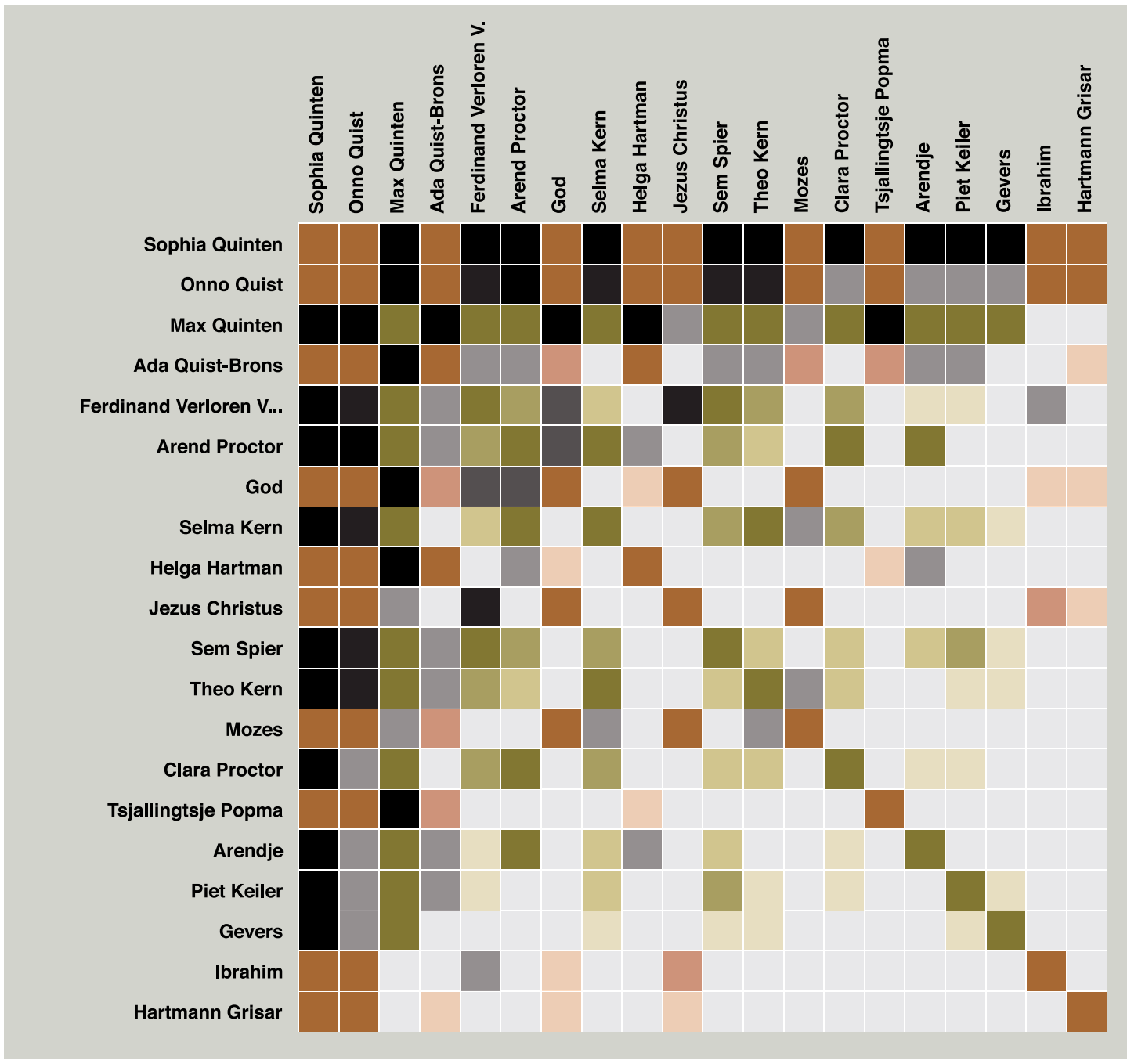
Namescape past named entity recognition toe

Letterkundig namenlandschap

Wat kunnen we leren over romans door te kijken naar het gebruik van namen? En zijn er voldoende gegevens beschikbaar voor betrouwbare uitspraken? Het Namescape-project brengt het in kaart. *Peter Boot*



Namen in Mulisch' De ontdekking van de hemel Personages die in het boek samen voorkomen, hebben dezelfde kleur. De intensiteit van de kleur is gebaseerd op de frequentie waarmee ze samen voorkomen. Zichtbaar is ook dat de naamherkenner soms personages combineert (Quinten is een afzonderlijk personage, niet de achternaam van Max of Sophia).

Namen hangen direct samen met een belangrijk aspect van de inhoud (personages en plaatsen) en belichamen ook stilistische keuzes van de auteur. Gebruik van voornamen versus familienamen zegt bijvoorbeeld iets over afstand en intimiteit. Maar voor betrouwbare uitspraken daarover moeten gegevens over een groot aantal romans beschikbaar zijn. In het onlangs afgesloten Namescape-project (een CLARIN-*demonstrator project*) is van 1.129 literaire werken in kaart gebracht welke namen er voorkomen en hoe die met elkaar samenhangen.

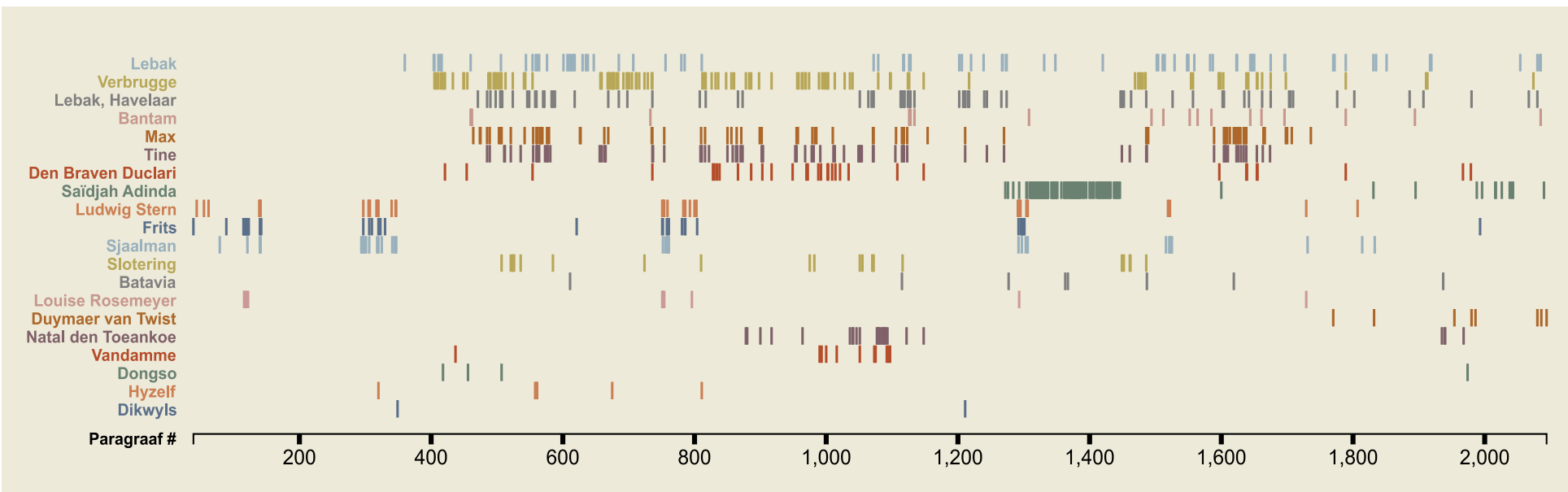
Namen in context

De *Named Entity Recognition* (NER) werd uitgevoerd door een team bij het Instituut voor Nederlandse Lexicologie (INL). Het bleek dat verhalende teksten toch wel anders zijn dan het materiaal waar NER-programmatuur meestal voor wordt ontwikkeld. Jesse de Does (INL): "De namen in literaire teksten zijn vaak nieuw, nog niet uit andere bronnen bekend. De namendichtheid is er ook anders. Voor Namescape werd gebruik gemaakt van een bestaande NER-*tagger* (de Stanford NER-tagger, nlp.stanford.edu/ner/), getraind op een speciaal ontwikkeld trainingcorpus van ongeveer 1 miljoen tokens. Maar er werd ook een eigen tagger ontwikkeld, die op dit materiaal nog iets beter presteert." De kwaliteit van de uitvoer van een statistische tagger hangt af van de hoeveelheid trainingsmateriaal, waaruit de tagger namen in context leert herkennen. De Does: "Eigenlijk hebben we een samenhangende collectie nodig van al het Nederlandstalig trainingsmateriaal voor naamsherkenning". Het INL ontwikkelde ook een web service waarmee onderzoekers hun eigen teksten kunnen laten analyseren op het gebruik van namen. Onderzoekers kunnen er bestanden uploaden, verwijzen naar een webpagina, of de service aanroepen via eigen scripts.

Visualisatie van de resultaten

Onder leiding van Maarten Marx werkten studenten aan de UvA aan de visualisatie van de resultaten. Ze ontwikkelden onder andere een 'karakterbundel' die clusters van personages en de samenhang daartussen toont, een 'matrixweergave' van welke personages voorkomen in dezelfde passage, en een 'barcodeweergave' van de personages door het boek. "Het is belangrijk dat we nu deze gegevens ter beschikking hebben", aldus Karina van Dalen-Oskam (Huygens ING), de projectleider van Namescape. Maar het project genereert ook een methodologisch resultaat. Van Dalen: "Hoe goed de taggers ook zijn, er gaat naar mijn gevoel nog erg veel mis. Letterkundig onderzoekers zijn daar niet aan gewend. Een belangrijke vervolgvraag is hoe we in geestes-wetenschappelijk onderzoek moeten omgaan met *noise*."

namescape.nl
visualizer.namescape.nl
ner.namescape.nl



Barcodeweergave van de namen in Multatuli's Max Havelaar Elk streepje in deze grafiek staat voor een alinea waarin de betreffende persoon verschijnt. De structuur van het boek is goed zichtbaar, met afwisselend hoofdstukken in Amsterdam (Stern, Frits en Sjaalman) en hoofdstukken in Indië. Ook de ingelaste vertelling over Saïdjah en Adinda is goed zichtbaar. Helaas zien we ook nogal wat ruis: de personages Lebak Havelaar, Dikwyls en Hyzelf treffen we in de roman niet aan.