(verschenen in: Alex Reuneker, Ronny Boogaart en Saskia Lensink (eds.), *Aries netwerk - een constructicon*. Columns aangeboden aan Arie Verhagen, september 2016).

Van A naar B

Door: *Jelle Zuidema* - Institute for Logic, Language & Computation, Universiteit van Amsterdam

In 2004 verhuisde ik voor de zoveelste keer *van A naar B*: van Edinburgh naar Amsterdam, voor een baan aan de Universiteit van Amsterdam. Ik kwam te wonen in de Czaar Peterstraat in Amsterdam, in een klein appartement in een huizenblok met mooie bakstenen gevels van rond 1900, dat dringend aan renovatie toe was. Aan de overkant waren de oude huizen al afgebroken, en verrees in rap tempo een nieuw complex, in de typische bouwstijl van begin 21e eeuw: een efficiënt betonnen casco, met een bakstenen voorgevel die beter past bij de oorspronkelijke panden in de straat. Van de bouw overdag maakte ik niet veel mee, maar elke avond bij thuiskomst kon ik de vorderingen van die dag goed opnemen. Daarbij viel iets op: de bakstenen gevel vorderde verdacht snel, en met verdacht grote rechthoeken van enkele meters bij enkele meters tegelijk. Nooit zag ik een rafelrandje, een stukje muur waar nog enkele bakstenen ontbraken. Waren de bouwvakkers enorm gedisciplineerd, en metselde ze elke dag door totdat de op de tekentafel bepaalde doelstelling van de dag gehaald was? Of was er iets anders aan de hand?

Na enige tijd realiseerde ik me dat een veel waarschijnlijkere hypothese was dat de bakstenen niet meer één voor één ter plekke door metselaars op elkaar werden gestapeld, maar dat de bouwvakkers kant-en-klare, rechthoekige stukken muur aanvoerden en op het betonnen casco bevestigden. Ook al was ik niet bekend met het bestaan van grote kant-en-klare muurstukken in de bouw (en achtte ik het klassieke metselwerk a-priori daarom veel waarschijnlijk), mijn observaties van de vorderingen in de avonden waren zoveel waarschijnlijker onder de muurstukken-hypothese dat ik die spoedig als de enige redelijke verklaring begon te zien. Daarmee volgde ik impliciet de wet van Bayes: de kans op een hypothese *gegeven* de data is proportioneel aan de kans op de hypothese *vóór* het zien van data maal de waarschijnlijkheid van de data gegeven de hypothese.

Achteraf realiseer ik me dat ik overdag als postdoctoraal onderzoeker aan de UvA met een heel vergelijkbare vraagstuk bezig was: worden zinnen in natuurlijke taal woord voor woord aan elkaar gemetseld, of halen taalgebruikers grote kant-en-klare constructies op uit hun geheugen? Die vraag was gemotiveerd door de populariteit van de constructie-grammatica-school in de taalwetenschap (Verhagen, 2005). Heel wat mensen hadden al hun eigen observaties gepubliceerd met voorbeelden van constructies zoals ‘zich een weg banen’ en ‘jij altijd met je X’, die taalgebruikers verdacht vaak in hun geheel in een zin stopte, of verdacht snel produceerden of begrepen, of die een verdacht afwijkende betekenis of fonologie hadden (het equivalent van een scheef hangend stuk muur). Maar daarmee was ik nog niet tevreden.

Ik wilde de wet van Bayes zo expliciet mogelijk toepassen. Kon je aan de hand van data van tienduizenden taaluitingen laten zien dat die data echt veel minder waarschijnlijk was onder de woord-voor-woord-hypothese dan onder de grote-brokstukken-hypothese? Hoe ga je van een sterke intuïtie naar een formeel model? Het werk aan het zogeheten 'data-oriented parsing' model door Remko Scha en Rens Bod en collega's aan de UvA (Bod, Scha & Sima'an, 2003) leek een fantastisch startpunt, en dat was één van de belangrijkste redenen voor mij om aan de UvA te willen werken (en van *Montpellier Park* naar *de Czaar Peterstraat* te verhuizen). Taal is echter een stuk taaier probleem dan een gemetselde muur, en dat is het vooral omdat de bouwstenen zo enorm divers zijn (er zijn tienduizenden verschillende woorden), zo enorm veel complexer zijn (elk van die bouwstenen heeft semantische, syntactische en fonetische eigenschappen) en zo enorm van elkaar verschillen in frequentie ('de' komt tienduizend keer vaker voor dan 'baksteen'). En als we grote brokstukken in beschouwingen willen nemen, zonder van te voren vast te stellen hoe groot of klein ze mogen zijn of van welke vorm (het equivalent van rechthoekige muurstukken), dan loopt het helemaal uit de hand: het aantal potentiële brokstukken groeit exponentiëel met de lengte van de zinnen.

Ik ontwikkelde een algoritme dat de meest waarschijnlijke afleidingen voor een corpus van zinnen kon berekenen, maar alleen met relatief kleine brokstukken (diepte-4-subbomen) en op relatief kleine corpora. Het werd een mooi algoritme, en leidde tot een paar mooie publicaties (Zuidema, 2007; Borensztajn, Zuidema & Bod, 2009) maar 'zich een weg banen' en 'jij altijd met je X' zijn relatief zeldzaam en komen in de corpora waar ik mee kon werken niet of nauwelijks voor. Ik draaide mijn algoritme op corpora zoals het 'Openbaar Vervoer Informatie Systeem', met zinnen als 'Wanneer gaat de volgende trein van Amsterdam naar Utrecht?'. Tot mijn grote frustratie rolde daar geen van de populaire constructies uit de literatuur uit, alleen keer op keer de suggestie dat taalgebruikers 'van A naar B' als kant-en-klaar brokstuk gebruiken. Zou het?

Referenties

Rens Bod, Remko Scha, and Khalil Sima'an (eds.), Data-Oriented Parsing (2003). University of Chicago Press / CSLI Publications.

Gideon Borensztajn, Willem Zuidema and Rens Bod (2009), Children's grammars grow more abstract with age - Evidence from an automatic procedure for identifying the productive units of language. TopiCS in Cognitive Science 1(1):175-188

Verhagen, Arie. "Constructiegrammatica en 'usage based' taalkunde." *Nederlandse taalkunde* 10 (2005): 197-222.

Willem Zuidema (2007), *Parsimonious Data-Oriented Parsing*, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007), pp551-560.