Over DIRT: achtergrond en gebruik van het DIRT-corpus

Gauthier Delaby, Lien Hellebaut & Ulrike Vogl

Universiteit Gent

Versie 1.0, 30 oktober 2025

Inhoudsopgave

1 Informatie over DIRT

1.1 Over het DIRT-corpus

Het **DIRT-corpus** (**D**utch **I**n **R**eality **T**V) bestaat uit transcripties van Nederlandstalige realityseries zoals *De Mol*, *Chateau Meiland* en *Temptation Island*. Dat zijn programma's waarin niet geacteerd wordt en waarin we relatief **spontaan gesproken, informeel Nederlands** terugvinden. Het corpus bevat zowel oudere als actuele programma's in **zowel Belgisch Nederlands als Nederlands Nederlands**. Daarnaast werd het corpus verrijkt met **metadata**: het bevat informatie over bijvoorbeeld de regionale afkomst, gender, opleiding en leeftijd van de sprekers.

DIRT is geconcipieerd als een **groeiend corpus**: er zullen regelmatige nieuwe versies van het corpus verschijnen die zijn aangevuld met nieuw getranscribeerd materiaal.

1.2 Doelstellingen van het DIRT-corpus

Het DIRT-corpus wil in de eerste plaats een leemte vullen in de bestaande corpora voor het Nederlands. Het Nederlands in Vlaanderen en Nederland wordt gekenmerkt door een situatie van diaglossie (Auer 2005), met een continuüm van de standaardtaal tot de dialecten, met daartussen een intermediaire vorm (tussentaal). Bestaande grote corpora omvatten echter niet alle taalvormen op dit continuüm:

- Het SoNaR-corpus (Oostdijk et al. 2013) bevat zowel formele als informele taal, maar uitsluitend in geschreven form.
- Het Corpus Gesproken Nederlands (Oostdijk 2000) bevat wel gesproken taal, zowel uit voorbereide en formelere situaties als uit spontane en informele situaties. De sprekers wiens taal in het corpus verzameld is, kregen echter de instructie om standaardtaal te spreken (Vandekerckhove 2004: 983; Ghyselen et al. 2020a: 6). Alhoewel niet elke spreker daar in dezelfde mate in slaagde en er wel degelijk een aantal niet-standaardtalige elementen te vinden zijn in het CGN, is het dus vooral representatief voor (bedoelde) standaardtaal.
- De Vlaamse dialecten worden dan weer vertegenwoordigd in het recent gelanceerde Gesproken Corpus van de zuidelijk-Nederlandse Dialecten (GCND; Breitbarth et al. 2024).

Er is dus nood aan een corpus met spontaan, informeel en gesproken Nederlands. Het DIRT-project wil daar een antwoord op bieden in de vorm van een corpus met gesproken conversaties uit realityprogramma's. In reality-tv worden "gewone" Vlamingen en Nederlanders gefilmd in vrij spontane en informele situaties, waardoor we kunnen verwachten dat de sprekers zich niet genoodzaakt voelen om voor een standaardtaliger register te kiezen, zoals in het CGN. Tegelijk zal, gezien het groeiende dialectverlies (Vandekerckhove 2009), het taalgebruik van de doorsnee deelnemer aan reality-tv minder dialectisch zijn dan in het GCND. Het taalgebruik in het DIRT-corpus kan dus hoofdzakelijk gekarakteriseerd worden als informeel gesproken Nederlands, tussen de polen van het dialect en het Standaardnederlands in.

Het genre van reality-tv vormt om nog een aantal bijkomende redenen een interessante bron voor het bestuderen van informeel gesproken Nederlands. Ten eerste is er in reality-tv een rijke sociale stratificatie van de sprekers, aangezien sprekers van bijvoorbeeld verschillende regio's, leeftijden en sociale achtergronden worden vertegenwoordigd (Zenner et al. 2009: 28). Ten tweede beschikken we bij reality-tv over opnames in een hoge kwaliteit met daarbij ook videomateriaal, wat toelaat om bijvoorbeeld de buitentalige context van bepaalde uitingen te analyseren (Zenner

& Van De Mieroop 2017: 79). Ten derde worden deelnemers van realityprogramma's vaak langere tijd gevolgd, in verschillende contexten en in situaties met veranderende sociale dynamieken, wat toelaat om de sociale context van talige uitingen in kaart te brengen. Ten vierde worden er jaarlijks heel wat nieuwe realityprogramma's uitgezonden, waardoor er dus ook steeds nieuw potentieel materiaal voor het DIRT-corpus is.

1.3 Geschiedenis van het DIRT-project

Het DIRT-project ontstond in het kader van het mastervak "Nederlandse taalkunde: het hedendaagse Nederlandse taalsysteem", gedoceerd door Ulrike Vogl en Gauthier Delaby aan de Universiteit Gent. Voor dit vak moesten studenten een academische paper schrijven waarin ze rapporteerden over een kort empirisch onderzoek over interjecties in het Nederlands. De studenten konden vrij hun onderzoeksopzet en -methodologie bepalen, maar rond het academiejaar 2020-2021 werd vastgesteld dat elk jaar steeds meer studenten beslisten om realityprogramma's te transcriberen. Deze transcripties bleken taalmateriaal te bevatten dat op velerlei vlakken erg interessant was, maar door het gebrek aan een uniforme manier van transcriberen en een gebrek aan consistent bijgehouden metadata, waren die transcripties niet bruikbaar voor verder onderzoek.

Er werd daarom beslist om in het academiejaar 2021-2022 een project op te starten waarin studenten via een gestandaardiseerd protocol transcripties zouden maken van realityprogramma's. Dit gebeurde zowel door jobstudenten als door studenten in het mastervak "Nederlandse taalkunde: het hedendaagse Nederlandse taalsysteem" en bachelorstudenten in een nieuwe onderzoekslijn "Taalgebruik in reality-tv", onder begeleiding van Ulrike Vogl en Gauthier Delaby. Voor het transcriptieprotocol werd gebruik gemaakt van het al bestaande protocol van het Gesproken Corpus van de zuidelijk-Nederlandse Dialecten (Ghyselen et al. 2020a, 2020b).

Reeds in de loop van het voorjaar van 2022 werd duidelijk dat dit project leidde tot een bruikbare verzameling aan transcripties van spontaan gesproken, informeel Nederlands, die de basis vormde voor een aantal interessante onderzoeken van studenten naar onder andere de semantisch-pragmatische functies van interjecties en naar de volgorde in werkwoordsclusters. Op 27 maart 2022 kreeg het corpus ook zijn naam: Dutch In Reality TV – afgekort: DIRT.

In de daaropvolgende jaren werd het corpus verder aangevuld met nieuwe transcripties. Vanaf juli 2024 vervoegde Lien Hellebaut het DIRT-team als projectcoördinator en werden alle transcripties van een grondige controle voorzien.

Op 30 oktober 2025 werd de eerste versie van het DIRT-corpus gelanceerd.[1]

1.4 Financiering

DIRT wordt van juli 2024 tot en met juni 2026 ondersteund door de Basisfinanciering van het Bijzonder Onderzoeksfonds (BOF) van de Universiteit Gent.

[1] We bedanken graag alle studenten die transcripties hebben gemaakt voor het DIRT-corpus, alsook Anne-Sophie Ghyselen, Melissa Farasyn en Anne Breitbarth voor hun advies in de loop van het project.

1.5 Gebruiksrechten

1.6 Contact

- Ulrike Vogl Ulrike.Vogl@UGent.be
- Gauthier Delaby Gauthier.Delaby@UGent.be
- Lien Hellebaut Lien.Hellebaut@UGent.be

2 Inhoud en structuur van het corpus

Het DIRT-corpus wordt aangeboden als een zip-archief. Dit archief bevat de volgende elementen:

- een ReadMe-bestand <Read Me.txt>;
- een document met informatie over het DIRT-corpus en het DIRT-project <Over DIRT.pdf> (i.e., dit document);
- een document met een aantal statistieken over de huidige versie van het DIRT-corpus <Statistieken.xlsx>;
- een map "Corpus", waarin alle getranscribeerde afleveringen zijn opgenomen als aparte txt-bestanden;
- een map "Metadata", met daarin drie xlsx-bestanden met metadata over respectievelijk de programma's, afleveringen en sprekers in het DIRT-corpus.

De structuur van de txt-bestanden met de transcripties wordt toegelicht in §3.2. De structuur en de gegevens van de metadatabestanden worden toegelicht in §4.

3 Transcripties

3.1 Transcriptieprotocol

Alle afleveringen van de realityprogramma's in het DIRT-corpus werden manueel getranscribeerd volgens het transcriptieprotocol dat werd ontwikkeld voor het Gesproken Corpus van de zuidelijk-Nederlandse Dialecten (GCND; Ghyselen et al. 2020a, 2020b). Dit protocol werd ontwikkeld met als doel verschillende Nederlandse dialecten zo uniform mogelijk weer te geven in orthografische transcripties (Ghyselen et al. 2020b: 89). Daarom werd ervoor geopteerd om in twee lagen te transcriberen: elke uiting wordt getranscribeerd in een "lichte vernederlandsing", die dichter aanleunt bij het dialect, en in een "zware vernederlandsing", die meer aanleunt bij het Standaardnederlands.

In de lichte vernederlandsing – zie (1) voor een voorbeeld – wordt er abstractie gemaakt van dialectische klanken door ze te vervangen door hun standaardtalige "tegenhangers". Dialectische lexemen, functiewoorden, morfemen en syntaxis worden wel getranscribeerd, maar dan volgens de spellingsregels van het Standaardnederlands. Zo wordt in (1) de dialectische vorm van persoonlijk voornaamwoord voor de eerste persoon meervoud *wijder* behouden, alsook de subjectsverdubbeling. Clitisering wordt weergegeven met een #-teken tussen de verschillende

delen, zoals bij *m#en* in (1). In de zware vernederlandsing – zie (2) voor een voorbeeld – wordt clisis opgelost en worden niet-standaardtalige morfemen en functiewoorden vernederlandst: zo is in (2) het dialectische voornaamwoord *wijder* vervangen door de corresponderende standaardtalige vorm *wij* en is de clisis *m#en* opgelost tot *we hebben*. Dialectische lexemen worden behouden in zowel de lichte als zware vernederlandsing, zoals het geval is voor *klappen* 'praten' in de lichte vernederlandsing in (3) en in de zware vernederlandsing in (4). Hetzelfde geldt voor dialectische syntactische structuren: zo wordt in (2) de subjectsverdubbeling (*we hebben wij*) bewaard en wordt het temporele hulpwerkwoord *hebben* niet vervangen door het standaardtalige equivalent *zijn*.

(1) m#en wijder zeker (...) vijf jaar wegewist wi. (Ghyselen et al. 2020b: 110)
(2) we hebben wij zeker (...) vijf jaar weggeweest wi. (Ghyselen et al. 2020b: 110)
(3) ma ja ze klappen al nie lijk ik zeker? (Ghyselen et al. 2020b: 106)
(4) maar ja ze klappen al niet gelijk ik zeker? (Ghyselen et al. 2020b: 106)

De zware vernederlandsing heeft een dubbel doel: enerzijds laat die gebruikers toe om de transcripties te doorzoeken zonder specifieke dialectkennis, anderzijds kan die als basis dienen voor bepaalde annotatietools die op standaardtaal zijn getraind (Ghyselen et al. 2025: 210).

Het DIRT-corpus bevat materiaal uit realityseries, die zijn opgekomen in de jaren 90 van de vorige eeuw. Gezien het voortschrijdende dialectverlies in Vlaanderen en Nederland (Willemyns 1979; Van Keymeulen 1993; Vandekerckhove 2009; Van Hoof & Jaspers 2012; Ghyselen & Van Keymeulen 2014), zijn er ook veel minder dialectische kenmerken aanwezig in de transcripties. Zodoende stelt zich de vraag of het voor het DIRT-corpus wel nodig is om met twee transcriptielagen te werken. Elke uiting transcriberen in twee lagen is immers een stuk tijdrovender, terwijl er ook veel minder dialectische taal is waarvoor transcriberen in twee lagen net wenselijk is. De beslissing om desondanks het GCND-transcriptieprotocol te hanteren, is tweevoudig gemotiveerd. Ten eerste is gebleken dat dit protocol behoorlijk vlot kan toegepast worden door transcribeerders: Ghyselen et al. (2025) voerden een foutenanalyse uit waarbij taalkundigen met moedertaalkennis van specifieke dialecten transcripties controleerden en stelden vast dat de word error rate van gecontroleerde transcripties 1.29% was, wat wijst op een hoge betrouwbaarheid. Ten tweede bevat het materiaal in DIRT minder dialectische taal dan in een dialectcorpus zoals het GCND, maar dat betekent zeker niet dat er geen dialect aanwezig is. Er zijn dus wel degelijk de nodige talige uitingen waarvoor een transcriptie in twee lagen een meerwaarde heeft met het oog op een efficiënte doorzoekbaarheid en op verwerking door annotatietools.

Naast de bovenstaande principes met betrekking tot het transcriberen in twee lagen, dient ook nog gewezen te worden op enkele bijzondere tekens die volgens het protocol worden gehanteerd:

???	De transcribeerder begrijpt niet wat er gezegd wordt, maar denkt dat iemand anders het misschien wel zou kunnen begrijpen.
xx	Een bepaald woorddeel is onverstaanbaar, bijvoorbeeld door achtergrondlawaai.
xxx	Een grotere passage is onverstaanbaar, bijvoorbeeld door achtergrondlawaai.
ggg	Sprekersgeluiden, zoals gelach of gehuil.

Voor een uitgebreide bespreking van het transcriptieprotocol en de motivering van bepaalde keuzes, verwijzen we naar Ghyselen et al. (2020a, 2020b). Hieronder sommen we een aantal punten op waarbij werd afgeweken van het GCND-transcriptieprotocol.

[A] In de zware vernederlandsing vervangen we niet-standaardtalige syntactische structuren door hun standaardtalige equivalent indien dat kan zonder de woordvolgorde en het aantal woorden te veranderen. Deze beslissing werd genomen om in de toekomst eventuele parsing van de zinnen op basis van de zware vernederlandsing door annotatietools te vergemakkelijken. Zo wordt presentatief *het* in (5) in de zware vernederlandsing vervangen door het standaardtalige presentatief *er* (6) en wordt het dialectische temporele hulpwerkwoord *zijn* in (7) vervangen door *hebben* in de zware vernederlandsing (8). Daarmee wordt er in het DIRT-corpus in hogere mate vernederlandst in de zware vernederlandsing dan in het GCND-corpus (vergelijk (7) en (8) met (1) en (2)). Dialectische en tussentalige syntactische structuren zoals een dubbele negatie (9) en een V3-structuur (11) worden niet vervangen door de standaardtalige structuur, omdat dat niet kan zonder de woordvolgorde of het aantal elementen te vervangen. De woordvolgorde en het aantal elementen worden steeds behouden in de zware vernederlandsing, zodat de koppeling tussen elementen in de lichte en zware vernederlandsing mogelijk blijft.

(5) ja t#sta voor jou ietske meer op#t spel natuurlijk ggg. (lichte vernederlandsing)
(6) ja er staat voor jou ietsje meer op het spel natuurlijk ggg. (zware vernederlandsing)
(7) ik zijn naar school geweest. (lichte vernederlandsing)
(8) ik heb naar school geweest. (zware vernederlandsing)
(9) ik en heb niet gedaan. (lichte vernederlandsing)
(10) ik en heb niets gedaan. (zware vernederlandsing)
(11) morgen ik ga dat doen. (lichte vernederlandsing)
(12) morgen ik ga dat doen. (zware vernederlandsing)

[B] Volgens het GCND-protocol wordt het beletselteken gebruikt op het einde van incomplete woorden en zinnen en om pauzes aan te duiden. In DIRT wordt het beletselteken enkel gebruikt voor incomplete woorden en zinnen. Pauzes worden aangeduid met de code ppp.

[C] Woorden en grotere gehelen die de normale zinsstructuur doorbreken, worden tussen vierkante haken gezet. Het gaat hierbij onder meer om incomplete woorden, verbeteringen en herhalingen, zie (13)-(16) voor voorbeelden. Dit laat toe om in een later stadium annotatietools de instructie te geven om die woorden/groepen te negeren bij bijvoorbeeld het parsen van een dependentieboom. Merk op dat een dubbele negatie niet tussen vierkante haken wordt gezet (17), aangezien dit grammaticale constructies zijn in de dialecten waarin ze voorkomen.

(13) ik eet [ap…] appelsienen.
(14) en ons Marie die [schreef] vulde die kaart in.
(15) hé [dat] dat is mijn geluk geweest hé.
(16) ik eet dat [heel heel] heel graag.
(17) [ik] ik en ga dat niet doen.

[D] Volgens het GCND-protocol worden er drie leestekens gebruikt op het einde van een zin (punt, vraagteken of beletselteken), maar worden er geen leestekens gebruikt binnen de zin (met uitzondering van het beletselteken na incomplete woorden). In de transcripties voor DIRT werden drie soorten leestekens toegelaten binnen de zin. Ten eerste kan het beletselteken gebruikt worden bij incomplete woorden (cf. supra). Ten tweede worden komma's gebruikt waar de transcribeer dat nodig acht. Ten derde worden intercalaties tussen gedachtestreepjes geplaatst.

3.2 Transcriptieproces

Alle transcripties werden gemaakt met de transcriptiesoftware ELAN. Tijdens het transcriberen worden er "annotaties" aangemaakt: dat zijn, volgens de terminologie die ELAN hanteert, korte stukjes transcriptie die verbonden zijn met tijdscodes die aangeven op welk tijdstip in het realityprogramma een bepaalde uitspraak werd gedaan. Deze annotaties bevatten normaal gezien één zin, maar in sommige gevallen ook meerdere zinnen indien verschillende zinnen van een spreker elkaar snel opvolgen en het daardoor moeilijk is om ze nauwkeurig af te bakenen op de tijdsas. Hoewel wij geen audio- of videobestanden aanbieden, kunnen gebruikers van het corpus zelf de afleveringen in kwestie online opzoeken (indien ze nog beschikbaar zijn) en op basis van de tijdscodes de relevante passages in de afleveringen terugvinden.

De transcripties werden enerzijds gemaakt door studenten Taal- en Letterkunde aan de Universiteit Gent, en anderzijds door de projectmedewerkers.

De meeste transcripties die door studenten werden gemaakt, werden nadien nog gecontroleerd door een projectmedewerker. Of een transcriptie al dan niet werd gecontroleerd, wordt aangegeven in het afleveringenoverzicht (zie §4.2).

Merk op dat bepaalde delen van realityprogramma's niet werden getranscribeerd:

- achtergrondmuziek;
- dialogen en/of scènes die volledig in een andere taal zijn dan het Nederlands (i.e., waarin geen enkel woord Nederlands voorkomt);
- intro's (die elke aflevering worden herhaald), voorbeschouwingen en terugblikken.

3.3 Structuur van de transcripties

De transcripties bestaan uit aparte txt-bestanden per getranscribeerde aflevering. Elk txt-bestand bestaat uit een tab-separated tabel, die onderstaande kolommen bevatten:

#	Kolom	Uitleg
1.	**Programmacode**	Een unieke code per seizoen van een programma, zie §4.1 voor meer toelichting.
2.	**Afleveringcode**	Een unieke code per aflevering van een programma, zie §4.2 voor meer toelichting.
3.	**Sprekerscode**	Een unieke code per spreker, zie §4.3 voor meer toelichting.
4.	**LVN**	De transcripties in lichte vernederlandsing.
5.	**ZVN**	De transcripties in zware vernederlandsing.
6.	**Starttijd**	De starttijd van de getranscribeerde tekst in de aflevering.
7.	**Eindtijd**	De eindtijd van de getranscribeerde tekst in de aflevering.

4 Metadata

4.1 Overzicht programma's

Het programmaoverzicht bevat één rij per seizoen van een programma dat opgenomen is in het corpus. Onderstaande tabel biedt een overzicht van de variabelen in het programmaoverzicht:

#	Variabele	Uitleg en mogelijke waarden
1.	**Programmacode**	De programmacode is een unieke code per seizoen van een programma. De code bestaat uit drie delen die van elkaar gescheiden worden door een underscore: • een code die aangeeft of het programma uit België (BE) komt, uit Nederland (NL) komt of een coproductie tussen Belgische en Nederlandse zenders is (BENL); • de naam van het realityprogramma in CamelCase; • het jaar van de uitzending. Voorbeelden van programmacodes zijn BE_BoerZoektVrouw_2020 en NL_ChateauMeiland_2023.
2.	**Programma**	De volledige naam van het realityprogramma. Om de spelling consistent te houden, krijgen enkel het eerste woord en eigennamen hoofdletters.
3.	**Zender**	De zender waarop het programma werd uitgezonden. Aangezien zenders regelmatig van naam veranderen, gebruiken we bij voorkeur de naam die de zender had op het moment van de uitzending.
4.	**Seizoen**	Het nummer van het seizoen van het realityprogramma dat getranscribeerd werd.
5.	**Jaar**	Het jaar waarin het realityprogramma werd uitgezonden. (Let op: het gaat hier dus niet om het jaar waarin het programma werd opgenomen, aangezien dit meestal niet bekend is.)
6.	**Land**	Het land waar het realityprogramma oorspronkelijk werd uitgezonden. Deze variabele heeft een van de volgende waarden: ✓ BE ✓ NL ✓ BE+NL: voor coproducties tussen Belgische en Nederlandse zenders ✓ ANDERS
7.	**Type_deelnemers**	Realityprogramma's worden soms ingedeeld volgens hun focus op bekende personen of op onbekende personen. Aangezien bekende personen over het algemeen meer media-ervaring en -training hebben (en in sommige gevallen daarbij de instructie hebben gekregen om een standaardtalig register te hanteren), kan hun taalgebruik eventueel verschillen van dat van onbekende personen. Daarom werd per programma bijgehouden of de voornaamste "deelnemers" aan het programma bekende of onbekende personen zijn. Deze variabele heeft een van de volgende waarden: ✓ BEKENDE_PERSONEN: de deelnemers aan het programma zijn voornamelijk bekende personen (bv. in Chateau Meiland, De Verhulstjes of Over De Oceaan); ✓ ONBEKENDE_PERSONEN: de deelnemers aan het programma zijn voornamelijk onbekende personen (bijvoorbeeld voor een spelprogramma waarbij enkel

#	Variabele	Uitleg en mogelijke waarden
		de presentator een bekende persoon is, maar de andere deelnemers niet); ✓ BEIDE: de deelnemers zijn zowel bekende als onbekende personen; ✓ ONDUIDELIJK.
8.	**Aantal_afleveringen_totaal**	Het aantal afleveringen dat het seizoen van het realityprogramma telt.
9.	**Aantal_afleveringen_getranscribeerd**	Het aantal afleveringen van het seizoen van het realityprogramma waarvoor er een transcriptie in het DIRT-corpus is opgenomen.
10.	**Aantal_woorden**	Het aantal woorden dat de transcripties voor het seizoen van het realityprogramma in totaal bevatten.
11.	**Getranscribeerde_tijd**	De duur van het getranscribeerde deel van het realityprogramma, uitgedrukt in uu:mm:ss.msmsms.

4.2 Overzicht afleveringen

Het afleveringenoverzicht bevat één rij per aflevering die opgenomen is in het corpus – in tegenstelling tot het programmaoverzicht dat één rij per seizoen van een programma bevat. Onderstaande tabel biedt een overzicht van de variabelen in het programmaoverzicht:

#	Variabele	Uitleg en mogelijke waarden
1.	**Programmacode**	Idem als bij het programmaoverzicht, zie §4.1.
2.	**Afleveringcode**	De afleveringcode is een unieke code per getranscribeerde aflevering. De code bestaat uit de programmacode, gevolgd door een underscore en het afleveringnummer (bestaande uit drie cijfers). Voorbeelden van afleveringcodes zijn BE_BoerZoektVrouw_2020_001 en NL_ChateauMeiland_2023_003.
3.	**Programma**	Idem als bij het programmaoverzicht, zie §4.1.
4.	**Zender**	Idem als bij het programmaoverzicht, zie §4.1.
5.	**Seizoen**	Idem als bij het programmaoverzicht, zie §4.1.
6.	**Jaar**	Idem als bij het programmaoverzicht, zie §4.1.
7.	**Land**	Idem als bij het programmaoverzicht, zie §4.1.
8.	**Type_deelnemers**	Idem als bij het programmaoverzicht, zie §4.1.
9.	**Aantal_afleveringen_totaal**	Idem als bij het programmaoverzicht, zie §4.1.
10.	**Nummer_aflevering**	Het nummer van de getranscribeerde aflevering binnen het seizoen van het realityprogramma.
11.	**Volledig_getranscribeerd**	Deze variabele geeft aan of de aflevering volledig getranscribeerd is, of slechts gedeeltelijk. Deze variabele heeft een van de volgende waarden: ✓ VOLLEDIG_GETRANSCRIBEERD ✓ DEELS_GETRANSCRIBEERD
12.	**Aantal_woorden**	Het aantal woorden dat de transcriptie van de aflevering in totaal bevat.
13.	**Getranscribeerde_tijd**	De duur van het getranscribeerde deel van de aflevering, uitgedrukt in uu:mm:ss.msmsms.
14.	**Transcribeerders**	Een identificatienummer van 7 cijfers dat aangeeft wie de transcriptie heeft gemaakt. Indien meerdere

#	Variabele	Uitleg en mogelijke waarden
		transcribeerders aan een transcriptie hebben gewerkt, worden hun identificatienummers gescheiden door een puntkomma.
15.	**VerwerktDoor**	Een identificatienummer van 7 cijfers dat aangeeft wie de transcriptie in het corpus heeft verwerkt. Indien meerdere personen hiervoor verantwoordelijk waren, worden hun identificatienummers gescheiden door een puntkomma.
16.	**Controle**	Deze variabele geeft aan of de transcriptie werd gecontroleerd door een projectmedewerker. Dit gebeurt steeds volgens het vierogenprincipe: indien een transcriptie werd gemaakt door een bepaalde projectmedewerker, kan die niet zelf de controle voor deze transcriptie uitvoeren. Deze variabele heeft een van de volgende waarden: ✓ JA: de transcriptie werd gecontroleerd; ✓ NEE: de transcriptie werd niet gecontroleerd.
17.	**ControleDoor**	Een identificatienummer van 7 cijfers dat aangeeft wie de controle van de transcriptie heeft uitgevoerd. Indien meerdere personen hiervoor verantwoordelijk waren, worden hun identificatienummers gescheiden door een puntkomma.
18.	**Update**	De datum waarop de transcriptie voor het laatst werd geüpdatet.

4.3 Overzicht sprekers

Het sprekersoverzicht bevat één rij per spreker die opgenomen is in het corpus.

Daarnaast werd er voor verschillende programma's ook een extra spreker in de transcripties aangemaakt, waaraan alle uitspraken werden toegekend waarvoor het niet duidelijk was door wie ze precies werden uitgesproken. De zinnen en andere uitingen bij deze spreker zijn in werkelijkheid dus niet noodzakelijk allemaal door dezelfde persoon uitgesproken. Deze spreker heeft steeds volgcijfer 000 (zie variabele #2 hieronder) en bevat logischerwijs geen verdere metadata.

De metadata voor de sprekers werden in de mate van het mogelijke gepseudonimiseerd.

Onderstaande tabel biedt een overzicht van de variabelen in het programmaoverzicht:

#	Variabele	Uitleg en mogelijke waarden
1.	**Code_programma**	De programmacode, zie variabele #1 in §4.1 voor toelichting.
2.	**Code_spreker**	De sprekerscode is een unieke code per spreker. De code bestaat uit de programmacode, gevolgd door een underscore en een volgcijfer per spreker (bestaande uit drie cijfers). Voorbeelden van sprekerscodes zijn BE_BoerZoektVrouw_2020_001 en NL_ChateauMeiland_2023_003. Let op: • De afleveringcodes en sprekerscodes zijn dus identitek opgebouwd. Zo kan een code met "003" verwijzen naar de derde aflevering van een programma of naar de derde spreker. Het is dus

#	Variabele	Uitleg en mogelijke waarden
		belangrijk om te controleren dat de codes uit de juiste kolom en uit het juiste metadatabestand worden gebruikt. • Sprekerscodes met volgnummer 000 verwijzen niet naar één specifieke spreker, maar geven alle getranscribeerde tekst weer waarvoor onduidelijk was door wie ze precies werd uitgesproken.
3.	**RolInProgramma**	Geeft weer in welke hoedanigheid de spreker voorkomt in het realityprogramma. Deze variabele heeft een van de volgende waarden: ✓ PRESENTATOR; ✓ VOICE-OVER; ✓ interviewer: een productiemedewerker van het programma die de deelnemers van het programma interviewt; ✓ DEELNEMER: dit label wordt breed geïnterpreteerd en wordt gebruikt voor alle personen die een centrale rol hebben in een realityprogramma (bijvoorbeeld kandidaten in een spelprogramma, een familie wiens leven wordt gevolgd in een programma etc.); ✓ NIET-PROMINENT: alle personen die geen centrale rol hebben in het realityprogramma en gewoonlijk slechts (heel) kort aan bod komen; ✓ NA: wordt enkel gebruikt voor spreker 000 (zie variabele #2 voor toelichting). Let op: indien de persoon die verantwoordelijk was voor de voice-over ook in een andere hoedanigheid in het programma voorkomt (bijvoorbeeld als presentator), worden beide voorkomens als aparte sprekers behandeld. Deze sprekers zijn dan steeds aan elkaar gelinkt, zie variabele #18.
4.	**Land**	Deze variabele geeft weer in welk land de spreker woont op het moment van de opnames van het programma. Deze variabele heeft een van de volgende waarden: ✓ BE: de spreker woont in België; ✓ NL: de spreker woont in Nederland; ✓ BE+NL: de spreker woont in België en Nederland; ✓ ANDERS: de spreker woont in een ander land dan België en Nederland. ✓ NA: wordt enkel gebruikt voor spreker 000 (zie variabele #2 voor toelichting). Let op: de waarde van deze variabele staat los van de landcode in de programma- en sprekerscode, aangezien die verwijst naar het land waar het *programma* gemaakt werd. Zo is het bijvoorbeeld mogelijk dat in een Belgisch programma (= landcode “BE” in de programmacode) een deelnemer uit Nederland (= “NL” als waarde bij deze variabele) meedoet.

#	Variabele	Uitleg en mogelijke waarden
5.	**Leeftijd**	De leeftijd van de spreker op het moment van de opnames van het programma. Voor sprekers waarvoor het geboortejaar bekend was, maar de exacte leeftijd niet, werd een "geschatte leeftijd" berekend, door het geboortejaar af te trekken van het jaar van de uitzending van het programma (i.e., het jaar dat vermeld is in de programmacode). Deze geschatte leeftijd zal niet helemaal nauwkeurig zijn doordat gewoonlijk niet bekend is (i) of de spreker al dan niet al verjaard is op het moment in het jaar dat de opnames gemaakt zijn en (ii) hoeveel tijd er zat tussen de opnames en de effectieve uitzending van het programma. Vermoedelijk zal de geschatte leeftijd in de meeste gevallen met maximaal 2 jaar afwijken van de werkelijke leeftijd. De geschatte leeftijden worden voorafgegaan door een asterisk. Indien de leeftijd (en ook het geboortejaar) niet bekend is, wordt de waarde NA toegekend.
6.	**Geboortejaar**	Het geboortejaar van de spreker. Voor sprekers waarvoor de leeftijd bekend was, maar het geboortejaar niet, werd een "geschat geboortejaar" berekend, door de leeftijd af te trekken van het jaar van de uitzending van het programma (i.e., het jaar dat vermeld is in de programmacode). Dit geschatte geboortejaar zal niet helemaal nauwkeurig zijn doordat gewoonlijk niet bekend is (i) of de spreker al dan niet al verjaard is op het moment in het jaar dat de opnames gemaakt zijn en (ii) hoeveel tijd er zat tussen de opnames en de effectieve uitzending van het programma. Vermoedelijk zal het geschatte geboortejaar in de meeste gevallen met maximaal 2 jaar afwijken van het werkelijke geboortejaar . De geschatte geboortejaren worden voorafgegaan door een asterisk. Indien het geboortejaar (en ook de leeftijd) niet bekend is, wordt de waarde NA toegekend.
7.	**Geslacht**	Het geslacht of gender van de spreker. Deze variabele heeft een van de volgende waarden: ✓ M: mannelijk; ✓ V: vrouwelijk; ✓ X: non-binair of andere; ✓ NA: geslacht/gender onbekend.
8.	**Provincie_huidig**	De provincie waar de spreker woont op het moment van de opnames van het programma. Deze variabele heeft een van de volgende waarden: ✓ WEST-VLAANDEREN ✓ OOST-VLAANDEREN ✓ ANTWERPEN ✓ VLAAMS-BRABANT

#	Variabele	Uitleg en mogelijke waarden
		✓ LIMBURG ✓ BRUSSEL ✓ WALLONIË ✓ GRONINGEN ✓ FRIESLAND ✓ DRENTHE ✓ OVERIJSSEL ✓ FLEVOLAND ✓ GELDERLAND ✓ UTRECHT ✓ NOORD-HOLLAND ✓ ZUID-HOLLAND ✓ ZEELAND ✓ NOORD-BRABANT ✓ NA: de provincie waar de spreker woont is niet bekend.
9.	**Woonplaats_huidig**	De huidige woonplaats (gemeente of stad) van de spreker. Indien de huidige woonplaats niet bekend is, wordt de waarde NA toegekend.
10.	**Provincie_jeugd**	De provincie waar de spreker tijdens zijn jeugd heeft gewoond. Indien de spreker tijdens zijn jeugd in meerdere provincies heeft gewoond, worden de verschillende provincies door een puntkomma gescheiden. Indien de provincie(s) waar de spreker tijdens zijn jeugd heeft gewoond niet bekend zijn, wordt de waarde NA toegekend.
11.	**Woonplaats_jeugd**	De woonplaats (gemeente of stad) waar de spreker tijdens zijn jeugd heeft gewoond. Indien de spreker tijdens zijn jeugd in meerdere plaatsen heeft gewoond, worden de verschillende woonplaatsen door een puntkomma gescheiden. Indien de woonplaatsen waar de spreker tijdens zijn jeugd heeft gewoond niet bekend zijn, wordt de waarde NA toegekend.
12.	**Beroep**	Het beroep van de spreker op het moment van de opnames van het programma. Indien de spreker meerdere beroepen heeft, worden ze door een puntkomma gescheiden. Indien het beroep niet bekend is, wordt de waarde NA toegekend. Let op: de weergave en schrijfwijzen van beroepen is niet geuniformiseerd in het protocol. Verschillende transcribeerders kunnen bepaalde beroepen dus op een andere manier noteren (bijvoorbeeld leerkracht vs. leraar).
13.	**Beroepsniveau**	Voor de sprekers waarvan een beroep bekend is, werd dat beroep ingedeeld in een beroepsniveau. De indeling is gebaseerd op degene die werd gehanteerd voor het Corpus Gesproken Nederlands (Oostdijk 2000): daarbij worden beroepen onderscheiden die ongeschoold zijn (bv. vuilnisman, poetsvrouw, taxi chauffeur, etc.), beroepen die een lagere opleiding veronderstellen (bv. mecanicien, kleuterleidster, bankemployee, etc.), beroepen die een middelhoge opleiding veronderstellen (bv. leraar, journalist,

#	Variabele	Uitleg en mogelijke waarden
		politicus, acteur, artiest, etc.) en beroepen die een hogere opleiding veronderstellen (bv. arts, advocaat, hoger management, bestuur, etc.). Sprekers die niet beroepsmatig actief zijn, worden ingedeeld in de categorieën “leerling/student”, “werkloos”, “huisman/-vrouw”, “arbeidsongeschikt” en “anders, bv. met pensioen”. Indien het beroep niet bekend is, wordt de waarde NA toegekend. Deze variabele heeft een van de volgende waarden: ✓ 0_ARBEIDSONGESCHIKT ✓ 0_HUISMANHUISVROUW ✓ 0_LEERLINGSTUDENT ✓ 0_PENSIOEN ✓ 0_WERKLOOS ✓ 1_ONGESCHOOLDBEROEP ✓ 2_LAGEREOPLEIDING ✓ 3_MIDDELHOGEOPLEIDING ✓ 4_HOGEREOPLEIDING ✓ NA Let op: de indeling van beroepen in een van deze niveaus is niet altijd vanzelfsprekend en vereist soms een subjectief oordeel van de transcribeerder. Wie deze variabele gebruikt voor onderzoeksdoeleinden, controleer best of de indeling overeenkomt met de verwachtingen voor het onderzoek.
14.	**Opleiding**	Het hoogste diploma dat de spreker reeds behaald heeft. Indien de opleiding niet bekend is, wordt de waarde NA toegekend. Deze variabele heeft een van de volgende waarden: ✓ LAGERONDERWIJS ✓ MIDDELBAARONDERWIJS_ASO/TSO/VWO/HAVO ✓ MIDDELBAARONDERWIJS_BSO/KSO/VMBO ✓ MIDDELBAARONDERWIJS_BUSO/PRAKTIJKSCHOOL ✓ MIDDELBAARONDERWIJS_SUBNIVEAU_ONBEKEND ✓ HOGERONDERWIJS_DOCTORAAT ✓ HOGERONDERWIJS_MASTER ✓ HOGERONDERWIJS_ACADEMISCHEBACHELOR /BACHELORWO ✓ HOGERONDERWIJS_PROFESSIONELEBACHELOR/ GRADUAAT/HBO/PABO ✓ HOGERONDERWIJS_SUBNIVEAU_ONBEKEND ✓ NA Enkele oude onderwijsvormen uit Nederland werden als volgt ingepast in bovenstaande indeling: • vbo (voorbereidend beroepsonderwijs) > vmbo • lbo (lager beroepsonderwijs) > vmbo • mavo (middelbaar algemeen voortgezet onderwijs) > vmbo

#	Variabele	Uitleg en mogelijke waarden
		• mulo (meer uitgebreid lager onderwijs) > vmbo • mms (middelbare meisjesschool) > havo
15.	**Nederlandstalig**	Deze variabele geeft weer of de spreker Nederlandstalig is, en zo ja, als moedertaal of als tweede taal. Deze variabele heeft een van de volgende waarden: ✓ NT1: de spreker spreekt Nederlands als moedertaal; ✓ NT2: de spreker spreekt Nederlands, maar niet als moedertaal; ✓ ANDERSTALIG: de spreker heeft geen kennis van het Nederlands; ✓ NA: de kennis van het Nederlands van de spreker is onbekend.
16.	**Talenkennis**	Geeft aan welke talen een spreker beheerst, naast het Nederlands. We veronderstellen dat een spreker een taal beheerst (i) als hij/zij aangeeft de taal te spreken of (ii) als hij/zij die taal effectief spreekt in het realityprogramma. Indien de spreker een bepaalde taal als moedertaal heeft, wordt dat weergegeven met een asterisk achter de taal. Indien de spreker meerdere talen beheerst, worden die gescheiden door een puntkomma. Bij sprekers die geen anderen talen beheersen of waarvoor geen informatie daarover bekend is, wordt de waarde NA toegekend.
17.	**Verblijf_buitenland**	Deze variabele codeert of een spreker een langere periode in het buitenland heeft doorgebracht. Dit kan bijvoorbeeld betrekking hebben op mensen die op latere leeftijd naar België/Nederland verhuisden en een deel van hun jeugd in een ander land doorbrachten, maar ook bijvoorbeeld op mensen die als expat enkele jaren in het buitenland werkten. Korte reizen en vakanties tellen hierbij niet mee. Indien een spreker in meerdere landen een langere periode heeft doorgebracht, worden die gescheiden door een puntkomma. Bij sprekers die geen langer verblijf hebben gehad in het buitenland of waarvoor geen informatie daarover bekend is, wordt de waarde NA toegekend.
18.	**Identificatienummer**	Sommige sprekers komen in meerdere realityprogramma's in het corpus voor. Voor sommige onderzoeksdoeleinden kan het nuttig zijn om de verschillende sprekerscodes van eenzelfde spreker aan elkaar te linken (bv. om het taalgebruik van een spreker over verschillende jaren heen te bestuderen). Daarom hebben alle sprekers die in meerder realityprogramma's in het corpus voorkomen in deze variabele een uniek identificatienummer gekregen.

#	Variabele	Uitleg en mogelijke waarden
		Aan sprekers die niet in meerdere programma's voorkomen, werd de waarde NA toegekend. Let op: het is, in uitzonderlijke gevallen, mogelijk dat een bepaalde spreker in meerdere programma's voorkomt, maar dat dit niet kon worden vastgesteld door een gebrek aan voldoende identificerende informatie in één van die programma's.
19.	**Aantal_woorden**	Het aantal woorden van de spreker in alle getranscribeerde afleveringen van het realityprogramma.
20.	**Getranscribeerde_tijd**	De totale tijd dat de spreker aan het woord is in de getranscribeerde afleveringen van het realityprogramma, uitdrukt in in uu:mm:ss.msmsms.
21.	**Transcribeerders**	Een identificatienummer van 7 cijfers dat aangeeft wie de sprekersinformatie heeft opgesteld. Indien meerdere transcribeerders hiervoor verantwoordelijk waren, worden hun identificatienummers gescheiden door een puntkomma.
22.	**VerwerktDoor**	Een identificatienummer van 7 cijfers dat aangeeft wie het sprekersoverzicht in het corpus heeft verwerkt. Indien meerdere personen hiervoor verantwoordelijk waren, worden hun identificatienummers gescheiden door een puntkomma.
23.	**Controle**	Deze variabele geeft aan of het sprekersoverzicht werd gecontroleerd door een projectmedewerker. Dit gebeurt steeds volgens het vierogenprincipe: indien een sprekersoverzicht werd opgesteld door een bepaalde projectmedewerker, kan die niet zelf de controle voor deze transcriptie uitvoeren. Deze variabele heeft een van de volgende waarden: ✓ JA: de transcriptie werd gecontroleerd; ✓ NEE: de transcriptie werd niet gecontroleerd.
24.	**ControleDoor**	Een identificatienummer van 7 cijfers dat aangeeft wie de controle van de sprekersinformatie heeft uitgevoerd. Indien meerdere personen hiervoor verantwoordelijk waren, worden hun identificatienummers gescheiden door een puntkomma.
25.	**Update**	De datum waarop het sprekersoverzicht voor het laatst werd geüpdatet.

5 Het corpus doorzoeken

Aangezien de transcripties bestaan uit eenvoudige tab-separated txt-bestanden, kan het doorzocht worden met alle gangbare software en methoden die geschikt zijn voor txt-bestanden. Daarnaast voorzien wij een eigen concordantietool waarmee het DIRT-corpus gemakkelijk kan doorzocht worden. De **DIRT concordancer** kan gedownload worden via de volgende link: https://doi.org/10.5281/zenodo.17469038.

De DIRT concordancer heeft een aantal handige functionaliteiten met het oog op de specifieke kenmerken van het DIRT-corpus:

- De concordantietool kan, indien gewenst, automatisch voor elke hit van een zoekopdracht de sprekersvariabelen annoteren.
- Met de concordantietool kan je kiezen of je wil zoeken in de lichte vernederlandsingen, de zware vernederlandsingen of in de sprekerscodes. In de output van de tool krijg je altijd beide transcriptielagen en de sprekerscodes te zien. Dit laat je toe om bijvoorbeeld in de zware vernederlandsing te zoeken naar het morfeem *-tje* voor verkleinwoorden en dan in de output na te gaan welke tussentalige/dialectische morfemen daarvoor werden gebruikt in de lichte vernederlandsing.
- De concordantietool kan "annotatie-eenheden" waarin meerdere hits zijn aangetroffen dupliceren in de output, zodat elke rij overeenkomt met één hit.
- De concordantietool laat toe om met RegEx "capturing groups" te bepalen om specifieke onderdelen uit een zoekopdracht in een aparte kolom weer te geven in de output.
- De concordantietool slaat de output onmiddellijk op als een xlsx-bestand.

De concordantietool wordt aangeboden als een portable exe-bestand en moet dus niet geïnstalleerd worden. Het programma wordt geopend via het exe-bestand en kan ook verwijderd worden door het exe-bestand te verwijderen.

Let op: de DIRT concordancer werkt enkel op Windows.

Onderstaande tabel biedt een overzicht van de opties en instellingen van de DIRT concordancer:

Selecteer de map waar de txt-bestanden van het DIRT-corpus zijn opgeslagen	Kies in de verkenner de map waar de txt-bestanden met transcripties zijn opgeslagen.
Selecteer de map waar de output van de DIRT concordancer moet worden opgeslagen	Kies in de verkenner de map waar het xlsx-bestand met de resultaten van je zoekopdracht mag worden opgeslagen door de concordancer.
Indien je de sprekersvariabelen wil laten annoteren, selecteer dan het sprekersoverzicht (optioneel)	Kies in de verkenner het xlsx-bestand met het sprekersoverzicht, indien je wil dat elke hit wordt geannoteerd voor de sprekersvariabelen in het sprekersoverzicht.
Zoekopdracht	Vul hier je zoekopdracht in. Hierbij kan je, indien gewenst, gebruik maken van reguliere expressies.
In welke laag van het DIRT-corpus wil je zoeken?	Kies hier in welke laag van het DIRT-corpus je wil zoeken: ✓ Lichte vernederlandsing (LVN) ✓ Zware vernederlandsing (ZVN) ✓ Sprekerscodes
Hoeveel voorafgaande en volgende zinnen (= annotatatieblokjes) moet de context van de hit bevatten?	Bij elke hit worden in de output ook twee kolommen toegevoegd met respectievelijk de linker- en rechtercontext van een hit. Deze instelling laat je toe om te bepalen hoeveel voorafgaande en volgende zinnen context je wil krijgen in de output.
Moet de context van de hit de sprekerscodes bevatten?	Bij elke hit worden in de output ook twee kolommen toegevoegd met respectievelijk

	de linker- en rechtercontext van een hit. Deze instelling laat je toe om te bepalen of die context ook de bijhorende sprekerscodes moet bevatten.
Moeten rijen met meerdere hits gedupliceerd worden per hit?	Kies hier of "annotatie-eenheden" moeten gedupliceerd worden in de output indien ze meerdere hits bevatten. De output variabele bevat een variabele "duplicaat", die aangeeft of een hit al dan niet een duplicaat is van een andere.
Welke regex capturing groups moeten in de dataset in aparte kolommen weergegeven worden? Vul de nummers van de groepen in gescheiden door een puntkomma, bv. 1;3 voor groepen 1 en 3. Laat dit veld leeg als je deze optie niet nodig hebt.	De DIRT concordancer biedt de mogelijkheid om zogenaamde capturing groups in de output in een aparte kolom weer te geven. Capturing groups worden afgebakend met behulp van ronde haakjes. Wil je bijvoorbeeld zoeken naar de verschillende meervoudsvormen van de substantieven *brandweerman*, *vakman* en *cameraman*, dan kan de volgende zoekopdracht gebruikt worden: (brandweer|vak|camera) (mannen|lieden|lui) De tweede capturing group (i.e., wat tussen de tweede set ronde haakjes staat) bevat dus de aangetroffen meervoudsvorm: *mannen* vs. *lieden* vs. *lui*. Indien je bij deze optie "2" invult, zal de output dus een kolom bevatten waarin voor elke hit is aangegeven welke meervoudsvorm exact werd gevonden. Vul je "1;2" in, dan wordt ook de hit voor de eerste capturing group (i.e., het exacte substantief waarvoor een meervoudsvorm werd gevonden) in een aparte kolom getoond in de output.

6 Statistieken

Een aantal relevante statistieken over de huidige versie van het DIRT-corpus zijn opgenomen in het bestand <Statistieken.xlsx>.

7 Referenties

Auer, Peter. 2005. Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations. In Nicole Delbecque, Johan Van Der Auwera & Dirk Geeraerts (eds.), *Perspectives on variation*, 7–42. De Gruyter Mouton. https://doi.org/10.1515/9783110909579.7.

Breitbarth, Anne, Melissa Farasyn, Anne-Sophie Ghyselen, Lien Hellebaut, Frederic Lamsens, Katrien Depuydt, Jesse de Does, Jan Niestadt & Koen Mertens. 2024. Gesproken Corpus van de

zuidelijk-Nederlandse Dialecten. Dutch Language Institute. https://hdl.handle.net/10032/tm-a2-z8.

Ghyselen, Anne-Sophie, Anne Breitbarth, Melissa Farasyn, Jacques Van Keymeulen & Arjan Van Hessen. 2020a. Clearing the transcription hurdle in dialect corpus building: The Corpus of Southern Dutch Dialects as case study. *Frontiers in Artificial Intelligence* 3. 1–17. https://doi.org/10.3389/frai.2020.00010.

Ghyselen, Anne-Sophie, Chelsey Deklerck, Melissa Farasyn, Timothy Colleman, Lien Hellebaut & Anne Breitbarth. 2025. Hoe betrouwbaar zijn manuele dialecttranscripties? Het GCND geëvalueerd. *Handelingen van de Koninklijke Commissie voor Toponymie en Dialectologie* 96(1). 203–227. https://doi.org/10.21825/hctd.94156.

Ghyselen, Anne-Sophie & Jacques Van Keymeulen. 2014. Dialectcompetentie en functionaliteit van het dialect in Vlaanderen anno 2013. *Tijdschrift voor Nederlandse Taal- en Letterkunde* 130(2). 117–139.

Ghyselen, Anne-Sophie, Jacques Van Keymeulen, Melissa Farasyn, Lien Hellebaut & Anne Breitbarth. 2020b. Het transcriptieprotocol van het Gesproken Corpus van de Nederlandse Dialecten (GCND). *Handelingen van de Koninklijke Commissie voor Toponymie en Dialectologie* 92(1). 83–115. https://doi.org/10.21825/hctd.88842.

Oostdijk, Nelleke. 2000. Het Corpus Gesproken Nederlands. *Nederlandse Taalkunde* 5(3). 280–284.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste & Ineke Schuurman. 2013. The construction of a 500-million-word reference corpus of contemporary written Dutch. In Peter Spyns & Jan Odijk (eds.), *Essential speech and language technology for Dutch*, 219–247. Springer. https://doi.org/10.1007/978-3-642-30910-6_13.

Van Hoof, Sarah & Jürgen Jaspers. 2012. Hyperstandaardisering. *Tijdschrift voor Nederlandse Taal- en Letterkunde* 128(2). 97–125.

Van Keymeulen, Jacques. 1993. Een verkennend taalgeografisch onderzoek naar dialectverlies in Nederlandstalig België. *Taal en Tongval* 6. 75–101.

Vandekerckhove, Reinhild. 2004. Waar zijn je, jij en jou(w) gebleven? Pronominale aanspreekvormen in het gesproken Nederlands van Vlamingen. In Johan De Caluwe, Georges De Schutter, Magda Devos & Jacques Van Keymeulen (eds.), *Taeldeman, man van de taal, schatbewaarder van de taal: liber amicorum Johan Taeldeman*, 981–993. Academia Press.

Vandekerckhove, Reinhild. 2009. Dialect loss and dialect vitality in Flanders. *International Journal of the Sociology of Language* 2009(196–197). 73–97. https://doi.org/10.1515/IJSL.2009.017.

Willemyns, Roland. 1979. Bedenkingen bij het taalgedrag van Vlaamse universiteitsstudenten uit Brussel-Halle-Vilvoorde. *Taal en sociale integratie* 2. 141–159.

Zenner, Eline, Dirk Geeraerts & Dirk Speelman. 2009. Expeditie Tussentaal - Leeftijd, identiteit en context in “Expeditie Robinson”. *Nederlandse Taalkunde* 14(1). 26–44. https://doi.org/10.5117/NEDTAA2009.1.EXPE340.

Zenner, Eline & Dorien Van De Mieroop. 2017. The social and pragmatic function of English in weak contact situations: Ingroup and outgroup marking in the Dutch reality TV show Expeditie Robinson. *Journal of Pragmatics* 113. 77–88. https://doi.org/10.1016/j.pragma.2017.02.013.