

SABeD: Een corpus en
woordenschatlijst
gesproken academisch
Belgisch-Nederlands
voor het hoger onderwijs



Jolien Mathysen, Vincent Vandeghinste,
Serge Verlinde, Elke Peters

State-of-the-Art

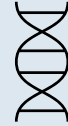
- Engels: luisterbegrip neemt toe met aantal gekende woorden (Durbahn et al., 2020; van Zeeland & Schmitt, 2013)
- 95-98% gekend → redelijk tot gedetailleerd inzicht (Dang, 2022; Dang & Webb, 2014)
- Academische woordenschat als **struikelblok voor studenten** (Deygers, 2017; Deygers et al., 2017; Deygers & Malone, 2019).
- **Problematisch:** link (academische) taalvaardigheid en success in hoger onderwijs (Heeren et al., 2021a; 2021b; Milton & Treffers-Daller, 2013; Trenkic & Warmington, 2019)

Doelstellingen

1. **Multimodaal (tekst, audio, video) corpus**
2. **Effectiviteit ASR (Automatic Speech Recognition)**
 - **Automatische transcriptie van Vlaamse hoorcolleges** en gesproken teksten
3. **Woordfrequentielijst**
4. **Woordenschattoets**

Methodologie

Stap 1:
Hoorcolleges
verzamelen



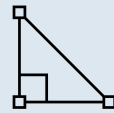
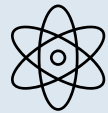
Biomedische wetenschappen

7 sprekers, 307 opnames



Geesteswetenschappen

10 sprekers, 210 opnames



Exacte wetenschappen

18 sprekers, 406 opnames



Sociale wetenschappen

21 sprekers, 98 opnames

Methodologie

Stap 2: Genereren ASR transcripties

Stap 3: Manuele correctie ASR transcripties o.b.v. protocol (cf. CGN, Oostdijk et al. 2002)

Stap 4: *Part-of-speech tagging + parsing* met FROG (Van den Bosch et al., 2007)

'Instituut voor de Nederlandse taal' + *CLARIN Virtual Language Observatory*

Woordfrequentielijst

- Frequentie, spreiding (Dang et al., 2017; Szudarski, 2017)
- Vergelijking stoplijst + frequentielijst met “A Frequency Dictionary of Dutch” (Tiberius & Schoonheim, 2013)
- Onderscheiden eigennamen, algemene academische woordenschat, domeinspecifieke woordenschat
- Sublijsten van 50 woorden volgens frequentie (Dang et al., 2017)
- Validatie

- vraag (1189)
- belangrijk (1009)
- bepaald (1009)
- element (642)
- vorm (570)
- betekenen (527)
- vragen (517)
- variabele (501)
- situatie (468)
- probleem (445)
- bepalen (439)
- relatie (415)
- functie (393)
- procent (378)
- typisch (371)
- structuur (356)
- type (349)
- normaal (335)
- effect (320)
- systeem (316)
- term (310)
- specifiek (282)
- vormen (281)
- positief (280)
- betekenis (279)
- informatie (275)
- principe (275)
- bespreken (260)

Meer info ...



<https://www.arts.kuleuven.be/ling/taalenonderwijs/projecten/sabed>



jolien.mathysen@kuleuven.be

Referenties

- Dang, T. N. Y. (2022). Vocabulary in academic lectures. *Journal of English for Academic Purposes*, 58, 101–123.
- Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The Academic Spoken Word List. *Language Learning*, 67(4), 959-997.
- Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, 33, 66–76.
- Deygers B. (2017). Validating university entrance policy assumptions. Some inconvenient facts. In Gutiérrez Eugenio, E. (Ed.), *Learning and Assessment: Making the Connections –Proceedings of the ALTE 6th International Conference* (pp. 46-50). Cambridge: ALTE.
- Deygers, B., & Malone, M. (2019). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing*, 36(3), 347–368.
- Deygers B., Van den Branden K., Peters E. (2017). Checking assumed proficiency: comparing L1 and L2 performance on a university entrance test. *Assessing Writing*, 32, 43-56.
- Durbahn, M., Rodgers, M., & Peters, E. (2020). The relationship between vocabulary and viewing comprehension. *System*, 88.
- Heeren, J., Speelman, D., & De Wachter, L. (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. *International Journal of Bilingual Education and Bilingualism*, 24(10), 1458–1473.
- Heeren, J., Speelman, D., & De Wachter, L. (2021b). Bepaalt taal wie het haalt? De samenhang tussen een academische taalvaardigheidscreening en het behalen van een bachelordiploma aan de universiteit. *Tijdschrift voor Hoger Onderwijs*, 39(1), 39-54.
- Mathysen, J., Vandeghinste, V., Peters, E., Wambacq, P. (2024). A Spoken Academic Belgian Dutch Corpus. *CLARIN2023: Selected papers in Post-Conference Proceedings*.
- Milton, J. & Treffers-Daller, J. (2013). Vocabulary Size Revisited: The Link between Vocabulary Size and Academic Achievement. *Applied Linguistic Review*, 4(1), 151–172.
- Oostdijk, N., Goedertier, W. Van Eynde, F., Boves, L., Martens, J-P., Moortgat, M., & Baayen, H. (2002). Experiences from the Spoken Dutch Corpus Project. *Proceedings of the Third International Conference on Language Resources and Evaluation*, European Language Resources Evaluation, Paris, pp. 340-347.
- Szudarski, P. (2017). *Corpus linguistics for vocabulary. A guide for research*. Routledge.
- Thompson, P., & Nesi, H. (2001). The British Academic Spoken English (BASE) Corpus Project. *Language Teaching Research*, 5, 263-264.
- Tiberius, C., & Schoonheim, T. (2013). *A frequency dictionary of Dutch: Core vocabulary for learners*. Routledge.
- Trenkic, D., & Warmington, M. (2019). Language and literacy skills of home and international university students: How different are they, and does it matter? *Bilingualism: Language and Cognition*, 22, 349–365.
- van den Bosch, A., Busser, G., Daelemans, W., & Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. van Eynde, P. Dirix, I. Schuurman, & V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting* (pp. 99–114). Centre for Computational Linguistics.
- Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied linguistics*, 34(4), 457–479.