Corrigeren van lemma’s en POS-tags

Lemma’s

De bedoeling is dat aan alle Groningse woorden lemma’s worden toegekend. Daardoor wordt het mogelijk om Groningse varianten van hetzelfde woord met elkaar te koppelen. Als iemand dan bijvoorbeeld zoekt op ‘huis’, worden zowel ‘hoes’ als ‘huus’ gevonden.

Het Groningse materiaal is al gelemmatiseerd met een lemmatizer, maar dit is niet altijd goed gegaan, hier en daar zitten er fouten in. Het is nu jouw taak om die fouten op te sporen en te corrigeren. Er is een speciale app waarmee je die kunt doen. Die staat op:

https://woordwaark.housing.rug.nl/gronings/

Gebruikersnaam: woordwaark
Wachtwoord: w00rdwaark

Lees vooraf het document lemma’s.pdf aandachtig door.

POS-tags

De bedoeling is dat aan alle Groningse woorden woordsoorten worden toegekend. Dat is nodig omdat hetzelfde woord verschillende woordsoorten kan hebben. Bijvoorbeeld "eten" kan bijvoorbeeld een werkwoord zijn, maar ook een zelfstandig naamwoord. De woordsoort hangt af van de context waarin het woord voorkomt.

De Engelse benaming voor ‘woordsoort’ is ‘part-of-speech’ (POS) en een computerprogramma dat parts-of-speech toekent aan woorden heet een part-of-speech-tagger (POS-tagger).

Het Groningse materiaal is al met een POS-tagger ‘getagd’, maar dit is niet altijd goed gegaan, hier en daar zitten er fouten in. Het is nu jouw taak om die fouten op te sporen en te corrigeren. Hiervoor gebruik je dezelfde app als die je gebruikt voor het corrigeren van de POS-tags.

Om te kunnnen corrigeren moet je POS-tags kennen en weten hoe ze gebruikt worden. Lees daarom aandacht het document op https://universaldependencies.org/u/pos/all.html aandacht door evenals het document POS-tags.pdf.

N.B.: wanneer token_GN begint met ‘$’ of ‘#’ zijn lemma_NL en upos leeggelaten.