/instituut voor de Nederlandse taal/

Stage corpuslinguïstiek

Contactpersonen: Katrien Depuydt (taalkundige verrijkingen) en Vivien Waszink (onderzoek)

Opdracht: Het Moroccorp (samengesteld door Freek Van de Velde en Tom Ruette (KU Leuven)) is een corpus van computergemedieerde communicatie in het Nederlands door MarokkaansNederlandse taalgebruikers en het bevat tien miljoen woorden chatmateriaal. Het Moroccorp werd automatisch verrijkt met lemma's en woordsoortinformatie. De kwaliteit van die automatische linguïstische verrijkingen kan verbeterd worden met behulp van een trainings- en evaluatiecorpus, wat gecreëerd wordt door het manueel controleren van de automatisch gegenereerde annotaties.

Methode: Vaststellen van de te verrichten werkzaamheden (zie hieronder). Vertrouwd raken met het Moroccorp en de automatischeannotatiemethodes. Controleren van automatisch gegeneerde annotaties.

Mogelijke deelwerkzaamheden tijdens de stage zijn (afhankelijk van de duur van de stage en het niveau van de stagiair):

- Controle verrijkingen
- Analyse veelgemaakte fouten
- Optioneel onderzoek: nagaan of de bevindingen in het artikel "Turks- en MarokkaansNederlands" (Dorleijn et al.) ook uit het corpus af te leiden zijn.

Literatuur:

- Ide, W. & Pustejovsky J. (2017). *Handbook of Linguistic annotation*. Springer, Dordrecht.
- Teubert, W. & Čermáková, A. (2007). *Corpus Linguistics: A short introduction*. Continuum, Londen. Ruette, T. & Van de Velde, F. *Moroccorp: tien miljoen woorden uit twee MarokkaansNederlandse chatkanalen*.
- Dorleijn, M., J. Nortier, A. El Aissati, L Cornips & L. Boumans (2005), *Turks- en Marokkaans Nederlands*, in N. vd Sijs (red.),Wereldnederlands: 149-184, Den Haag: SDU