De compilatie van het Dutch C-CLAMP corpus (Dutch Corpus of Contemporary & Late Modern Periodicals)

In deze presentatie zal een recent samengesteld corpus, het Dutch Corpus of Contemporary & Late Modern Dutch, voorgesteld worden. Het corpus bestaat uit 52,931 artikels die gepubliceerd werden in Vlaamse en Nederlandse culturele tijdschriften tussen 1837 en 1999. In totaal beslaat het corpus ongeveer 200 miljoen tokens. Het eerste deel van de presentatie focust op de methodologie, de dataverzameling en het compilatieproces. In dit deel komen achtereenvolgens de tekstuele mark-up van de data, het verwerken van het materiaal en het verrijken van zowel de tekstuele data als de metadata aan bod. In het tweede deel van de presentatie worden twee tests voorgesteld die enerzijds de stabiliteit van het corpus en anderzijds de betrouwbaarheid van de PoS-tagger (Frog, Hendrickx et al. 2016) controleren. Allereerst wordt er naar het verloop van functiewoorden gekeken, met de veronderstelling dat de ratio van functiewoorden min of meer stabiel blijft doorheen de tijd. Daarnaast wordt aan de hand van een diachrone sentiment analysis (met behulp van Pattern, De Smedt & Daelemans 2012) nagegaan of de polariteit en subjectiviteit van het materiaal toe- of afneemt doorheen de tijd. Tot slot stel ik twee syntactische veranderingen voor die plaatsgevonden hebben in de 19de en 20ste eeuw, namelijk de teloorgang van het lidwoord in vaste voorzetseluitdrukkingen (zoals *onder (de) leiding van*) en de positie van het voorzetselvoorwerp bij *verliefd*. Die case studies geven de representativiteit van het C-CLAMP corpus weer en illustreren de verschillende onderzoeksmogelijkheden.

De Smedt, T. & Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research* 13: 2031–2035.

Hendrickx, I., van den Bosch, A., van Gompel, M., & van der Sloot, J. (2016). Frog, A Natural Language Processing Suite for Dutch. *Language and Speech Technology Technical Report Series*, Radboud University Nijmegen.