

Gaandeweg werd ik opgezogen door de computerlinguïstiek

Rens Bod, voormalig Vidi-laureaat, ontving afgelopen december een Vici-subsidie voor zijn onderzoek *Integrating Cognition: Unsupervised Learning with the DOP model*. Rens Bod (1965) is verbonden aan het Institute for Logic, Language and Computation, Universiteit van Amsterdam. Daarnaast is hij hoogleraar Kunstmatige intelligentie aan de University of St Andrews (GB).

U hebt een deel van uw studie in Italië gevolgd. Op uw website schrijft u dat het Italiaanse onderwijsstelsel "gave me a truly humanist background on which I can still build today."

Toen ik midden jaren tachtig in Rome studeerde kon ik alle vakken kiezen die ik maar wilde zonder me zorgen te maken over een coherent curriculum. Zo studeerde ik dwars door elkaar heen wiskunde, muziekwetenschappen, filosofie, kunstgeschiedenis en natuurkunde. Dit bedoel ik met "truly humanist background": een brede opleiding die het woord "universitas" eer aan doet en die me zowel een alfa- als een beta-achtergrond heeft opgeleverd. Ondertussen is dat niet meer mogelijk: je kiest een bachelor met een voorgeschreven programma. Dwarsverbanden tussen de humaniora en exacte wetenschappen zijn er nauwelijks, een enkele studierichting als computerlinguïstiek daargelaten.

Hoe kwam u vanuit Italië naar de computerlinguïstiek in Amsterdam?

Ik kwam in 1988 naar Nederland en wilde me richten op wat ik noemde 'computationele stilistiek'. In Italië was ik geïnteresseerd geraakt in computationele methoden voor (visuele) stijlherkenning. Echter, ik kon niemand vinden die daar mee bezig was. Totdat ik ontdekte dat er zoiets bestond als Alfa-Informatica, de studie van informatica in de context van de alfa-wetenschappen. Prof. Remko Scha was iemand die dat vak echt serieus nam, en hij wist me te overtuigen van het belang van computationele linguïstiek. Aanvankelijk wilde ik me zo snel mogelijk aan die stilistiek wijden, maar gaandeweg werd ik opgezogen door de computerlinguïstiek. Bovendien was het een spannende tijd: er was in de computerlinguïstiek een soort revolutie gaande waar statistische in plaats van deterministische grammatica's werden gepropageerd en waarbinnen Amsterdam een aanzienlijke rol speelde. Dit soort

grammatica's kon ook worden uitgebreid naar muziekwaarneming en visuele perceptie, waarbij ik weer uitkwam bij mijn aanvankelijke interesse.

In 2000 ontving u een Vidi-subsidie. Wat wilde u onderzoeken en wat hebt u gevonden?

Tijdens mijn promotieonderzoek heb ik een data-georiënteerd model van taalverwerking ontwikkeld ('DOP') dat gebruik maakt van concrete taalervaringen in plaats van grammaticale regels. Nieuwe taaluitingen konden worden geanalyseerd aan de hand van zo groot en zo frequent mogelijke fragmenten van eerdere taaluitingen. Grammaticale regelmatigheden emergeerden uit de data zonder dat deze expliciet geformuleerd hoefden te worden. De vraag deed zich voor in hoeverre dit model toepasbaar was op andere vormen van menselijke perceptie. Dat was wat ik heb onderzocht tijdens mijn Vidi. Ik heb een overkoepelend model ontwikkeld dat kan worden ingezet voor zowel taalanalyse, muzikanalyse als automatisch redeneren.

Eind 2006 ontving u een Vici-subsidie. Hoe sluit deze aan op uw Vidi-onderzoek?

Een grote uitdaging voor datageoriënteerde modellen is het acquisitieprobleem, niet alleen in taal maar ook in andere modaliteiten. Als ik me even mag beperken tot een taalvoorbeeld: hoe leren kinderen regelmatigheden zoals de inversie bij vraagzinnen, en waarom doen ze dit bijna altijd correct? Er zijn grofweg twee visies in de hedendaagse taalkunde, een 'nativistische' die aanneemt dat veel taalkennis is aangeboren, en een 'empiristische' die er vanuit gaat dat alle taalkennis uit linguïstische data kan worden geleerd. Ik heb recentelijk een derde visie onderzocht, namelijk dat regelmatigheden kunnen emergeren uit een statistisch matchingproces zonder dat deze in de data hoeven voor te komen en zonder dat we ze als aangeboren hoeven te veronderstellen. Een van de doelstellingen van het huidige Vici-project is te onderzoeken of een nieuw ongesuperviseerd DOP model (U-DOP) het taalacquisitie-probleem kan oplossen. Preciezer gezegd: het project beoogt een generalisatie van een gesuperviseerd naar een ongesuperviseerd leermodel.

Waarom betreft u ook muziek in het onderzoek? Is taal alleen niet al complex genoeg?

Uit mijn eerdere onderzoek is gebleken dat inzichten uit de ene modaliteit (bijvoorbeeld muziek) van belang kunnen zijn voor het modelleren van een andere modaliteit (bijvoorbeeld taal of redeneren). Er zijn ook aanwijzingen uit de neurowetenschappen en de cognitieve psychologie dat muziek- en taalverwerking gebruik maken van gemeenschappelijke neurale processen. De uitdaging is om deze processen computationeel te modelleren zodat er concrete voorspellingen mee kunnen worden gedaan. Tijdens mijn Vidi heb ik een model ontwikkeld dat voor drie verschillende modaliteiten de beste analyse kan voorspellen. Voor mijn Vici hoop ik aan te tonen dat we ook een algemeen model kunnen ontwikkelen dat het leerproces beschrijft. Hoewel ik me concentreer op het leren van taal, wil ik het ook testen voor muziek en redeneren.



Het U-DOP model zal getest worden aan de hand van concrete toepassingen. Hoe gaat dat in zijn werk?

Deze concrete toepassingen zijn ondermeer spraakherkenning, machinaal vertalen, pitch spelling (muziek) en probleem-oplossen. We maken gebruik van internationale 'benchmarks' waar het grootste deel van de wetenschappelijke wereld op test. Zo is het gebruikelijk om bij automatisch vertalen te testen op het zgn. Europarl corpus (teksten van het Europees parlement). Er is de laatste 10 jaar een groot aantal modellen op de 'markt' gekomen die net

als DOP probabilistisch van aard zijn. En hoewel er meetbare vooruitgang is in nauwkeurigheid, blijft het nodig om nieuwe modellen te ontwikkelen. U-DOP is zo'n nieuw leermodel dat zijn sporen heeft verdiend op het gebied van automatisch leren van syntactische structuur. De uitbreiding naar ondermeer automatisch vertalen zie ik als een spannende testcase voor een concrete toepassing.

Het NWO Geesteswetenschappen is van mening dat al het onderzoek, ook fundamenteel, nieuwsgierigheidgedreven onderzoek, mogelijkheden voor benutting kent en dat het van belang is die mogelijkheden te identificeren en als het kan ook te benutten. Hoe staat u daar tegenover?

Ik kan niet overzien of al het onderzoek mogelijkheden voor benutting kent, maar bij mijn onderzoek is dat inderdaad het geval. We moeten echter niet vergeten dat het omgekeerde ook vaak het geval is: veel nuttigheidsgedreven onderzoek heeft direct weerslag op fundamenteel onderzoek. In mijn eigen vakgebied is dit recentelijk gebeurd op het gebied van statistische grammatica's. Aanvankelijk werden zulke grammatica's alleen bestudeerd vanuit de toepassingsgerichte computerlinguïstiek en spraaktechnologie, maar sinds een paar jaar verschijnen er ook artikelen en debatten in tijdschriften als *Language* over het belang van statistische grammatica's voor de taaltheorie. Dus hoewel ik het met bovenstaande mening eens ben, kan deze ook worden omgedraaid: al het toepassingsgerichte onderzoek kent mogelijkheden voor fundamentele reflectie.

Wat komt er na het Vici-onderzoek? Wat zijn belangrijke onderzoeksvragen voor de toekomst?

Stel dat het ons lukt om een computationeel model van taalacquisitie te ontwikkelen dat ook bruikbaar is voor enkele andere modaliteiten. Dan doet de vraag op: kan dit model gegeneraliseerd worden naar menselijke cognitie in het algemeen? Dit brengt ons bij Allen Newell's "Grand Challenge": bestaat er een unificerend model voor alle menselijke cognitie? Er zijn recentelijk initiatieven op dit gebied ontplooid, zoals het National Initiative on Brain & Cognition die bijzonder veelbelovend zijn.