## Ronde 5

*Jolien Mathysen, Vincent Vandeghinste & Elke Peters*
*KU Leuven*
*Contact: jolien.mathysen@kuleuven.be*

# SABeD: Wat kan een corpus gesproken academisch Belgisch-Nederlands betekenen in het hoger (taal)onderwijs?

## 1. Inleiding

SABeD (*Spoken Academic Belgian Dutch*) vormt een interdisciplinair onderzoeksproject waarbij een corpus gesproken academisch Belgisch-Nederlands samengesteld wordt. Concreet heeft het SABeD-project als doel (1) een corpus gesproken academisch Belgisch-Nederlands samen te stellen, (2) hierbij de effectiviteit van spraaktechnologie te onderzoeken voor automatische transcriptie van gesproken teksten, om nadien (3) een woordfrequentielijst en, op termijn, (4) een woordenschattoets gesproken academisch Nederlands te ontwikkelen.

Tijdens deze presentatie gaan we dieper in op de noden waaruit het project ontstaan is, de theorie achter de corpusontwikkeling en hoe een dergelijk corpus ingezet kan worden voor (het ontwikkelen van *tools* voor) taalondersteuning in het hoger onderwijs.

## 2. Waarom is een corpus gesproken academisch Belgisch-Nederlands nodig/ relevant?

Het SABeD-project is ontstaan vanuit twee concrete noden in het Vlaamse hoger onderwijs. Allereerst heeft onderzoek aangetoond dat studenten, en dan specifiek internationale studenten, gesproken academische woordenschat als een struikelblok ervaren bij (het begrijpen van) hoorcolleges (Deygers 2017; Deygers e.a. 2017). Dat is problematisch, aangezien studenten hoger onderwijs veelvuldig met dit taalregister in contact komen, onder andere in hoorcolleges, en zij dus nieuwe lesinhouden moeten leren in een taalregister waarmee ze weinig vertrouwd zijn. Het hoeft dan ook niet te verbazen dat 'taalvaardigheid' en 'woordenschatkennis' significante voorspellers zijn van de slaagkansen in het hoger onderwijs (Heeren e.a. 2021a; 2021b; Milton & Treffers-Daller 2013).

Woordenschatkennis is daarnaast ook van belang voor de slaagkansen op Nederlandse taaltesten, zoals de Interuniversitaire Taaltest Nederlands voor Anderstaligen (ITNA) of het Certificaat Nederlands als Vreemde Taal (CNaVT) (Heeren e.a. 2021a; Trenkic & Warmington 2019). Zulke testen zijn verplicht voor internationale studenten die toegang willen tot het Nederlandstalige hoger onderwijs. Bovendien weten we uit onderzoek (Deygers e.a. 2018) dat studenten aangeven dat de luistertaken op zulke testen als gemakkelijker gepercipieerd worden dan de echte hoorcolleges. Concreet halen zij aan dat 'het spreektempo', 'de uitspraak of de accentkleur', alsook 'de variatie in intonatie' van bepaalde docenten de begrijpbaarheid van echte hoorcolleges kan bemoeilijken (Deygers e.a. 2018).

Een corpus gesproken academisch Belgisch-Nederlands biedt een antwoord op beide vraagstukken. Het corpus stelt ons allereerst in staat om de academische woordenschat in hoorcolleges in kaart te brengen. Ten tweede laat het corpus toe om de taal uit de luistertaken van de bestaande taaltesten te vergelijken met die uit echte hoorcolleges. Verder maakt het corpus het ook mogelijk om corpusgebaseerd leermateriaal te ontwikkelen om de studenten hun academische woordenschatkennis te helpen bijspijkeren.

## 3. Hoe maak je een corpus gesproken academisch Belgisch-Nederlands?

Door het individueel contacteren van docenten hebben we in totaal 1.028 hoorcolleges uit Vlaamse universitaire eerstejaars bacheloropleidingen verzameld. Uit deze 1.028 opnames maakten we een selectie van 200 hoorcolleges, evenredig verspreid over vier academische domeinen ('Biomedische wetenschappen', 'Exacte wetenschappen', 'Sociale wetenschappen' en 'Geesteswetenschappen'). Er werd voorrang gegeven aan hoorcolleges van minstens 30 minuten die live en *on campus* gedoceerd werden. Dit alles om het corpus zo representatief en gebalanceerd mogelijk te houden.

Voor de effectieve transcriptie van het corpus wordt er in de eerste plaats beroep gedaan op automatische spraakherkenning (ASR). De spraak in de hoorcolleges wordt dus in eerste instantie naar tekst omgezet door een reeds bestaand AI-taalmodel dat is afgestemd op Belgisch-Nederlands (Van Dyck e.a. 2021). Het model is getraind op basis van data uit het Corpus Gesproken Nederlands (CGN, Oostdijk e.a. 2002). Omwille van praktische redenen en om verschillen in lengte tussen de verschillende hoorcolleges te compenseren, wordt er per hoorcollege 30 minuten aan ASR-tekst manueel nagekeken en gecorrigeerd. Dit proces gebeurt met behulp van de ELAN-software (Wittenburg e.a. 2006). De deels manueel gecorrigeerde teksten kunnen dan weer gebruikt worden voor het verder trainen en optimaliseren van de ASR, zodat de 'spraak-naar-teksttranscripties' van steeds betere kwaliteit worden.

Het manuele correctieproces gebeurt aan de hand van een CGN-gebaseerd protocol en omvat drie grote stappen of fases.

1. Tijdens de segmentatiefase splitst men de audio-opnames op in korte, behapbare stukjes, genaamd '*chunks*'. Dat doen we om ervoor te zorgen dat manuele correcties niet op woordniveau moeten gebeuren. Zo kunnen fouten in de ASR-transcripties die over de woordgrenzen heen gaan vlot verbeterd worden.
2. Tijdens de eerste transcriptiefase wordt alles op het orthografische niveau aangepakt. Dat houdt in dat de spelling gestandaardiseerd wordt, er interpunctie wordt toegevoegd, en woorden afkomstig uit een vreemde taal en tussenwerpsels worden gesignaleerd door middel van codes. Ook spraak van studenten en persoonsgegevens ('eigennamen van docenten', 'studenten', 'vakken') die te horen zijn op de opnames worden geanonimiseerd.
3. Tijdens de tweede transcriptiefase controleren we alles op akoestisch niveau. Dat houdt in dat we 'reducties', 'dialectwoorden', 'versprekingen', 'afgebroken woorden of zinnen', 'onverstaanbare spraak en sprekersgeluiden' (zoals kuchjes of niezen) correct overnemen en aanduiden met bepaalde codes.

Na de manuele correcties passen we nog een aantal automatische processen toe. Aan de hand van FROG (Van den Bosch e.a. 2007) worden de aanwezige 'tokens' (*tokenization*), 'lemma's' (*lemmatization*) en 'eigennamen' (*named entity labeling*) geïdentificeerd, alsook 'de morfemen of woorddelen' (*morphological segmentation*) van de woorden van het corpus. Verder wordt het corpus ook voorzien van '*part-of-speech* tags' en worden de (types) relaties tussen verschillende woorden blootgelegd (*dependency parsing*). Intussen wordt ook een interface gemaakt, zodat onderzoekers het corpus zullen kunnen raadplegen via de website van het Instituut voor de Nederlandse taal (INT) en *CLARIN Virtual Language Observatory*.

## 4. Wat zijn de toepassingen van het corpus in het hoger onderwijs en wat brengt de toekomst?

Onze huidige prioriteit is om, op basis van het corpus, een woordenschatlijst gesproken academisch Belgisch-Nederlands te ontwikkelen. De voornaamste parameters voor deze lijst zijn 'frequentie' (wat zijn de meest voorkomende woorden in het corpus?), 'bereik (*range*) (welke woorden komen voor in meerdere verschillende subdomeinen/-onderdelen van het corpus?) en 'spreiding' (*dispersion*) (welke woorden komen voor in meerdere verschillende subdomeinen/-onderdelen van het corpus én komen ongeveer even vaak voor in al deze subdomeinen/-onderdelen?) (Dang e.a. 2017; Szudarski 2017). Ook zal onze basislijst vergeleken worden met een woordenlijst van Tiberius & Schoonheim (2013), die de meest frequente woorden in het (Belgisch-)Nederlands in het algemeen bevat. Zo kunnen de woorden die niet enkel frequent zijn in academisch, maar ook in algemeen taalgebruik eruit gefilterd worden. Daarnaast zullen we, aan de hand van een nieuwe termextractie*tool* van het INT, ook een onderscheid maken tussen 'eigennamen', 'domeinoverschrijdende academische woordenschat' en 'domein- of vakspecifieke woordenschat'. Naar voorbeeld van de *English Academic Spoken Word List* (Dang e.a. 2017) zal de lijst ook in sublijsten van 50 woorden opgedeeld worden naargelang woordfrequentie.

Met behulp van de woordenschatlijst kan ook een woordenschattest gemaakt worden. Deze zou nagaan of studenten de uitgesproken woordvorm en betekenis van een academisch woord kunnen herkennen. De test zou online beschikbaar zijn en een meerkeuzeformat hebben. Hij zou ook onderverdeeld zijn in secties die overeenkomen met de sublijsten van de woordenschatlijst. De eerste beschikbare versie zou voorgelegd worden aan een kleine groep van ongeveer 30 Nederlandstalige studenten voordat die onderworpen zou worden aan een grootschaliger validatieproces. Het is onze hoop dat deze hulpmiddelen breed ingezet zouden worden, niet enkel om uit te zoeken met welke academische woordenschat studenten het meest worstelen, maar ook als leidraad voor zowel studenten als docenten om zich bewust te zijn van hun academische woordenschatkennis en die, zo nodig, bij te spijkeren.

## Referenties

Dang, T.N.Y., A. Coxhead & S. Webb (2017). "The Academic Spoken Word List". In: *Language Learning*, 67 (4), p. 959-997.

Deygers B. (2017). "Validating university entrance policy assumptions. Some inconvenient facts". In: *Learning and Assessment: Making the Connections – Proceedings of the ALTE 6th International Conference*, p. 46-50.

Deygers B., K. Van den Branden & E. Peters (2017). "Checking assumed proficiency: comparing L1 and L2 performance on a university entrance test". In: *Assessing Writing*, 32 (4), p. 43-56.

Deygers, B., K. Van den Branden & K. Van Gorp (2018). "University entrance language tests: A matter of justice". In: *Language Testing*, 35 (4), p. 449-476.

Dyck, B. Van, B. BabaAli & D. Van Compernolle (2021). "A Hybrid ASR System for Southern Dutch". In: *Computational Linguistics in the Netherlands Journal*, 11, p. 27-34.

Heeren, J., D. Speelman & L. De Wachter (2021a). "A practical academic reading and vocabulary screening test as a predictor of achievement in first-year university students: implications for test purpose and use". In: *International Journal of Bilingual Education and Bilingualism*, 24 (10), p. 1458-1473.

Heeren, J., D. Speelman & L. De Wachter (2021b). "Bepaalt taal wie het haalt? De samenhang tussen een academische taalvaardigheidscreening en het behalen van een bachelordiploma aan de universiteit". In: *Tijdschrift voor Hoger Onderwijs*, 39 (1), p. 39-54.

Milton, J. & J. Treffers-Daller (2013). "Vocabulary Size Revisited: The Link between Vocabulary Size and Academic Achievement". In: *Applied Linguistic Review*, 4 (1), p. 151-172.

Oostdijk, N., W. Goedertier, F. Van Eynde, L. Boves, J.-P. Martens, M. Moortgat & H. Baayen (2002). "Experiences from the Spoken Dutch Corpus Project". In: *Proceedings of the Third International Conference on Language Resources and Evaluation*, p. 340-347.

Szudarski, P. (2017). *Corpus linguistics for vocabulary. A guide for research*. Oxon: Routledge.

Tiberius, C. & T. Schoonheim (2013). *A frequency dictionary of Dutch: Core vocabulary for learners*. Oxon: Routledge.

Trenkic, D. & M. Warmington (2019). "Language and literacy skills of home and international university students: How different are they, and does it matter?". In: *Bilingualism: Language and Cognition*, 22 (2), p. 349-365.

Wittenburg, P., H. Brugman, A. Russel, A. Klassmann & H. Sloetjes (2006). "ELAN: a Professional Framework for Multimodality Research". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, p. 1556-1559.