 – ook voor literatuur-referenties) als:

- de spreker een **zin begint in de standaardvolgorde** subject–persoonsvorm (soms wel met inversie als er een zinsmodificeerder voorop staat),
- vervolgens een **bepaling** toevoegt,
- en daarna het **bouwplan** van de zin als het ware **herziet** en de bepaling herinterpreteert als topic op de eerste zinspositie, en daarom nog eens de persoonsvorm en het onderwerp laat volgen.

(24) a. je gebruikt nog alsan diezelfde potees gebruik je.  
b. en da wordt alle zes maanden wordt dat vernieuwd.  
c. het is toen een eindje verder is het.  
d. de Bredabaan heb ik ook verschillende jaren heb ik ik daar gewerkt.

Om van een spiegelconstructie te spreken, moet het onderwerp exact hetzelfde zijn. In het GCND, net als in het CGN, heeft de dependentiestructuur van de zin dan (uitzonderlijk) twee hoofden en twee onderwerpen.

#### 8.5.4 Apokoinouconstructies

Een apokoinouconstructie is een constructie waarbij een woord of woordengroep tegelijkertijd deel uitmaakt, eerst als staart en dan als kop, van twee onafhankelijke constructies.

- (25) a. ik maakte een tots zeiden wij daartegen  
b. en dat was van de zaterdagochtend reden we wij mekaar de kop in

In het GCND wordt hier het categorielabel **apokoinou** gebruikt, en in het tweede zinsdeel is er een lege knoop die met het woord dat twee rollen heeft gecoördineerd is (=dezelfde index heeft).

#### 8.5.5 Opsomming van cijfers met betekenis ‘ongeveer’

##### Type 1: *dertig veertig vijftig boterhammen*

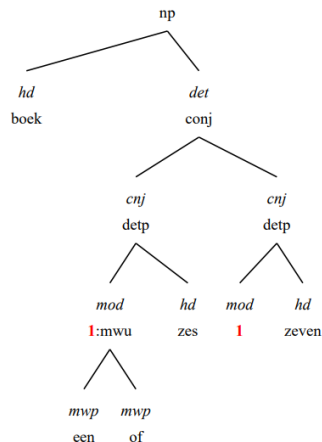
- (26) en we hebben dagen geweest dat we dertig veertig vijftig boterhammen hadden voor ons [ @alt hè hé ] .

Dit soort constructies analyseren we als een **asyndetische nevenschikking** van twee of meer determinatoren: dependentielabel **det**, categorielabel **conj**, met als conjuncten (**cnj**) meerdere telwoorden **tw** (cf. ook Lassy-aanpak bij constructies als *een man of twee drie*, zie ook hieronder).

##### Type 2: *Een jaar/man/... of vijftien zestien*

Volgende de Lassy-manual (p. 120) worden dit soort telwoorden als complexe determinator (**det|conj**) met als conjuncten twee **detp**'s (determinator-frase) behandeld.

(363) een boek of zes zeven



Figuur 18: Dependentieboom van *een boek of zes, zeven*.

Andere voorbeelden van complexe determinatoren zijn *van die bananen*, *een heleboel flessen*<sup>11</sup>, *een stuk of zes pintjes*.

### 8.5.6 Discontinue telwoorden *drie jaar en half*, *drie kilo een half*

Dit soort constructies wordt in het GCND als een multi-word-unit (categorielaabel **mwu**) geanalyseerd met als leden (dependentielabel **mwu** *drie*, *en*, *half* of *drie*, *een*, *half*).

### 8.5.7 Tijdsbepalingen met (*van*) *te*: *te zessen*, *van te zessen*

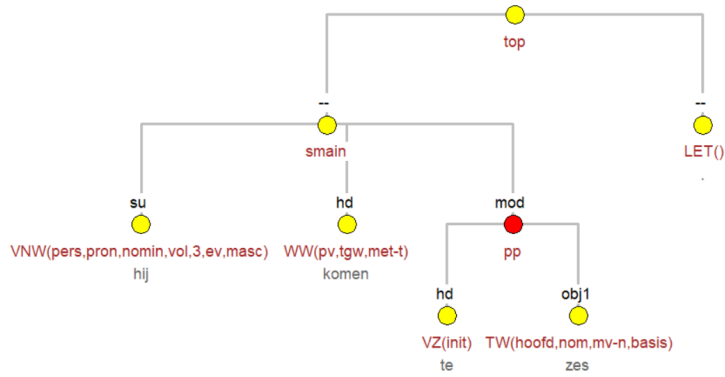
Dit soort tijdsbepalingen behandelen we als een **complexe voorzetselconstituent** die als **mod** functioneert en die bestaat uit een voorzetsel *van* en een voorzetselconstituent *te zessen*.

### 8.5.8 *De die*

*De die* zoals in (27) wordt in het GCND als **np** met *de* als determinator en *die* als hoofd geanalyseerd.

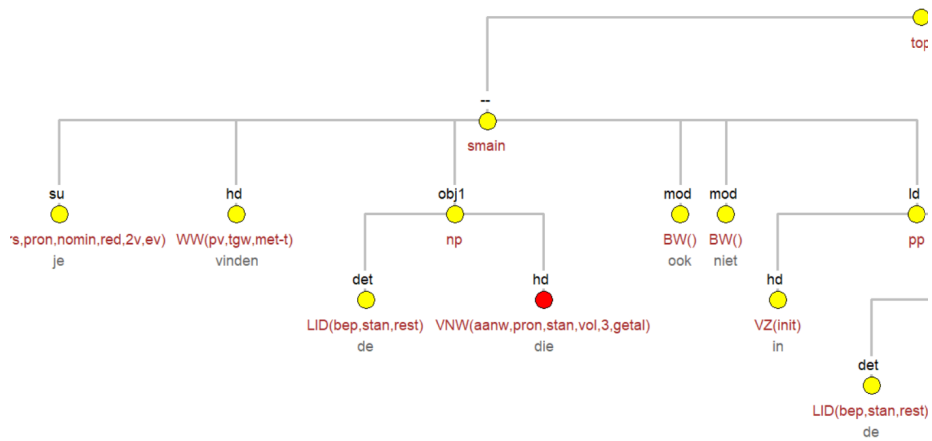
<sup>11</sup>Let wel: niet hetzelfde als *een kop koffie*, *een aantal mensen*!





Figuur 19: Dependentieboom van *hij komt van te zessen*.

- (27) a. je vindt **de die** ook niet in de gewone kruisweg.  
 b. en de mensen zeggen, wat lopen **de die** hier nu zo rond de kerk en ze komen weer op hetzelfde uit.

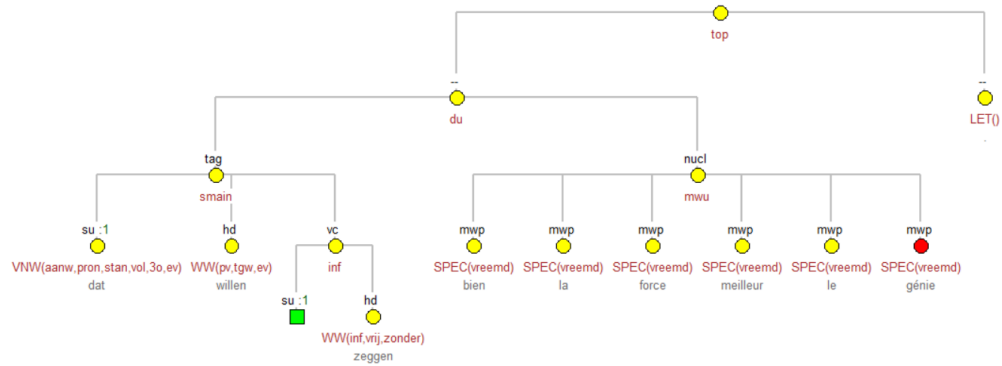


Figuur 20: Dependentieboom van (27a).

### 8.5.9 Code-switches naar het Frans/Engels/...

Wanneer het om een enkel woord gaat, is het in het GCND als leenwoord geannoteerd (POS-tag/categorieelabel *n*), meerdere woorden (bvb. in *dat wil zeggen bien*

*la force meilleur le génie.*) worden als multi-word unit geanalyseerd, waarbij elk woord de POS-tag **spec(vreemd)** krijgt. We analyseren zinnen of zinsdelen in het Frans/Engels/... dus niet intern.



Figuur 21: Dependentieboom van *dat wil zeggen bien la force meilleur le génie*.

### 8.5.10 Geluiden en klanknabootsingen

Uitgeschreven geluiden en klanknabootsingen analyseren we als tussenwerpsels.

- (28) a. en als je voeten zweetten **zwiep** zat je kleine teen erdoor.  
 b. zodus iedere keer dat hij sloeg hé dat was ... **djoef**.

Als ze geen zinsfunctie hebben (28a), staan ze zoals andere tussenwerpsels (**tsw**) bovenaan in de boom. Als ze wel een zinsfunctie hebben (28b) krijgen ze ook de POS-tag **tsw** maar de zinsfunctie moet wel benoemd worden. Dat is **predc** in het geval van (28b).

### 8.5.11 Exclamatieve infinitiefzinnen

Zinnen zoals (29)

- (29) a. wij maar **werken!**  
 b. en ik **zoeken** maar!

worden in het GCND als infinitieven (categorielabel **inf**) geanalyseerd, maar mét een overt subject (**su**). Deze infinitief wordt i.p.v. een **smain** gebruikt.

### 8.5.12 Circumpositie

In zinnen zoals

(30) want je moest **achter** niets **achter** gaan

wordt de circumpositie *achter* N *achter* als voorzetselphrase (pp) geanalyseerd met als **hd** *van*, *niets* als **obj1**, en *achter* als **hdf** (afsluitend element van een circumpositie).

## Referenties

- Barbiers, Sjef, Hans Bennis, Gunther De Vogelaer, Johan van der Auwera e.a. (2008). *Syntactische Atlas van de Nederlandse Dialecten. Deel II*. Amsterdam University Press.
- Barbiers, Sjef, Hans Bennis, Gunther De Vogelaer, Magda Devos e.a. (2005). *Syntactische Atlas van de Nederlandse Dialecten. Deel I*. Amsterdam University Press.
- Barbiers, Sjef, Johan Van der Auwera e.a. (2005). *Syntactic atlas of the Dutch dialects*. Deel 2. Amsterdam University Press.
- Belemans, Rob e.a. (2000). „Woordenboek van de Brabantse dialecten. 3: Inleiding; klankgeografie van de Brabantse dialecten [Dictionary of the Brabantine Dialects. 3: Introduction, sound geography of the Brabantine dialects]”. In: *Assen: Koninklijke Van Gorcum*.
- Benincà, Paola en Cecilia Poletto (2005). „The ASIS enterprise: a view on the construction of a syntactic atlas for the Northern Italian dialects”. In: *Nordlyd* 34, p. 35–52.
- Blancquaert, Edgard (1948). „Na meer dan 25 jaar dialect-onderzoek op het terrein”. In: *Koninklijke Vlaamse Academie voor Taal- en Letterkunde III*, p. 5–62.
- Brandner, Ellen (2012). „Syntactic Microvariation”. In: *Language and Linguistics Compass* 6/2, p. 113–130.
- (2015). „Syntax des Alemannischen (SynAlm): Tiefenbohrungen in einer Dialektlandschaft”. In: *Projekte und Perspektiven*. Red. door Roland Kehrein, Alfred Lameli en Stefan Rabanus. Berlin, München, Boston: De Gruyter, p. 289–322. DOI: doi:10.1515/9783110363449-014. URL: <https://doi.org/10.1515/9783110363449-014>.
- De Vogelaer, Gunther en Magda Devos (2008). „On geographical adequacy, or: How many types of subject doubling in Dutch”. In: *Syntax and Semantics* 36, p. 251–276.
- Devos, Magda en Reinhild Vandekerckhove (2005). *West-Vlaams. Taal in stad en land Vol. 7*. Tiel: Lannoo.
- Driel, Lo Van (2004). *Zeeuws: De dialecten van Zeeuws-Vlaanderen tot Goeree-Overflakkee*. Den Haag: SDU. ISBN: 9012090172. URL: <https://lib.ugent.be/catalog/rug01:000905917>.
- Farasyn, Melissa e.a. (2022). „Challenges in tagging and parsing spoken dialects of Dutch”. In: *Journal of Historical Syntax* 6.4-11, p. 1–36.
- Fleischer, Jürg, Alexandra N. Lenz en Helmut Weiß (2015). „Syntax hessischer Dialekte (SyHD)”. In: *Projekte und Perspektiven*. Red. door Roland Kehrein, Alfred Lameli en Stefan Rabanus. Berlin, München, Boston: De Gruyter, p. 261–288. DOI: doi:10.1515/9783110363449-013. URL: <https://doi.org/10.1515/9783110363449-013>.
- Ghyselen, Anne-Sophie, Anne Breitbarth e.a. (2020). „Clearing the transcription hurdle in dialect corpus building: The corpus of southern Dutch dialects as case study”. In: *Frontiers in artificial intelligence* 3, p. 10.

- Ghyselen, Anne-Sophie, Chelsey Deklerck e.a. (2024). *Hoe betrouwbaar zijn manuele dialecttranscripties? Het GCND geëvalueerd*. Koninklijke Commissie voor Toponymie en Dialectologie; 14/10/2024.
- Ghyselen, Anne-Sophie, Jacques Van Keymeulen e.a. (2020). „Het transcriptieprotocol van het Gesproken Corpus van de Nederlandse Dialecten (GCND)”. In: *BULLETIN DE LA COMMISSION ROYALE DE TOPONYMIE & DIALECTOLOGIE (PRINTED)= HANDELINGEN VAN DE KONINKLIJKE COMMISSIE VOOR TOPONYMIE & DIALECTOLOGIE* 92, p. 83–115.
- Haegeman, Liliane, Ciro Greco en Trang Phan (2017). „Expletives and Speaker-Related Meaning”. In: *Order and Structure in Syntax II: Subjecthood and Argument Structure*. Red. door Michelle Sheehan en Laura Bailey. Berlin: Language Science Press, p. 69–93.
- Hoekstra, A.W. e.a. (2002). „Syntactische annotatie voor het Corpus Gesproken Nederlands (CGN)”. Dutch. In: *Nederlandse Taalkunde* 4. Nella 19 dec., p. 335–352.
- Kayne, Richard (1996). „Microparametric syntax: some introductory remarks”. In: *Microparametric syntax and dialect variation*. Red. door J. R. Black en V. Motapanyane. Amsterdam: Benjamins, p. ix–xviii.
- Kroch, Anthony (2020). *Penn Parsed Corpora of Historical English. LDC2020T16*. Philadelphia: Linguistic Data Consortium. <https://doi.org/10.35111/4hzx-5483>.
- Kruijssen, Joep (2001). *Woordenboek van de Limburgse Dialecten (WLD), deel III, Algemene Woordenschat, Inleiding*. Assen: Kon. van Gorcum.
- Lindstad, Arne Martinus e.a. (2009). „The Nordic Dialect Database: Mapping Microsyntactic Variation in the Scandinavian Languages”. In: *NEALT Proceedings Series. Volume 4*. Red. door Kristiina Jokinen en Eckhard Bick.
- Noord, Gertjan van e.a. (2013). „Large Scale Syntactic Annotation of Written Dutch: Lassy”. In: *Essential Speech and Language Technology for Dutch: Results by the STEVIN programme*. Red. door Peter Spyns en Jan Odijk. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 147–164. ISBN: 978-3-642-30910-6. DOI: 10.1007/978-3-642-30910-6\_9. URL: [https://doi.org/10.1007/978-3-642-30910-6\\_9](https://doi.org/10.1007/978-3-642-30910-6_9).
- Rensink, W. G. (dec 1962). „Dialecten op band”. In: *Taal en tongval* 14.4. (No summary available), p. 184–196. URL: [www.bntl.nl/bntl/publicatie/dialecten\\_op\\_band\\_w\\_g\\_rensink/remsink\\_w\\_g](http://www.bntl.nl/bntl/publicatie/dialecten_op_band_w_g_rensink/remsink_w_g).
- Ryckeboer, Hugo (1986). „Het hulpwerkwoord doen in replieken”. In: *Vruchten van z'n akker. Opstellen van (oud-)medewerkers en oud-studenten voor Prof. V.F. Vanacker hem aangeboden bij zijn afscheid van de Rijksuniversiteit Gent*. Red. door Magda Devos en Johan Taeldeman. Gent: Seminarie voor Nederlandse Taalkunde en Vlaamse Dialectologie, Rijksuniversiteit Gent, p. 321–337.
- (2004). *Frans-Vlaams*. Tiel: Lannoo.
- (2013). „A West Flemish dialect as a minority language in the north of France”. In: ... Red. door Frans Hinskens en Johan Taeldeman, p. 782–799.

- Taeldeman, Johan (1979). „Op fonologische verkenning in Zeeuws-Vlaanderen”. In: *Taal en Tongval* 31, p. 143–193.
- (2005). *Oost-Vlaams. Taal in stad en land Vol. ??* Tielt: Lannoo.
- Van Craenenbroeck, Jeroen (2004). „Ellipsis in Dutch dialects”. LOT dissertation 96. Proefschrift. Leiden.
- Van Craenenbroeck, Jeroen, Marjo van Koppen en Antal van den Bosch (2019). „A quantitative-theoretical analysis of syntactic microvariation: Word order in Dutch verb clusters”. In: *Language* 95, p. 333–370.
- Van Noord, Gertjan (2006). „At last parsing is now operational”. In: *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Conférences invitées*, p. 20–42. URL: <https://aclanthology.org/2006.jeptalnrecital-invite.2.pdf>.
- Vanacker, V. F. en Georges de Schutter (apr 1967). „Zuidnederlandse dialecten op de band”. In: *Taal en tongval* 19.1-2. Also published in *Gentse Bijdragen 1967*, pp. 85-101, p. 35–51. URL: [www.bntl.nl/bntl/publicatie/zuidnederlandse\\_dialekten\\_op\\_d/vanacker\\_valeer\\_frits](http://www.bntl.nl/bntl/publicatie/zuidnederlandse_dialekten_op_d/vanacker_valeer_frits).
- Vanacker, Valeer-Frits (1967). „Syntaktische Daten aus französisch-flämischen Tonbandaufnahmen”. In: *Verhandlungen des 2. Internationalen Dialektologenkongresses. Band II*, p. 844–855.
- Vandeweghe, Willy (1971). „Om en rond de (om) te-konstruktie”. In: *Studia germanica gandensia* 13.1. Hulde-aflevering Willem Pée, p. 37–61. URL: [https://www.bntl.nl/bntl/publicatie/om\\_en\\_rnd\\_de\\_om\\_te\\_-\\_konstrukt/vandeweghe\\_willy](https://www.bntl.nl/bntl/publicatie/om_en_rnd_de_om_te_-_konstrukt/vandeweghe_willy).
- Weijnen, Antonius Angelus en Jan J. A. Van Bakel (1967). *Voorlopige inleiding op het woordenboek van de Brabantse dialecten*. Assen: Van Gorcum.