

Pas op met gebruikersvriendelijke corpustoepassingen!

Jan Odijk

Ik heb in het kader van CLARIN-NL en CLARIAH de afgelopen jaren het creëren van gebruikersvriendelijke applicaties (bijv. GrETEL, PaQu, OpenSoNaR) om corpora te raadplegen, en ook het gebruik ervan, enorm gestimuleerd. Dit stelt veel taalkundigen in staat op eenvoudige manier gebruik te maken van informatie die bevat is in corpora. Maar het gebruik van corpora in het algemeen en het gebruik van deze gebruikersvriendelijke applicaties in het bijzonder moet wel met de nodige omzichtigheid gebeuren, anders zal het snel tot onbetrouwbare conclusies leiden, in het bijzonder wanneer het onderzoek plaatsvindt op volledig automatisch verrijkte corpora. Ook kunnen de gebruikersvriendelijkheid en het raadplegen van corpora niet altijd de verhoopte verwachtingen inlossen. Ik schets aan de hand van een aantal case studies enkele beperkingen van het gebruik van corpora, en enkele methodologische valkuilen, en hoe die vermeden kunnen worden. De meeste zullen voor taalkundigen die ervaren zijn in het raadplegen van corpora niet als een verrassing komen, maar nieuwe taalkundigen die enthousiast raken door de mogelijkheden van gebruikersvriendelijke toepassingen zoals GrETEL, PaQu, OpenSoNaR en Nederlab moeten daarover wel correct geïnformeerd worden, daarin onderwezen worden, en de gebruikersvriendelijke tools moeten faciliteiten bieden hiervoor (wat ze tot nu toe onvoldoende doen).

Ik zal een en ander illustreren aan de hand van voorbeelden onttrokken aan case studies die ik de afgelopen jaren heb uitgevoerd. Deze betreffen het modificatiepotentieel van de woorden *heel*, *erg* en *zeer* [2,3]; de complementatiepatronen van de koppelwerkwoorden *raken* en *worden* [4], normatieve en niet-normatieve varianten van een tiental constructies [1], werkwoordclusters [5], en de volgorde van werkwoorden in werkwoordclusters.

- [1] Noord, Gertjan van & Odijk, Jan (06.02.2016). *Goed of fout? Wat gebruikt men feitelijk?*. Presentation held at Grote Taaldag 2016, Utrecht.
- [2] Odijk, J.E.J.M. (14.11.2015). *Linguistic Research with PaQU*. *Computational Linguistics in The Netherlands journal*, 5, (pp. 3-14).
- [3] Odijk, J.E.J.M. (11.04.2016). *A Use Case for Linguistic Research on Dutch with CLARIN*. In Koenraad De Smedt (Eds.), *Selected Papers from the CLARIN Annual Conference 2015* (pp. 45-61) (17 p.). Linköping: Linköping University Electronic Press, Linköpings universitet.
- [4] Odijk, Jan, Noord, Gertjan van, Kleiweg, Peter & Tjong Kim Sang, Erik (28.12.2017). *The Parse and Query (PaQu) Application*. In Jan Odijk & Arjan van Hessen (Eds.), *CLARIN in the Low Countries* (pp. 281-297) (17 p.). London, UK: Ubiquity Press, DOI: <http://dx.doi.org/10.5334/bbi.23> . License: CC-BY 4.0.
- [5] Odijk, Jan, Klis, Martijn van der & Spoel, Sheean (01.01.2018). *Extensions to the GrETEL Treebank Query Application*. In Eduard Bejcek (Eds.), *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)* (pp. 46-55) (10 p.). Prague, Czech Republic: Charles University, <http://aclweb.org/anthology/W/W17/W17-7608.pdf>.