kennislink.nl maakt nieuwsgierig

Bereken de woordvolgorde

Wiskunde in de taalwetenschap

Een boodschap kun je op verschillende manieren overbrengen. Je kunt bijvoorbeeld de woordvolgorde van een zin veranderen zónder dat de betekenis echt verandert. Maar hoe kies je eigenlijk (onbewust) de volgorde die uiteindelijk uit je mond of op papier komt? Taalwetenschappers gebruiken wiskundige functies om dit te onderzoeken.

Wanneer je naar het onderstaande plaatje van Sneeuwwitje kijkt, kun je op twee manieren beschrijven wat je ziet:

A. *De boze koningin geeft Sneeuwwitje de giftige appel.*
B. *De boze koningin geeft de giftige appel aan Sneeuwwitje.*

Welke van de twee zinnen zou jij gebruiken?

Dit zijn dus twee zinnen die precies hetzelfde zeggen, en die allebei grammaticaal goed zijn. In de eerste zin noem je eerst de ontvanger (Sneeuwwitje) en daarna het thema (de giftige appel), in de tweede zin andersom. Onbewust maak je een keuze tussen deze twee volgordes. Waarop is deze keuze eigenlijk gebaseerd?

Stukjes van de puzzel

Er zijn al veel onderzoekers geweest die naar de keuze voor woordvolgorde hebben gekeken. Ze vonden elk een deel van de oplossing, een stukje van de puzzel. De één ontdekte bijvoorbeeld dat mensen vaak eerst de dingen noemen die al bekend zijn bij henzelf en de mensen met wie ze communiceren. Daarna noemen ze pas de nieuwe informatie. Zo kost het zo min mogelijk denkkracht om te communiceren. Je zult dus eerder ‘Ik geef jou een boek’ zeggen dan ‘Ik geef een boek aan jou’, omdat de jij-persoon al bekend is, maar het boek nog niet.

Weer een andere onderzoeker toonde aan dat ook de ‘bezieldheid’ of *animacy* een rol speelt: je noemt meestal eerst

de mensen en de dieren, en daarna pas de dingen. En zo blijken er nog veel meer eigenschappen van de woordgroepen een rol te spelen bij de keuze die we maken.

Klare taal

Omdat het kiezen tussen twee woordvolgordes meestal helemaal onbewust gaat, is het lastig om zelf te bedenken waarom je soms voor de ene zin kiest, en soms voor de andere. Taalwetenschappers gebruiken daarom vaak andere manieren om deze keuze te onderzoeken. Een veelvoorkomende manier is het kijken naar taal die al 'klaar' is: taal die al gezegd of opgeschreven is. Een verzameling van bestaande taal noemen we een corpus (meervoud: corpora).

Een corpus dat taalwetenschappers vaak gebruiken is het 'International Corpus of English' of korter ICE. In dit corpus is gesproken en geschreven Engelse taal verzameld uit allerlei bronnen: van sportcommentaar bij een voetbalwedstrijd tot een literaire roman, en van een persoonlijke brief tot een politieke debat. Voor het Nederlands bestaat er het Corpus Gesproken Nederlands. Hierin vind je gesprekken en toespraken van mannen en vrouwen van allerlei leeftijden uit verschillende delen van Nederland en Vlaanderen. Er bestaat ook een corpus met daarin filmpjes van dove Nederlanders die met elkaar praten in gebarentaal: het Corpus Nederlandse Gebarentaal.

Wiskundige functie

Maar hoe kunnen we die corpora nu gebruiken om de woordvolgorde-puzzel op te lossen? Heel simpel: door te tellen! Taalwetenschappers hebben bijvoorbeeld alle zinnen van het type zin (A) en (B) opgezocht in het Brits-Engelse deel van het ICE-corpus. Van de 930 gevonden zinnen hebben ze vervolgens allerlei eigenschappen geteld. Hoe vaak staat het thema vóór de ontvanger? En hoe vaak is de ontvanger animate? En zo nog ongeveer tien eigenschappen meer.

Er zijn nu dus behoorlijk wat eigenschappen bekend over de zinnen. We willen weten welke eigenschappen ervoor zorgen dat we kiezen voor de woordvolgorde in zin (A) in plaats van (B), of juist andersom. Maar het is voor mensen erg moeilijk om de samenhang te zien als er zoveel eigenschappen bekend zijn. We zien door de bomen het bos niet meer. Daarom gebruiken taalwetenschappers vaak wiskundige functies die hierbij helpen. In plaats van zélf op zoek te gaan naar de rol van de eigenschappen, laten we de computer het werk doen. Want een computer heeft veel meer reken- en denkkracht dan een mens. We noemen het gebruik van computers voor het oplossen van zulke puzzels ook wel patroonherkenning of machinaal leren.

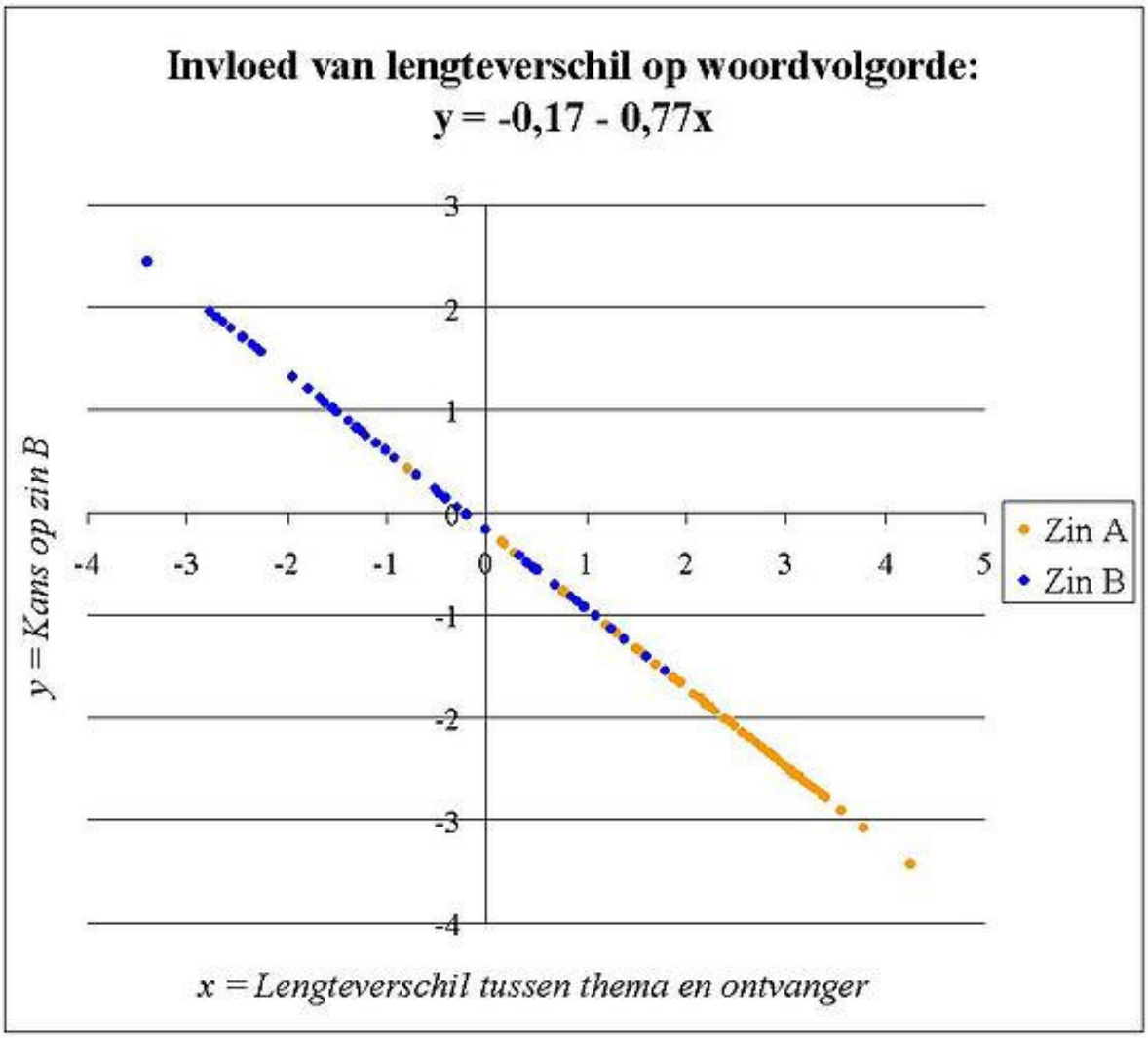
Woordvolgorde voorspellen

Laten we nu een functie opstellen voor het kiezen tussen de woordvolgordes in zin (A) en (B) in het voorbeeld. We nemen hiervoor alle eigenschappen van de zinnen op als termen in de functie. De uitkomst van de functie voorspelt de kans dat we de woordvolgorde in 'De boze koningin geeft de giftige appel aan Sneeuwwitje' (zin B) kiezen, en niet die in 'De boze koningin geeft Sneeuwwitje de giftige appel' (zin A). Als de uitkomst bijvoorbeeld 4 is, dan is de kans op de woordvolgorde in zin B vier keer zo groot als de kans op de woordvolgorde in zin A. Als de uitkomst negatief is, is de kans juist groter op de volgorde in zin A.

In de grafiek hieronder zie je een functie met maar één eigenschap: het lengteverschil (in woorden) tussen het thema en de ontvanger. Als dit lengteverschil positief is, is het thema langer dan de ontvanger, en als het negatief is andersom. In voorbeeldzin (A) bestaat het thema uit drie woorden ('de giftige appel'), en de ontvanger uit één ('Sneeuwwitje'). Het lengteverschil is dan dus (3-1=) 2. Het lengteverschil is -2 in de zin 'De boze koningin geeft appels aan de mooie Sneeuwwitje', omdat het thema nu één woord bevat ('appels') en de ontvanger drie ('de mooie Sneeuwwitje'). In de zin 'De boze koningin geeft appels aan Sneeuwwitje', met één woord in het thema en één in de ontvanger, is het lengteverschil 0.

De kleuren van de punten in de grafiek zijn de échte woordvolgordes zoals ze in het corpus voorkomen, de waarde op de y-as geeft de voorspelde kans dat het de volgorde uit zin (B) is. In de grafiek zie je dat als het lengteverschil tussen het thema en de ontvanger groter wordt, de kans op zin (B) kleiner wordt, en wel met een factor -0,77. Dat

betekent dat mensen de langere woordgroepen meestal achteraan in de zin zetten. Ook zie je dat de punten niet door de oorsprong gaan, maar door (0, -0,17). Als het thema en de ontvanger even lang zijn, is de kans op de woordvolgorde in zin A dus iets groter dan de kans op die in zin B.



De kans op zin (B) is afhankelijk van het lengteverschil (in woorden) tussen het thema en de ontvanger. Dit verschil is positief als het thema uit meer woorden bestaat dan de ontvanger, en negatief als de ontvanger juist uit meer woorden bestaat. Je ziet dat zin (B) vooral voorkomt als de ontvanger uit meer woorden bestaat dan het thema.

We kunnen de functie uitbreiden met de andere eigenschappen. Elk daarvan krijgt een eigen as in een multi-dimensionale ruimte, net als het lengteverschil in de grafiek. De computer tekent de 930 zinnen uit het corpus in de ruimte en gaat vervolgens op zoek naar een functie die de punten het beste beschrijft. Alleen eigenschappen die écht (significant) van invloed zijn op de keuze worden in de functie opgenomen. De functie die de computer vindt kan goed voorspellen welke keuze mensen maken: voor 88 procent van de 930 zinnen kiest de functie dezelfde woordvolgorde als degene die de zin heeft gezegd of geschreven.

Sneeuwwitje en de giftige appel

In het begin van dit artikel heb je gekozen welke van de twee zinnen over Sneeuwwitje jij zou gebruiken. De gevonden functie voor het Brits-Engels kunnen we gebruiken om te kijken of de functie dezelfde woordvolgorde voorspelt. Dus voor iedere eigenschap van de zin vullen we in de functie een waarde in. Als de eigenschap er is, is die waarde 1, en als deze er niet is, is de waarde 0. Voor het lengteverschil vullen we het verschil in aantal woorden in.

Eigenschap	**Factor x waarde**
De kruising met de x-as	1,88
Het lengteverschil tussen thema en ontvanger	-0,60 × 2

De ontvanger is in de 3e persoon (‘zij’)	1,08 × 1
De ontvanger in onbepaald (‘een…’)	1,00 × 0
De ontvanger is nieuwe informatie	0,65 × 0
De ontvanger is een voornaamwoord	-1,35 × 0
Het thema is onbepaald (‘een…’)	-1,12 × 0
Het thema is nieuwe informatie	-1,19 × 0
Het thema is iets abstracts	-1,37 × 0

Je ziet dat de factor voor het lengteverschil tussen het thema en de ontvanger van -0,77 veranderd is naar -0,60. Blijkbaar beïnvloeden de andere eigenschappen de rol van het lengteverschil. Je ziet dus dat het nodig is om de computer te laten rekenen, want als mens kunnen we onmogelijk zulke moeilijke verbanden zien.

Om de uitkomst van de functie te vinden tellen we de vermenigvuldigingen tussen de factoren en de waarden op: 1,88 – 1,20 + 1,08 = 1,76. Deze uitkomst betekent dat de kans op zin (B) 1,76 keer zo groot is als zin (A). Mensen zouden dus geneigd zijn te kiezen voor de zin ‘De boze koningin geeft de giftige appel aan Sneeuwwitje’.

Heb jij dezelfde zin gekozen?

Lees ook:

- Hoe wordt een corpus samengesteld? (Kennislink)
- Wat bezielt talen eigenlijk? (Kennislink)

Auteur

Daphne Theijssen

Gepubliceerd door

Kennislink (correspondentennetwerk)

Publicatiedatum

dinsdag, 29 juni 2010

Kernwoorden

taal & spraak stevin

Meer Taal & Spraak

Meer Taal & Spraak
Dialectverschillen in het Zweeds nemen af
door Mathilde Jansen
-
2 uur geleden
Therese Leinonen deed onderzoek naar dialectverschillen in het Zweedse taalgebied. Er is veel variatie in de uitspraak van …
Nieuw model simuleert spraakherkenning