Biologisch rekentuig brengt taalkundigen op nieuwe gedachten

Eigenlijk spreken we

Otto Vork

Talen gedragen zich in menig opzicht als DNA. Maar terwijl biologen en informatici al jarenlang geavanceerde rekenmethoden gebruiken om stambomen te reconstrueren uit DNA-codes, nemen taalkundigen deze methoden pas de laatste jaren over. Met controversiële resultaten.

Bruno van Wayenburg

allemaal Turks

Talen zijn net levende wezens: ze leven, krijgen nakomelingen en sterven, ze hebben zusjes (Nederlands en Duits), verre achterneven (Nederlands en Russisch) en zelfs complete stambomen, zoals de Indo-Europese taalfamilie die nu over de hele wereld gesproken wordt.

Talen lijken zelfs zo erg op levende wezens, stelt een kleine groep biologen en taalkundigen, dat je hun familiegeschiedenis uit kunt pluizen met de allerlaatste DNA-analysetechnieken, die eigenlijk bedoeld zijn om de evolutie van biologische soorten in kaart te brengen. De resultaten zijn opmerkelijk: de Indo-Europese oertaal zou drieduizend jaar ouder zijn dan altijd gedacht, en niet uit de Oekraïne, maar uit Turkije komen. In zekere zin zijn alle Europeanen van origine Turk, wat de huidige discussie over het Turkse EU-lidmaatschap weer eens in een heel ander licht plaatst.

'Cladistische' methoden kunnen misschien nog wel dieper dan 10.000 jaar terug in het verleden van een taal kijken. Deze grens wordt door historisch taalkundigen meestal aangehouden omdat verwante woorden nog verder in het verleden onherkenbaar gemuteerd zijn. Daarom is de vraag, of alle grote taalfamilies van de wereld zijn terug te voeren op één enkele oertaal, waarschijnlijk nooit te beantwoorden.

De grens van 10.000 jaar is al ruim, gezien het feit dat woorden binnen enkele eeuwen grondig kunnen veranderen of verdwijnen. Dat het Nederlandse 'wiel' en het Engelse 'wheel' verwant zijn, hoort een kind nog, maar andere familieleden zijn het Oud-Indische 'chakra' en het Oud-Griekse 'kuklos' (vanwaar ons 'cyclisch'), die in de gereconstrueerde oertaal Proto-Indo-Europees (PIE) dezelfde stamvader hebben, het woord : *$x^{w}ex^{w}$los (het sterretje geeft aan dat het om een hypothetische, gereconstrueerde vorm gaat). Op één of andere manier heeft de 'x' uit het PIE op bepaalde plaatsen kans gezien om in een Griekse 'k' te veranderen, en in een Nederlandse 'w'. Ook de rest van het woord heeft in diverse familietakken verschillende 'klankmutaties' ondergaan.

Een ander alsmaar voortschrijdend proces is het verdwijnen en vervangen van woorden, nog een reden dat wij Nederlanders een Rus zoveel slechter verstaan dan een Duitser. Het Russisch voor vader is bijvoorbeeld 'otets', dat voorzover bekend niet verwant is aan ons 'vader'. Vader en father zijn terug te voeren op een zelfde Proto-Indo-Europese wortel, ph2ter (waarbij h2 een keelklank is), en otets niet.

Cladistische methoden om stambomen te maken testen duizenden tot een miljoen mogelijke stambomen om te kijken welke het best past bij de bekende gegevens. Eén criterium om ze allemaal te beoordelen is bijvoorbeeld mutatiezuinigheid, ofwel 'maximum parsimony': Stel, de talen A en B hebben twee verwante woorden ('vader' en 'father'), waar talen C en D twee verwante woorden hebben (het Russische 'otets' en het Poolse 'ojciec').

Dan is het waarschijnlijker dat A en B zusjes zijn, en C en D ook, terwijl A en C hoogstens nichten zijn. Een dergelijke stamboom is te verklaren met maar één woordvervanging ergens op de takken.

Een alternatieve stamboom (A en C zijn zusters, beiden nichten van zusters B en D) kan alleen verklaard worden met minstens twee woordvervangingen van vader/father naar otets/ojciec of andersom.

Stamboom In het algemeen geldt: hoe eerder twee talen uit elkaar zijn gegaan, hoe minder verwante woorden ze hebben. Door het vergelijken van talen en achterhalen van zulke veranderingen, de 'comparatieve methode', hebben taalkundigen de afgelopen tweehonderd jaar oertalen als het PIE in groot detail gereconstrueerd. Alleen de oudste regionen van de stamboom zijn in details nog onduidelijk. Zo weet men niet of Grieks en Armeens (of Romaans en Keltisch) apart zijn afgesplitst van de stamboom, of dat ze nog een poosje samen verder zijn gegaan.

En passant wordt er veel duidelijk over de prehistorische geschiedenis van de volken die deze talen spraken. De herkomst van de Roma, ('zigeuners') uit India is alleen ontdekt doordat ze een van oorsprong Indische taal bleken te spreken. Armeense, Iraanse, Griekse en Slavische leenwoorden illustreren hun lange trektocht naar het westen. Een handvol leenwoorden uit het Burushaski, alleen gesproken in een klein gebied in de Noord-Indiase Hindu Kush, pint één pleisterplaats op hun reis wel heel nauwkeurig vast.

"De stambomen en het evolueren van talen deden me erg denken aan mijn eigen werk ", zegt bioloog Russell Gray van Auckland University in Nieuw-Zeeland. Dat werk was het reconstrueren van de evolutionaire geschiedenis van de pinguïn aan de hand van DNA-gegevens.

Net zoals woorden, muteren DNA-codes in de loop van de tijd, iets wat vooral in het niet-coderende DNA ongestraft door kan gaan. Soms worden zelfs vreemde stukken in een genoom opgenomen, bijvoorbeeld als virussen DNA inbouwen in dat van hun gastheer. Dat is vergelijkbaar met een taal die leenwoorden overneemt. Door van twee verwante soorten of van hele stambomen te traceren op welke manier bepaalde stukken DNA verschillen, kunnen biologen stambomen reconstrueren die een goed idee geven van de evolutie van de soorten. Weliswaar is het vaak lastig om de stamboom precies te reconstrueren, maar de hoofdlijnen, en vaak ook belangrijke details, zijn met hoge waarschijnlijkheid vast te stellen.

Hittitische tempel in Anatolie Hittitisch is vermoedelijk de eerste afgesplitste Indo-Europese taal, 1800 tot 1300 v.Chr.

Aangezien de hoeveelheden gegevens daarbij enorm zijn, en de 'woorden' in het DNA zeer lang, wordt daarbij veelvuldig gebruik gemaakt van computers en geavanceerde rekentechnieken. In de afgelopen decennia is deze tak van DNA-rekenen uitgegroeid tot een compleet vakgebied: de bio-informatica.

Maximale zuinigheid Gray en zijn collega Fiona Jordan besloten eens te kijken wat het bioinformatica-arsenaal zou doen met het Austronesisch, de enorme familie van Stille-Zuidzeetalen waartoe ook het Maori behoort.

Van een taalkundige kregen de onderzoekers een database van 5185 Austronesische woorden in 77 talen, van Hawaiiaans en Maori tot het Taiwanese Paiwan en het Javaans, inclusief verwantschapsinformatie. Het ging om basiswoorden als 'vader' of 'water', van het genre dat vermoedelijk niet al te snel vervangen wordt door leenwoorden.

Uit de bioinformatica-receptendoos haalden de Nieuw-Zeelanders vervolgens de veelgebruikte *maximal parsimony* (maximale zuinigheid) rekenmethode. Deze methode probeert uit de miljoenen mogelijke stambomen de boom te vinden, die de DNA-volgorde verklaart met zo weinig mogelijk mutaties. Het idee hierachter is dat de eenvoudigste verklaring vermoedelijk een eind in de goede richting zal zitten.

Vertaald naar de taalkunde door Gray kwam dat neer op een stamboom waarin het aantal malen dat een woord vervangen wordt door een niet-verwant woord minimaal was.

Enigszins tot zijn verbazing kreeg Gray een heel behoorlijke talenboom die in grote lijnen overeenkwam met de in de taalkunde geaccepteerde indeling, maar die wel op een andere, tamelijk objectieve manier was samengesteld.

Daarnaast vormde de boom een duidelijke ondersteuning van de 'expressetrein'-theorie voor de kolonisatie van de eilanden rond en in de Stille Oceaan. Volgens deze theorie, ook ondersteund door archeologisch bewijsmateriaal en genetische gegevens van de Austronesiërs zelf, kwamen de oorspronkelijke Austronesiërs uit Taiwan. Geholpen door de uitvinding van de landbouw verspreidden ze zich vanaf 4000 v. Chr. in sneltreinvaart over de eilanden in de Stille Oceaan, door de Filippijnen, Indonesië, Oceanië, om uiteindelijk uit te komen in Hawaii, Nieuw-Zeeland en Paaseiland.

De onderverdelingen van Grays familiestamboom kwamen mooi overeen met de eilandengroepen waar de huidige talen nog gesproken worden, wat dit beeld bevestigde. "Het was geen controversieel resultaat", zegt Gray.

Controverse kwam er pas in 2003, toen hij een analyse publiceerde van het Indo-Europees, paradepaard van historisch-taalkundigen en de comparatieve methode, en met Engels, Spaans, Russisch en Frans de meest wijdverbreide taalfamilie op aarde.

Verdwijnkans Gray gebruikte een vernieuwde rekenmethode, die meer op een statistisch model van taalevolutie was gebaseerd. In plaats van een zo zuinig mogelijke boom streefde de rekenmethode een boom na die het best klopte met de aanname dat woorden per jaar een bepaalde kans hebben om uit de taal te verdwijnen. Deze verdwijnkans mocht over de loop van de stamboom wel variëren, omdat sommige talen conservatiever zijn dan andere, maar al te wilde fluctuaties kostten een getoetste boom 'strafpunten', zodat hij weer onwaarschijnlijker werd.

De exercitie eindigde met een verzameling van de meest plausibele stambomen van 87 talen. Deze bomen verschilden in de oudste details, maar waren allemaal min of meer acceptabel voor taalhistorici.

Revolutionairder was dat Gray er ook een dateringsmethode aan koppelde, geijkt aan bekende tijdsvensters als het uitsterven van het Latijn tussen 450 en 800 n.Chr. of het oudst bekende Grieks, vóór 1500 v. Chr. . Hij kwam uit op een leeftijd van 7800 tot 9800 jaar voor het Proto-Indo-Europees, in plaats van de geaccepteerde leeftijd van ongeveer 6000 jaar.

Deze gangbare datering hangt samen met de theorie dat het Indo-Europees zijn oorsprong heeft op de steppen ten noorden van de Zwarte Zee. Het Kurgan-ruitervolk, waarvan de archeologische sporen ruimschoots voorhanden zijn, zou vanaf het vierde millennium v.Chr. uitgezwermd zijn over Europa en Zuid-West-Azië tot India, met medeneming van hun taal en de technologie van het wiel en de wagen.

De veel eerdere datering van Gray spoort daar niet mee, en wijst eerder op de concurrerende 'Anatolische hypothese' van de archeoloog Colin Renfrew. Volgens die versie werd het PIE gesproken door een volk in Anatolië, in het huidige Turkije, in de 'vruchtbare halve maan' waar circa 9000 jaar geleden de landbouw uitgevonden werd. Het succes daarvan zou de taalfamilie hebben opgestoten in de vaart der volkeren. In plaats van een krijgshaftig ruitervolk zouden de eerste boeren aan de wieg gestaan hebben van de Indo-Europese expansie. "De reacties waren gemengd, om het vriendelijk uit

"De reacties waren gemengd. Een van de meest publiceerbare was junk science*."*

te drukken", zegt Gray onderkoeld, "een van de meest publiceerbare was *junk science*. Het was een nachtmerrie."

"Volstrekt belachelijk", vindt ook Indo-Europeanist Alexander Lubotsky van de universiteit Leiden het idee om biologiemethoden op taal toe te passen. "Het is duidelijk dat taal volgens heel andere mechanismen verandert dan DNA, dat toch min of meer mechanische slijtage ondergaat." Wat ook niet hielp om de harten en geesten van de taalkundigen te winnen, was dat een half jaar daarvoor het Amerikaanse tijdschrift *Proceedings of the National Academy of Science* een slecht onderbouwd artikel publiceerde over een door biologen bepaalde stamboom van het Keltisch, met talloze fouten en taalkundige slordigheden.

"Disciplines hebben hun eigen stijl, en taalkundigen kunnen nogal... agressief zijn in hun debat", weet Michael Dunn van het Max Planck-instituut in Nijmegen, een Australische taalkundige die met cladistische methoden werkt aan een groep eilandtalen van Papoea-Nieuw-Guinea. Bovendien wordt de taalkunde geplaagd door goedbedoelende maar fanatieke amateurs met onzinnige theorieën, en enig wantrouwen is wel verklaarbaar. "Zelf was ik ook extreem sceptisch."

Dunn beschouwt Gray nu als een van de biologen die taalkundige vragen met de nodige omzichtigheid en respect voor prestaties van de historische taalkunde aanpakken.

Ruitervolk De Scythen waren nauw verwant met de Kurgan-krijgers, die gezien worden als de sprekers van de Indo-Europeese oertaal.

Glottoklok Toch zijn de taalkundige bezwaren wel begrijpelijk. De gedachte om statistische technieken te gebruiken in plaats van de moeizame, maar degelijke comparatieve methode is al in de jaren vijftig gelanceerd door de Amerikaanse taalkundige Morris Swadesh, onder de naam glottochronologie. Dat is een soort taalkundige koolstof-14-methode, waarbij van een korte lijst met basiswoorden van het genre 'vader', de zogenaamde Swadesh-lijst, gekeken wordt hoeveel twee talen nog gemeen hebben. Onder de zogeheten 'glottoklok'-aanname dat talen per duizend jaar 68 procent van hun gezamenlijke woordenschat verliezen, kun je zo de datum berekenen dat de twee talen uit elkaar gingen.

Na aanvankelijk enthousiasme bleek de methode niet echt betrouwbare resultaten te geven. Ten eerste bleek de basisaanname eenvoudigweg niet te kloppen. Sommige talen, zoals het Fins, IJslands of Baskisch, zijn veel conservatiever dan andere, zoals Engels. "Daarnaast gooi je een heleboel informatie weg als je alleen met paarsgewijze overeenkomstpercentages werkt", zegt Gray. In zijn dateringsmethode gebruikte Gray een variabele 'kloksnelheid', die echter over de hele stamboom niet al te veel mocht schommelen.

Lubotsky blijft onverminderd kritisch: "De dateringen zijn gewoon veel te vroeg." Een van de doeltreffendste argumenten voor die stelling is wel het feit dat er een PIE-woord voor 'wiel' bestond, terwijl nergens 9000 jaar oude wielen zijn aangetroffen. De oudst bekende afbeelding van een wiel, uit het antieke Soemerië, is van 3500 v. Chr. Ook 'as', 'disselboom', 'juk' en 'wagen', andere technische termen voor mensen die met door paarden aangedreven wagens werken, zijn duidelijk te onderscheiden in het PIE-vocabulair. "Dat zijn woorden voor een manier van leven die 9000 jaar geleden nog niet bestond", zegt Lubotsky. Daarentegen zijn de aanwijzingen voor specifieke landbouwtermen, namen van granen, 'zaaien' of 'oogsten' in het PIE niet zomaar aan te wijzen.

Gray vindt het argument maar half overtuigend. Het woord voor 'wiel' kan best als leenwoord meegelift zijn met de technologie, denkt hij. Weliswaar zijn recentere leenwoorden vaak te herkennen omdat ze niet alle klankwisselingen hebben ondergaan, maar hoe langer het lenen geleden is, hoe moeilijker het verschil te herkennen is. Een alternatieve verklaring is dat het woord in verschillende talen is afgeleid van een wortel die 'draaien' betekent. Dat verklaart overigens nog niet waarom landbouwwoorden in het PIE zo ondervertegenwoordigd zijn.

De Kurgan- en de Anatolische hypothese hoeven elkaar niet helemaal uit te sluiten, omdat ook in Grays plaatje de meeste takken zich pas 6000 jaar geleden afsplitsen, wat overeen zou kunnen komen met de Kurgan-cultuur.

De Indo-Europeanist Don Ringe van Pennsylvania State University was aanvankelijk zeer kritisch over de date-

Anatolië Volgens bioloog Gray ligen in dit Turkse landschap de wortels van de Indo-Europeese taal. Niet ruiters maar landbouwers zouden de eerste sprekers van deze taalstam zijn.

ring. Ook Ringe experimenteert, met informaticus Tandy Warnow, met manieren om theorieën over taalafstamming met de computer te te toetsen, maar dan zonder de statistische aanpak die grote aantallen verschillende stambomen toetst. Ringe gaat uit van een handvol plausibele stambomen, en test een beperkt aantal mogelijke mutaties hierop, een soort tussenvorm van de comparatieve en cladistische methoden.

Hij noemde statistici die zich met taalkunde bemoeien ooit 'de bezitters van een klein aantal hamers waarmee ze de wereld afzoeken naar schroeven waar ze op kunnen hameren.'

Inmiddels is Ringe iets positiever. In een recente vergelijking van verschillende cladistische technieken, waaronder die van hemzelf en Gray, beveelt hij ze aan om 'op een maximaal rigoureuze manier de consequenties van hun oordelen uit te werken.' Veelbetekenender is misschien nog wel het recente werk van Ringe aan de evolutionaire geschiedenis van een lintwormsoort: als biologen zich met talen mogen bemoeien, waarom dan niet andersom?

Ook bij andere wetenschappers lijkt de biologen-aanpak gewoner te worden. Een vergelijkbare exercitie, door Clare Holden die de Afrikaanse Bantoe-taalfamilie ordende, ondersteunde ook een geleidelijke verspreiding door het continent, door archeologen weer toegeschreven aan de uitvinding van de landbouw. De bekende Nijmeegse taalkundige Pieter Muysken werkt aan plannen om de Amazone-talen met cladistische methoden te onderzoeken.

Ook antropologen en handschrift-deskundigen besteden inmiddels aandacht aan cladistische methoden uit de biologie, om stambomen van tradities of (overgeschreven) manuscripten te reconstrueren.

Vragenformulier Vernieuwend is de opzet van taalkundige Michael Dunn van het Max Planck-instituut in Nijmegen, en collega's. Zij gebruikten cladistische technieken juist om een verwantschap aan te tonen tussen vijftien talen van eilanden ten oosten van Papoea-Nieuw-Guinea, met fraaie namen als Rotokas en Lavukaleve. Deze 'Papoea-talen' horen niet bij de Austronesische taalfamilie waardoor ze omringd worden, maar een onderlinge verwantschap kon ook niet aangetoond worden. 'Er zijn niet genoeg geloofwaardige woordovereenkomsten om te laten zien dat het een familie is, al zijn er wel opvallende structurele overeenkomsten', zegt Dunn.

Hij en collega's besloten daarom grammaticale eigenschappen te gebruiken, en stelden een soort vragenformulier samen voor alle vijftien talen waarin 108 eigenschappen worden aangevinkt, zoals 'de taal heeft een bepaald lidwoord', 'de taal heeft een aparte dualis', of 'de taal kent een verschil tussen "inclusief wij (ik en jij/jullie)" en "exclusief wij (ik en hij/zij)'.

Een voordeel van deze methode is dat er geen aannamen nodig zijn over het al of niet verwant zijn van woorden. Een taalkundige moet altijd besluiten of twee woordvormen op elkaar lijken door afstamming of ontlening (met inachtneming van de juiste klankveranderingen), dan wel door toeval. "Grammaticale eigenschappen kun je, hoewel het heel veel werk is, tamelijk objectief coderen", zegt Dunn, die gebruik maakte van publicaties over de Papoeatalen en eigen onderzoek van zijn vakgroep.

Op de resulterende eigenschappenmatrix lieten de taalkundigen een 'zuinige' stamboomgenerator los en zie: de hoofdtakken van de resulterende stamboom klopten keurig met de geografische verspreiding van de talen, zo meldden zij in *Science* van 23 september.

Eén tak wordt op Bougainville gesproken, een op de Lousiade-archipel, één op de Solomon-eilanden en één op de Bismarck-archipel. 'De structuur die we vonden was niet willekeurig maar geografisch plausibel', zegt Dunn, dus is een taalkundige verwantschap ook aannemelijk.

Het idee, dat ook klopt met archeologische vondsten, is dat de Papoeatalen afstammen van een oertaal die in de regio gesproken werd vóór de invasie door sprekers van het Austronesisch, die vanaf 4000 jaar geleden plaatsvond.

Aangezien er vrijwel geen verwante woorden zijn, is de aangetoonde verwantschap vermoedelijk zeer oud, misschien zelfs voorbij de magische grens van 10.000 jaar waarin woorden onherkenbaar eroderen. Als grammaticale structuren verwantschap kunnen laten zien voorbij deze grens, concludeert Dunn in zijn artikel hoopvol, 'wordt ook de mogelijkheid geopend om relaties te vinden tussen de 300 andere taalfamilies en geïsoleerde talen van de wereld.' Want, zoals de meeste taalkundigen vermoeden maar nog lang niet kunnen bewijzen, eigenlijk was er ooit, ver voorbij de 10.000 jaar-grens, één oertaal. Eigenlijk spreken we allemaal Afrikaans. ■

David Searls, 'Trees of life and language', *Nature*, 27 november 2003

Russell Gray, 'Pushing the Time Barrier in the Quest for Language Roots', *Science*, 23 september 2005