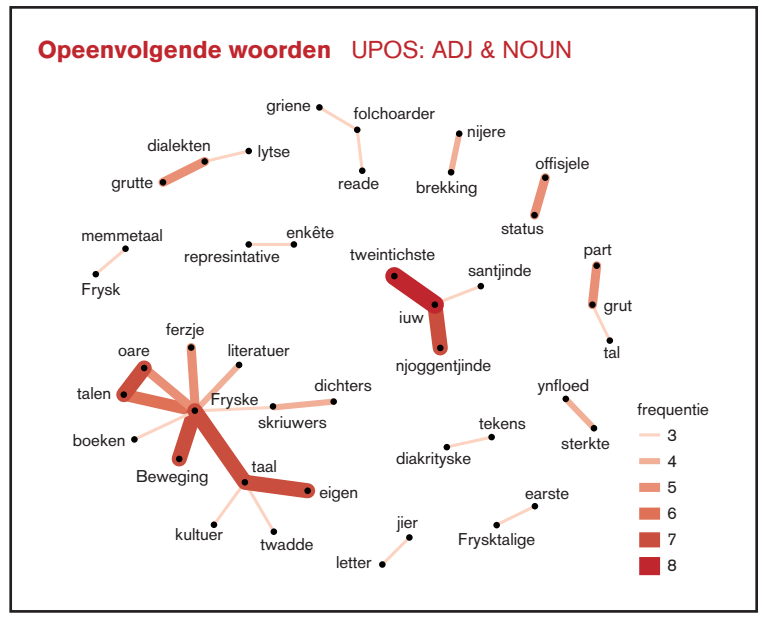
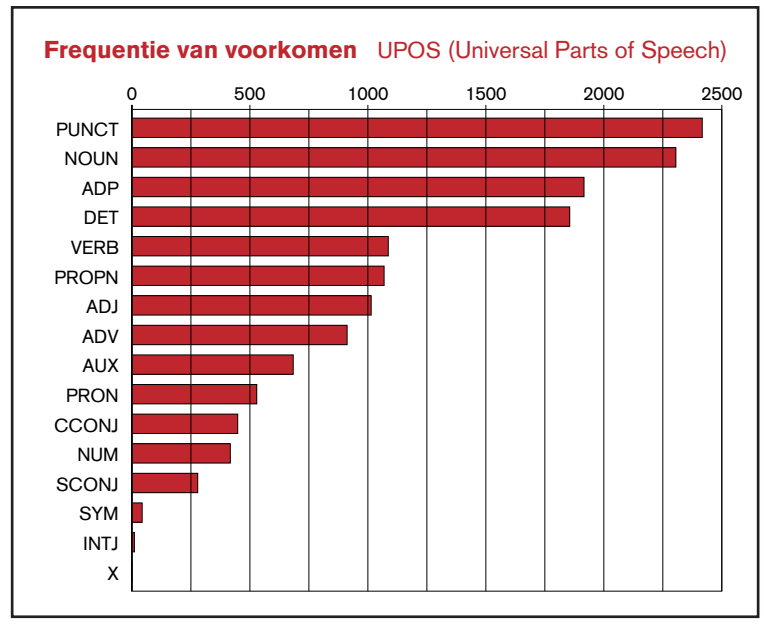
Friese teksten door online pijplijn

Tool voor taalkundig onderzoek Fries

De online tool UDPipe Frysk kent woordsoorten toe aan teksten in het Fries. Een dergelijke basistool ontbrak nog voor de tweede rijkstaal. *Erica Renckens*





Op basis van het Wikipedia-artikel 'Frysk' (nl.wikipedia.org/wiki/Westerlauwers_Fries) kan UDPipe Frysk teksten analyseren. Zo laat de rechter afbeelding zien dat zelfstandige naamwoorden (NOUN) het meest frequent zijn gebruikt, gevolgd door voorzetsels (ADP) en lidwoorden (DET). De linker grafiek laat de combinaties zien van bijvoeglijk naamwoord (ADJ) en zelfstandig naamwoord (NOUN). Uit de tekst zijn wel de titels, opschriften, bijschriften, tabellen, links en referenties weggelaten. *Credits: UDPipe Frysk*

Webapp

"Een dergelijke basistool voor taalkundig onderzoek bestond nog niet voor de tweede rijkstaal in Nederland, het Fries", vertelt Hans Van de Velde, die als projectleider aan de Fryske Akademy verantwoordelijk was voor de ontwikkeling van de tool. "POS-tags zijn belangrijk, omdat woordsoorten soms contextafhankelijk zijn. In de zin 'De bern krige iisfrij' ('De kinderen krijgen ijsvrij') is iisfrij bijvoorbeeld een zelfstandig naamwoord, maar in de zin 'De mar is hielendal iisfrij' ('Het meer is volledig ijsvrij') een bijvoeglijk naamwoord."

Onderzoekers kunnen de webapp gebruiken voor onderzoek naar bijvoorbeeld taalverandering, syntactische verhoudingen, auteursherkenning, sentiment-analyse of voor de ontwikkeling van automatische vraag-antwoordsystemen. Van de Velde: "De gebruiker typt zelf een Friese tekst in, uploadt deze of voert een Friese website in." Hierna verschijnt een tabel met voor elk token het lemma en de woordsoort. Deze output kan vervolgens in verschillende formaten (txt, excel, CoNLL-U) gedownload worden voor verdere analyse."

Wilbert Heeringa, programmeur bij de Fryske Akademy, ontwikkelde de pijplijn, samen met Gosse Bouma van de Rijksuniversiteit Groningen. Martha Hofman (Fryske Akademy) helpt bij het handmatig annoteren van het trainingscorpus. Heeringa gebruikte het project *Universal Dependencies* (UD). "Dat project ontwikkelt een universeel annotatieschema dat cross-linguïstisch vergelijken mogelijk maakt. Zo kunnen vergelijkbare constructies in verschillende talen op een consistente manier worden geannoteerd, terwijl ook taalspecifieke annotaties worden toegestaan als die nodig zijn."

1.547 zinnen

Heeringa trainde de UDPipe Frysk met 1.547 zinnen uit het Oersettercorpus. Dit corpus is in 2012 ontwikkeld voor *Oersetter*, een automatische vertaalservice voor het Fries en het Nederlands. Het bevat onder andere nieuwsberichten, romans, wetenschappelijke teksten en historisch-culturele teksten.

In de eerste update, die half mei verscheen, is dit trainingscorpus verder uitgebreid met meer zinnen. Bovendien zijn er data toegevoegd die *dependency parsing* mogelijk maken, zodat ook de grammaticale structuur van een zin met de onderlinge relaties tussen woorden in kaart gebracht kan worden. Ten slotte wordt ook gewerkt aan een analyse van de kwaliteit van de POS-tagging. UDPipe Frysk is mede tot stand gekomen dankzij financiering van CLARIAH-PLUS.

fryske-akademy.nl

Tool voor omzetten in machine-leesbare data

Historische biodiversiteit digitaal ontsluiten

Lise Stork helpt onderzoekers om moeilijk toegankelijke gegevens te verwerken met computationele technieken. Eind 2019 ontving ze hiervoor de Young eScientist Award. *Maarten Heerlien*

Winnaar van de Young eScientist Award 2019 Lise Stork helpt onderzoekers om moeilijk toegankelijke gegevens te verwerken met computationele technieken. *Credits: Thijs Stork Photography*

JONG TALENT

'Relatief eenvoudig historische manuscripten omzetten naar machine-leesbare data'

Onderzoeksinstellingen en natuurmusea herbergen veel archieven waarin gegevens over historische biodiversiteit zijn vastgelegd. Deze data zijn nog altijd relevant, maar de toepassing ervan wordt bemoeilijkt door het complexe, vaak ondoordringbare karakter van deze archieven. Lise Stork, PhD kandidaat aan het Leiden Institute of Advanced Computer Science (LIACS), combineert in haar onderzoek verschillende computationele modellen om dergelijke archieven toegankelijk te maken en zo het wetenschappelijk onderzoek te accelereren.

Digitaal vindbaar

Er zitten verschillende uitdagingen aan het ontsluiten van de informatie in het soort manuscripten dat Stork gebruikt, veelal soortbeschrijvingen en -schetsen van wetenschappers op onderzoeksexpedities in gebieden met een rijke flora en fauna. Voorbeelden van uitdagingen zijn bijvoorbeeld de kwaliteit van het handschrift, meertaligheid en verouderde terminologie. Stork gebruikt een innovatieve mix van methoden en technieken om belangrijke stukjes informatie in de manuscripten machine-leesbaar, en daarmee digitaal vindbaar te maken. "Eerst modelleer ik elementen die in de manuscripten voorkomen aan de hand van achtergrondkennis uit het domein, bijvoorbeeld taxonomie, anatomie en geografie. Vervolgens gebruik ik beeldherkenning om deze specifieke elementen - soortnamen, anatomische kenmerken en locaties - automatisch terug te vinden. Deze elementen maak ik vindbaar aan de hand van de standaarden van het kennisdomein, met behulp van semantische webtechnieken."

Naar een webomgeving

Haar onderzoek maakt deel uit van het NWO-project *Making Sense of Illustrated Handwritten Archives*. Stork heeft een workflow ontwikkeld waarmee onderzoekers relatief eenvoudig historische manuscripten kunnen omzetten in machine-leesbare data. De volgende stap is de ontwikkeling van een schaalbare en duurzame webomgeving, waarin wetenschappers historische onderzoeksarchieven betekenisvol kunnen ontsluiten. Zo kan er efficiënt door deze manuscripten worden gezocht en kunnen relaties worden blootgelegd. De ontwikkeling van deze omgeving wordt ondersteund door het eScience Center, dat in november 2019 de Young eScientist Award toekende aan Stork voor dit idee.

Mens centraal

Hoewel de focus in het onderzoek ligt op historische biodiversiteitsdata, zijn de resultaten van het onderzoek van Stork en haar medeonderzoekers breder toepasbaar: "De essentie van deze technieken en workflow is dat de mens centraal staat: we helpen mensen bij lastige keuzes door ze van de juiste informatie te voorzien en suggesties te doen, daar waar beelddata met gestructureerde, terugkerende informatie een rol speelt."

liacs.leidenuniv.nl/~storkl/