# Associatieve neurale netwerken voor de humaniora

Via associaties zoeken naar relevante tekstpassages waarin de meest voor de hand liggende trefwoorden niet voorkomen. Daaraan werkt de Utrechtse vakgroep Geschiedenis van de filosofie met een zoekmachine op basis van neurale netwerken.

ERIK-JAN BOS EN PAUL ZICHE

Als wij complexe teksten lezen, doen we dat associatief: we willen bijvoorbeeld weten hoe Kant over vrijheid denkt, of op welke manier Descartes reageert op de door Hobbes geformuleerde kritiek op zijn werk. Hiervoor kan het belangrijk zijn om alle vindplaatsen te kunnen zien waar bij Kant het woord 'vrijheid' voorkomt, of waar de termen uit Hobbes' kritiek bij Descartes te vinden zijn. Maar niet alle vindplaatsen zullen relevant zijn, en het is denkbaar dat Kant inhoudelijk relevante dingen over vrijheid te zeggen heeft zonder het woord zelf, of daarvan direct afgeleide woorden, te gebruiken. Deze inhoudelijk relevante passages waarin het trefwoord zelf niet voorkomt zijn alleen *associatief*, op basis van semantische criteria te vinden.

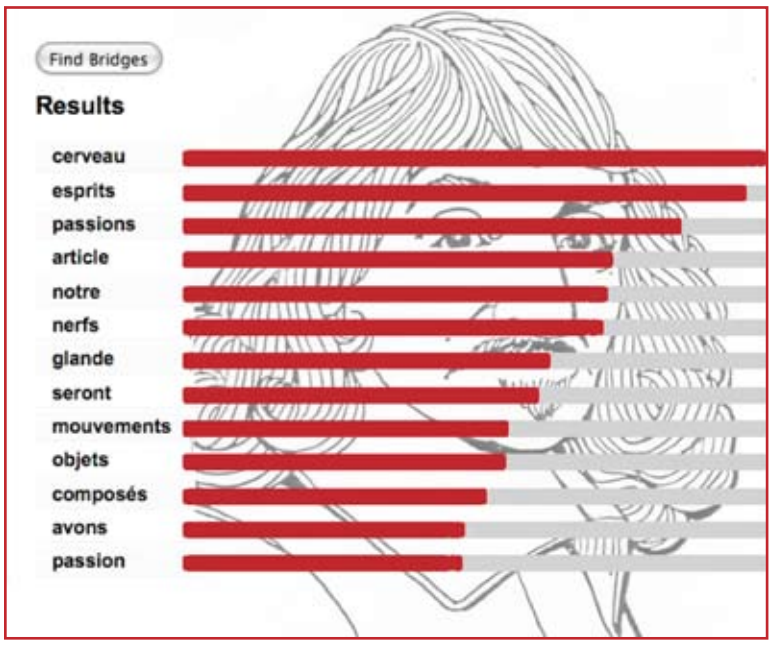
Dit idee ligt ten grondslag aan een zoekmachine die gebaseerd is op *associatieve neurale netwerken voor de humaniora (ANNH)*, en die door de vakgroep Geschiedenis van de Filosofie aan de Universiteit Utrecht wordt toegepast op wijsgerig-historisch onderzoek. In die netwerken interageren neuronen met elkaar op een manier die ontleend is aan het hersenonderzoek. In de hersenen zijn zenuwcellen - neuronen - met elkaar verbonden via synapsen. Op dezelfde manier zijn de onderdelen van een tekst met elkaar verbonden via de woorden in die teksten. Semantisch informatieve verbindingen zijn sterker dan andere, net zoals in de hersenen; op basis van de aan elk element van een tekst toegeschreven informatiewaarde, kunnen associaties tussen de tekstelementen worden berekend die voor inhoudelijk relevante resultaten zorgen.

Een belangrijk verschil met bestaande systemen om grote hoeveelheden tekst te analyseren is dat dit systeem geen informatie vooraf nodig heeft over de te doorzoeken teksten. Er ligt geen woordenlijst, thesaurus of expertsysteem aan ten grondslag. Het systeem kent zelf een relatieve informatiewaarde toe aan elk woord en maakt op basis daarvan associaties, afhankelijk van de teksten die doorzocht worden. Een zoekopdracht binnen één werk van een auteur levert dus heel andere associaties op dan diezelfde zoekopdracht binnen zijn gehele oeuvre. Het is ook denkbaar dat het systeem passages als resultaat levert waarin de oorspronkelijke zoekterm niet voorkomt, maar die toch op een semantisch niveau met de zoekterm verbonden zijn.

De zoekmachine is een belangrijke aanwinst voor de tekst- en editiewetenschap. Traditionele vragen uit dit vakgebied, die tot op heden alleen na bijzonder arbeidsintensief onderzoek beantwoord konden worden, kunnen nu snel worden opgelost. Wie is de auteur van een tekst die zonder auteursnaam werd overgeleverd? Kunnen we iets zeggen over de datering van een ongedateerde tekst? Door zulke teksten associatief te relateren aan teksten waarvan auteur en datum wel bekend zijn, levert ANNH hele belangrijke informatie over auteurschap en datering.

Daarenboven biedt ANNH een nieuwe, innovatieve benadering om (al dan niet wijsgerige) conceptuele systemen te interpreteren. Welke concepten verbinden twee zoekwoorden op associatieve manier (zie illustratie)? Welke stukken in het werk van een denker hangen met welke andere stukken samen? Zijn er bijvoorbeeld binnen één werk heterogene elementen te ontdekken die in verband staan met verschillende andere teksten in het oeuvre van de auteur? Of, kleinschaliger geargumenteerd: hoe groot is de mate van variatie in het taalgebruik van een filosoof? Leiden associatieve verbindingen steeds weer tot een relatief klein aantal van basistermen, of breidt zich het vocabulaire van een wijsgeer uit tot de meest exotische vertakkingen?

Op dit moment zijn de volledige werken van de filosofen Descartes, Schelling, Kant, Fichte, Schopenhauer, en andere belangrijke tekstbestanden uit de zeventiende eeuw aan het systeem toegevoegd. Ook bestaan er interfaces die het gemakkelijk mogelijk maken om nieuwe zoekfuncties te kunnen ontwikkelen, en nieuwe tekstbestanden aan het systeem toe te voegen.




BEELDBEWERKING ELLEN BOUMA

**Het resultaat van een zoekactie op geassocieerde woorden met het woordenpaar 'âme' (ziel) en 'matière' (materie) in het werk van Descartes.**

# Naar een betere toegang tot tekst- en spraaktools

MARC KEMPS-SNIJDERS EN INEKE SCHUURMAN

Het Nederlands-Vlaamse project 'Taal- en spraaktechnische Tools voor het Nederlands als Webservices in een Workflow (TTNWW)' maakt technische faciliteiten op het gebied van tekst en spraak toegankelijk voor een brede groep onderzoekers in de geestes- en sociale wetenschappen met geringe technische bagage. Daardoor moeten ze hun onderzoeksvragen beter en gemakkelijker aan kunnen pakken en vragen kunnen stellen die tot dusverre niet te beantwoorden waren.

TTNWW sluit aan bij CLARIN, een Europees project om een geïntegreerde en interoperabele infrastructuur te bouwen die onderzoekers eenvoudig toegang geeft tot taalkundige bronnen. Die bronnen kunnen geannoteerde mediabestanden zijn (vooreerst spraak en tekst), lexica en ontologieën, of het kan om technieken gaan zoals verwerkt in spraakherkenners, lemmatizers en parsers. Zulke bronnen, en de technieken om ze toegankelijk te maken, zijn voor het onderzoek van groot belang. Bij het beschikbaar maken ervan zijn vaak aanzienlijke kosten gemaakt voor het verzamelen, digitaliseren en annoteren van bijvoorbeeld grote tekst- of spraakcorpora – werk dat grote vaardigheid en expertise vereist.

Die bronnen en technologie kunnen alleen volledig worden benut als ze op grote schaal eenvoudiger worden ontsloten met een robuuste en duurzame infrastructuur, zodat de onderzoekers niet zelf over (taal)technische knowhow hoeven te beschikken. Zo'n infrastructuur behelst de opslag en toegankelijkheid van data en services op grote schaal, het onderling aansluiten van dataformaten en het vermogen van de gebruikte tools om onderling samen te werken. Hierbij kan gebruik worden gemaakt van middelen die beschikbaar worden gesteld door samenwerkende instituten uit heel Europa. Zo wordt voor onderzoekers de mogelijkheid geopend om een betere keuze te maken uit de beschikbare data en technieken om hun onderzoeksvragen te beantwoorden. Verschillende nationale initiatieven, waaronder CLARIN-NL en CLARIN-Vlaanderen, dragen inmiddels bij aan de realisatie van deze ideeën.

Naast CLARIN-NL wordt het project gefinancierd door het Vlaamse departement Economie, Wetenschap en Innovatie. In de afgelopen tien jaar was er al een nauwe samenwerking tussen Nederland en Vlaanderen op het vlak van taal- en spraaktechnologie. Die kreeg vorm in het project Corpus Gesproken Nederlands en daarna in het bilaterale STEVIN programma voor taal- en spraaktechnologie. Deze samenwerking wordt nu voortgezet in het TTNWW-project. Naast het toegankelijk maken van faciliteiten voor onderzoekers is een belangrijk doel om de in Nederland en Vlaanderen gangbare de facto standaarden voor dataformaten en interfaces tussen tools en protocollen te promoten en te toetsen aan Europese initiatieven zoals CLARIN.

Voor de tekstkant worden veel verschillende tools, waaronder voor alignering, automatische naamherkenning en coreferentie gecombineerd tot standaard workflows voor de beantwoording van onderzoeksvragen over historische teksten, forumdiscussies, vertaalde romans en archeologische data. Voor de spraakkant wordt gebruik gemaakt van spraakherkenners voor het ontsluiten van gesproken data op basis van toegankelijke en doorzoekbare tijdsynchrone transcripties.

Deelnemers aan TTNWW zijn de universiteiten van Tilburg, Antwerpen, Groningen, Utrecht en Twente, de Katholieke Universiteit Leuven, de Radboud Universiteit, Hogeschool Gent, het Max Planck Instituut voor Psycholinguistiek, het Instituut voor Nederlandse Lexicologie, het Huygens Instituut, het Katholiek Documentatie Centrum, het Katholiek Documentatie- en Onderzoekscentrum voor Religie, Cultuur en Samenleving en Aletta.

www.clarin.eu
http://taalunieversum.org/taal/technologie/stevin