Formaliseren in de taalkunde

Anneke Neijt, Radboud Universiteit

Samenvatting

Crisis in de taalkunde? Eerder een stelling dan een vraag voor van der Horst en Van de Velde. Ze onderbouwen de stelling door te verwijzen naar de dalende studentenaantallen en de groeiende invloed van gamma- en bètaonderzoek. Oplossingen? Veranderingen in het curriculum en weg met dat vooroordeel 'voor wie niets heeft met formele taalkunde'. Onderzoek de samenhang van de uiteenlopende componenten van het taalsysteem en de vraag hoe data en theorie samen kunnen gaan. Hieronder een schets van de verrassingen die volgen uit een nauwgezette beschrijving van het alfabetische schrift. De toekomst van de taalkunde ligt in toetsing van hypotheses en, zoals van der Horst en Van de Velde zelf ook al concluderen, in interdisciplinaire samenwerking.

Abstract

Crisis in linguistics? is a statement rather than a question for van der Horst and Van de Velde, substantiated by referring to student numbers and the growing influence of gamma and beta research. Solutions? Curriculum change and get rid of that prejudice 'for those who have nothing to do with formal linguistics'. Investigate the relations between the various components of the language system, and develop models that combine data and theory, as exemplified by a meticulous description of alphabetic writing. The crisis disappears for those who choose formal approaches and aim, as van der Horst and Van de Velde themselves conclude, at interdisciplinary teams.

1. Modelleren en formaliseren

Taalkundigen zijn meesters in het verzinnen van theorieën, dat blijkt wel uit het targetartikel van dit nummer. De meeste daarvan geven overzichten, zonder zich om de details te bekommeren, of laten mooie generalisaties zien op deelgebieden van het taalsysteem. Het wordt tijd om aan de vraag te beginnen hoe een en ander in een groter geheel past en aansluit op het moois dat de anderen leveren. Uiteindelijk zal er toch een verklarende beschrijving moeten overblijven van hoe de mens taal verwerft en gebruikt.

Verslagen en Mededelingen van de KANTL, vol. 130, nr. 1, 2020; 69-75

De weg daarnaartoe gaat via omtrekkende bewegingen. Bijvoorbeeld, in de jaren 80 is er binnen het project Eurotra gewerkt aan een vertaalcomputer. Het idee achter Eurotra was om voor elk van de toen vijftien talen regels op te stellen om in een universeel formaat de betekenis van uitingen weer te geven. Dus van concrete vormen naar een abstracte betekenisrepresentatie die als brug kon dienen voor het vertalen. Zoals de mens dat doet, zoals vertalers hun vertaling maken via het begrijpen van de brontaal.

Het Eurotra-project is jammerlijk mislukt, maar het heeft wel nuttige inzichten opgeleverd. Bijvoorbeeld het inzicht dat modellen die gebaseerd zijn op regels veel en veel te traag zijn. Over het vertalen van een ondubbelzinnige zin van twaalf woorden deden computers destijds een etmaal. Zelfs als elke stap in zo'n regelsysteem maar een fractie van een seconde zou kosten, dan nog evenaart het computermodel niet de snelheid van de mens. Bovendien, met ambiguïteit, modaliteit en ellipsen kon het systeem niet omgaan. Zinnen zoals '*To be or not to be'* werden uiteindelijk woord voor woord vertaald in '*Naar zijn of niet naar zijn'*. Dat doet Google Translate inmiddels beter, waarschijnlijk gewoon omdat de vertaling 'hard' is opgenomen, dus zonder dat het vertaalsysteem een inkijkje geeft in hoe menselijke vertalers het doen.

De invalshoek van het modelleren van taal blijft onderbelicht in het target-artikel. Modelleren vereist formalisatie en daarmee gaan van der Horst en Van de Velde zuinigjes om. Ze doen de 'micro-comparatieve benadering' (Van Craenenbroeck & Van Koppen, 2019) af met 'maar de micro-comparatieve benadering is theorie-intern, en heeft weinig te bieden voor wie niets heeft met formele taalkunde'. *Theorie-intern* klinkt denigrerend, terwijl er natuurlijk niks mis is met een aanpak waarin theorie-interne validiteit een rol speelt – elke theorie dient intern valide te zijn. En je moet kijken naar wat daarbuiten gebeurt. Vroeger of later zul je zulke uiteenlopende takken van de taalkunde als fonologie en pragmatiek via een of andere formalisatie met elkaar in verband moeten brengen.

2. DATABESTANDEN, LABORATORIA, GELD

Wat zeker niet gebeuren mag, is toekijken hoe 'niet-taalkundigen gretig gebruikmaken van gegevens die moeizaam verzameld zijn door taalkundigen en vervolgens de eer gaan opstrijken met in het oog springende publicaties, en methodes en statistiek gebruiken die veel taalkundigen vreemd zijn.' Het klopt ook niet dat taalkundigen het een doen (data verzamelen) en niet-taal-

kundigen het ander (onderzoek doen met die data), dus goed dat het targetartikel dat misverstand aan de kaak stelt.

Het ontwikkelen van databestanden gebeurt meestal in teamverband, op initiatief van iemand (taalkundige of niet) die de noodzaak van zo'n bestand onderkent, en de financiën weet te verwerven om eraan te gaan werken. Geld is echt een probleem bij het ontwikkelen en onderhouden van databestanden. En voor de verdere ontwikkeling van het vak. Er zijn laboratoria nodig om experimenten te kunnen doen met apparatuur die vele malen meer kost dan pen en papier. Er is alleen daarom al samenwerking nodig tussen alfa, bèta en gamma.

De ontwikkelingen in de richting van interdisciplinair teamwerk zijn al een eeuw aan de gang. Het is misschien een wat ongemakkelijke ontwikkeling voor onderzoekers met veel talent voor monodisciplinair onderzoek. Vruchtbare nieuwe inzichten zijn immers in het verleden onmiskenbaar ontwikkeld door wat gekscherend het leunstoelonderzoek wordt genoemd. Die tijd komt niet terug. Voor de taalkunde geldt, zoals voor elke empirische wetenschap: '*Out of the armchair, into the fire!*' Het is tijd om theorieën te toetsen.

3. De koppeling van data en regels

Het succes van systemen die van grote databestanden gebruikmaken, illustreert dat opslag van gegevens in het taalsysteem een grote rol speelt. Er zijn tevens aanwijzingen dat taal gebaseerd is op regels. Hoe koppel je data en regels? En hoe koppel je de verschillende lagen van het taalsysteem (ieder met een eigen interne systematiek waarvoor je specialistische kennis nodig hebt), nu gebleken is dat dat niet kan gebeuren door chomskyaanse transformaties?

Misschien gaat het wel om directe koppelingen van de lagen van het taalsysteem via de data, en is afzonderlijk daarvan ergens in het menselijke brein plaats voor regelmatigheden. Dat idee is gebaseerd op het onderzoek van Johan Zuidema naar het verband tussen de fonologie, de morfologie en de spelling van het Nederlands.

Toegegeven, spelling is niet, zoals gesproken taal, een natuurlijk verworven deel van de taal, maar de spelling lijkt bij volleerde gebruikers wel dezelfde status te hebben. Op een gegeven moment zijn schrijver en lezer zich niet meer bewust van regels, maar maken ze buiten de regels om koppelingen tussen klanken en letters. Dat kun je internaliseren noemen (voor de spelling meer precies omschreven in het 12321-model, zie Neijt, Peters en Zuidema,

2012). Een goede beschrijving van de spelling levert inzicht op in hoe dat precies verloopt. Zuidema heeft de afgelopen ruim tien jaar zo'n beschrijving ontwikkeld (ja, in de woorden van van der Horst en Van de Velde: 'moeizaam verzameld'). Een voorbeeldige combinatie van spellingdata en spellingtheorie.

In Zuidema en Neijt (2017) beschrijven we Zuidema's BasisSpellingBank (BSB), zo genoemd omdat het systeem de woorden bevat van de Basisspellinggids (Cranshoff en Zuidema, 2005/2010) en omdat het verwijst naar de spellingregels die daar uitgelegd worden. Van oorsprong beoogt de BSB een hulpmiddel te zijn voor wie de spelling moet leren, maar het systeem blijkt verbluffend interessante eigenschappen te hebben, interessant ook voor taalkundigen die gewoonlijk met een ruime boog om de spelling heen lopen.

3.1. HOE DE BSB DATA EN REGELS SAMENBRENGT

De bouwstenen van een realistisch systeem hoeven niet precies overeen te komen met de bouwstenen die taalkundigen aandragen. Bijvoorbeeld, in de fonetiek levert difoonsynthese vloeiender spraak op dan allofoonsynthese. Je koppelt dan de tweede helft van een klank aan de eerste helft van de volgende klank. In de BSB zijn de bouwstenen eveneens grootheden die in de theoretische taalkunde niet voorkomen, namelijk tripletten van letters, klanken en informatie over hun relatie. Drielagige bouwstenen dus: de spellinglaag, de morfofonologische laag, en een code die verwijst naar de beschrijving van de relatie tussen die twee lagen inclusief de volgorde van verwerving.

Neem het woord *hellinkje*. Dat omvat de drie tripletten *h, ell* en *inkje*. Het eerste triplet is het eenvoudigst 'h;h;1_4'. De puntkomma's scheiden de drie delen van het triplet. In dit geval geeft het triplet aan dat de letter *h* bij de klank *h* hoort volgens een regel met de code 1 (de code voor een eenduidige letter-klankverbinding; de koppeling die je het eerst leert) en subcode 4 (de *h* is niet de allereenvoudigste letter-klankverbinding). Meer dan dit valt er over de *h* van *hellinkje* niet te vertellen.

Het tweede triplet is iets ingewikkelder 'e#l=l;E=l;4.2c'. Het isgelijkteken staat voor een lettergreep- of syllabegrens. In de spelling 'e#l=l' ligt de lettergreepgrens op een andere plaats dan de syllabegrens, aangegeven met #. In de uitspraak 'E=l' gaat het om een beklemtoonde korte *e* (E onderstreept, hoofdletters voor korte klinkers in fonetisch schrift), een syllabegrens en één *l*. De regel die je moet leren heeft code 4 (verdubbeling na een korte klinker) met subtype 2c, nodig om eenvoudige en wat ingewikkelder verdubbelingsgeval-

len van elkaar te onderscheiden. In dit geval verwijst de code naar het achtervoegsel *-ing*. Naast dit alomvattende triplet zijn er tripletten voor de afzonderlijke delen, zoals ‘e;E;1_3’. De *e* van *hellinkje* is een spellingzuivere klinker die al vroeg wordt aangeleerd.

Het derde triplet omvat de rest van het woord, omdat de spelling van de twee achtervoegsels *-ing* en *-tje* in elkaar grijpt: ‘in=kje;IN=kj@;6.5&2.9&9.2’. Code 6.5 voor het achtervoegsel *-ing,* code 2.9 voor de digraafspelling *nk,* en code 9.2 voor het achtervoegsel *-tje*. Naast dit complexe alomvattende triplet zijn er weer tripletten die kleinere letter-klankkoppelingen beschrijven, zoals ‘e;@;2.1_2’. Geen spellingzuivere koppeling voor de sjwa.

3.2. Een voorbeeldige aanpak

De BSB beoogt te beschrijven hoe en in welke volgorde de mens leert schrijven. Het bestand laat zien hoe data en een theoretische beschrijving samen kunnen gaan. Vanuit de koppeling van klanken en letters wordt via codes een verband gelegd met kennis van de fonologie en morfologie van het Nederlands en de spellingregels die taalgebruikers al dan niet volkomen geleerd hebben. Haal je de codes die naar de regels verwijzen weg, dan is er een directe koppeling van klanken en letters, iets wat het geval moet zijn gegeven de snelheid van lezen en schrijven. Het BSB-model heeft vanuit dit perspectief een hogere realiteitswaarde dan derivationele modellen.

De aanpak van de BSB verdient navolging op een aantal punten: (a) de verschillende lagen van het taalsysteem zijn gekoppeld, (b) verwervingsstadia zijn erin verwerkt, (c) data en de beschrijving daarvan zijn gescheiden, maar toch met elkaar in verband gebracht en (d) als je het derde deel van een triplet weglaat, heb je een directe koppeling van letters en klanken, zoals de volleerde taalgebruiker zonder zich om de regels te bekommeren met spelling omgaat. Onduidelijk is nog of de bouwstenen intact blijven als je bij het beschrijven van de koppeling niet vanuit de schrijver, maar vanuit de lezer redeneert.

3.3. De winst

Je zou kunnen denken dat de BSB met die minutieuze aanpak van tripletten een onoverzichtelijk geheel is gaan vormen. Hoe grillig is immers de spelling van het Nederlands! De BSB omvat voor ongeveer 100.000 woorden (vervoegingen en verbuigingen van de woorden van de Basisspellinggids) inderdaad

heel veel tripletten, een tokentelling van meer dan een miljoen, maar een typetelling van 'slechts' 5000, netjes in een zipfiaanse verdeling: 30 tripletten voor de helft van de set, ruim 150 tripletten voor 80% van de set en 4000 tripletten voor 5%, de uitzonderingen. Generaliseren we over tripletten (het triplet dat verdubbeling van de letter *l* in het woord *helling* betreft, lijkt op het triplet van verdubbeling van de *k* in *takken*), dan blijkt het om een bestand van nog geen 200 tripletten te gaan. Uiteindelijk dus een overzichtelijk aantal. Zou zoiets niet ook ontwikkeld kunnen worden voor de lagen van het taalsysteem die dichter liggen bij de syntaxis en semantiek? Er zijn deelgebieden van de taalkunde waarvan we inmiddels heel veel weten, zoals tijd, aspect, modaliteit, kwantoren, valentie, enz.

4. Tot slot

De taalkunde is een empirische wetenschap die theorievorming doet op basis van waarneming. Daarbij zouden de hoeveelheden data die nu beschikbaar zijn en de inbreng van gamma en bèta als aanwinst ervaren moeten worden, niet als crisis. En die crisis is er ook niet voor wie 'iets heeft' met formaliseren en modelleren, want daarmee kun je de scheiding tussen de disciplines overbruggen. Die crisis is er ook niet bij de onderzoeksinstituten, want daar gebeurt het al, interdisciplinaire samenwerking. Daling van studentenaantallen los je er niet mee op. Daar heb je veranderingen in het curriculum van het voortgezet onderwijs voor nodig, plus in Nederland een ander eindexamen. Of veranderingen in het universitaire onderwijs, met taalkunde in het vakkenpakket van bepaalde gamma- en bètaopleidingen. Want waarom zou je taal slechts vanuit het alfa-perspectief willen bestuderen?

Noot: Johan Zuidema, psycholoog, onderwijskundige en lexicograaf van huis uit, wil ik bedanken voor zijn opmerkingen bij deze tekst, en voor zijn inbreng vanuit andere disciplines.

Literatuurlijst

Cranshoff, B. & Zuidema, J. (2005/2010). *Van Dale Basisspellinggids*. Utrecht, Antwerpen, Tilburg: Van Dale, Zwijsen.

Neijt, A, Peters, M. & Zuidema, J. (2012). 'The 12321-model of Dutch spelling acquisition'. *Linguistics in the Netherlands 2012,* 111-122.

Van Craenenbroeck, J. & Van Koppen, M. (2019). 'Theoretische taalkunde in het digitale tijdperk'. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 135(4): 416-432.

Zuidema, J. & Neijt, A. (2017). 'The BasisSpellingBank. A spelling database with knowledge stored as a lexicon of triplets'. *Written Language and Literacy* 20(1): 52-79.