Laura van Eerten & Griet Depoorter
Instituut voor Nederlandse Lexicologie

Het Corpus Gesproken Nederlands

Een waardevolle bron van hedendaagse spraak uit Nederland en Vlaanderen

Wetenschappers die onderzoek doen naar een taal, hebben bronmateriaal nodig. Bijvoorbeeld een verzameling van teksten of spraakopnames, in de taalkunde ook wel *corpus* genoemd. Het samenstellen van een corpus is een kostbare en tijdrovende klus die soms jaren in beslag kan nemen. Gelukkig worden er ook corpora speciaal gemaakt om voor verschillende soorten onderzoek dienst te kunnen doen. Het Corpus Gesproken Nederlands (CGN) is zo'n corpus. Nooit eerder werd er zo een grote hoeveelheid Nederlandse spraak in een corpus verzameld.

Onvolledige zinnen

Voordat het CGN gebouwd werd, waren er voornamelijk Nederlandstalige tekstcorpora beschikbaar. Hierdoor richtte het taalkundig onderzoek zich sterk op de beschrijving van het geschreven Nederlands. In de wetenschap ontstond er behoefte aan een groot spraakcorpus. Zoals iedereen die zich bezighoudt met taal, weet, is gesproken taal veel complexer dan geschreven taal. Mensen spreken namelijk niet in grammaticaal correcte zinnen: we haperen, maken zinnen niet af, beginnen opnieuw en spreken dialect. Een grote verzameling gesproken Nederlands bracht daarom voor taalonderzoekers veel nieuwe mogelijkheden met zich mee.

Nederlands in Europa

Ook de technologie had baat bij de aanleg van een groot Nederlandstalig spraakcorpus. In het veeltalige Europa concurreert het Nederlands met andere talen, met name het Engels. Het Engels had een belangrijke voorsprong in de taal- en spraaktechnologie. Dat kwam door de beschikbaarheid van grote databanken van gesproken en geschreven Engels. Voor het Nederlands ontbraken dergelijke bronnen nog. Een groot Nederlands spraakcorpus zou de positie van het Nederlands in de taal- en spraaktechnologie en daarmee de economische en culturele positie van het Nederlands in Europa versterken.

Nederland en Vlaanderen

Om het Nederlandse spraakcorpus daadwerkelijk te realiseren, sloegen Nederland en Vlaanderen de handen ineen. De Nederlandse en Belgische regering financierden het project en de bouw van het corpus vond plaats in beide landen. Medewerkers van tien Nederlandse en Vlaamse universiteiten en taalinstellingen werkten vanaf 1998 vijf jaar lang aan het opnemen, verzamelen en bewerken van bijna dertienduizend spraakfragmenten. Dat resulteerde in 2004 in het CGN: een gesproken verzameling van Standaardnederlands, waarvan een derde afkomstig uit Vlaanderen en twee derde uit Nederland.

Hoeveel woorden?

Het CGN bevat maar liefst 900 uur spraak. Uitgeschreven levert dat bijna 9 miljoen woorden op. Zo'n grote hoeveelheid data is zeer waardevol omdat het onderzoekers in staat stelt om statistisch onderzoek te doen. Een verschijnsel dat in een spontaan gesprek voorkomt, beurtwisselingen bijvoorbeeld, is beter te analyseren met behulp van honderd opnames in plaats van drie. En een individuele onderzoeker kan in zijn of haar eentje nooit zo veel data bij elkaar brengen voor een onderzoek. Het CGN is een kant-en-klare onderzoeksbron, waardoor wetenschappers alle tijd en geld kunnen investeren in het onderzoek zelf, in plaats van in het verzamelen van data.

Karakteristieke gesprekken

De data in het CGN zijn zeer divers. Dat maakt het corpus geschikt voor uiteenlopende onderzoeksrichtingen. Er zijn opnames van verschillende soorten spreeksituaties verzameld, waaronder spontane gesprekken, voorbereide publieke lezingen, officiële debatten en voorgelezen verhalen. De spreekstijlen in die situaties hebben een verschillend karakter. De woordkeuze in een spontaan telefoongesprek tussen twee vrienden is bijvoorbeeld heel anders dan in een officieel debat in de Tweede Kamer. En in tegenstelling tot spontane spraak bevat een voorgelezen tekst nauwelijks afgebroken zinnen en tussenwerpsels als *uh* en *oh*. Die voorgelezen, grammaticaal correcte teksten zijn goed bruikbaar voor het testen van spraakherkenners. De spontane gesprekken zijn juist weer interessant voor experimenteel taalkundig onderzoek.

Monnikenwerk

De spraakopnames zijn zorgvuldig en grotendeels handmatig bewerkt. Alle opnames zijn helemaal woord voor woord met de hand uitgeschreven. Het doel was de spraak zo waarheidsgetrouw mogelijk neer te schrijven. Dat betekent inclusief herhalingen, versprekingen en afgebroken woorden. Speciale woorden, zoals dialectwoorden en niet-Nederlandse woorden, zijn apart gemarkeerd. Zelfs zogenoemde sprekergeluiden, zoals gelach en gehoest, hebben een eigen codering. De onderstaande tekst is een voorbeeld van zo'n uitgeschreven geluidsfragment. Het is een stukje voetbalcommentaar van een lokale radiozender.

	Component	Aantal woorden	VL	NL
a.	Spontane conversaties	2.626.172	878.383	1.747.789
b.	Interviews met leraren Nederlands	565.433	315.554	249.879
c.	Telefoondialogen opgenomen m.b.v. platform	1.232.636	489.100	743.537
d.	Telefoondialogen opgenomen m.b.v. minidiskrecorder	853.371	343.167	510.204
e.	Zakelijke onderhandelingen	136.461	0	136.461
f.	Interviews en discussies vanop radio en tv	790.269	250.708	539.561
g.	Discussie, debatten, vergaderingen (m.n. politieke)	360.328	138.819	221.509
h.	Lessen	405.409	105.436	299.973
i.	Spontane commentaren vanop radio en tv	208.399	78.022	130.377
j.	Actualiteitenrubrieken en reportages vanop radio en tv	186.072	95.206	90.866
k.	Nieuwsbulletins vanop radio en tv	368.153	82.855	285.298
l.	Beschouwingen en commentaren vanop radio en tv	145.553	65.386	80.167
m.	Missen, lezingen, plechtige toespraken	18.075	12.510	5.565
n.	Colleges, voordrachten, lezingen	140.901	79.067	61.834
o.	Voorgelezen teksten	903.043	351.419	551.624
	Totaal	**8.940.098**	**3.285.631**	**5.654.644**

De verschillende soorten spreeksituaties die opgenomen zijn in het CGN

ja 't was in de eenentwintigste minuut. || toen uh brak op de rechterkant Bas Schaaij door || hij omspeelde z'n man mooi legde de bal terug. || in de zestien meter kwam uh Rikken || binnengelopen die werd aangetikt || tenminste zo oordeelde scheidsrechter uh Tempelaar. || hij gaf daarvoor een strafschop een gele paart*u || gele kaart voor Felibor Peters || en die werd uh de strafschop werd verzilverd || door uh Luc Van Raaij. || en nog geen twee minuten later || was de bal uh in één keer werd ie diep gegeven en || werd er wederom gescoord. || nu was het Mario Lammers met een uh knap afstandsschot. || het is dus nul twee voor Hatert.

Een uitgeschreven spraakfragment

Uitspraak en grammatica

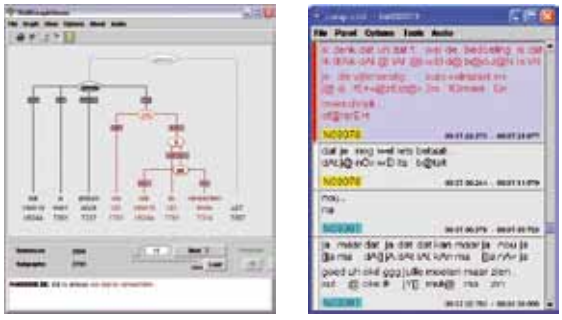
Bij een groot deel van de fragmenten is ook extra taalkundige informatie toegevoegd. Die extra informatielagen worden annotaties genoemd. De grammaticale ofwel syntactische annotatielaag bevat bijvoorbeeld informatie over de functie van elk woord en zinsdeel in een zin. De fonetische annotatielaag geeft de uitspraak in klanken weer. De klankweergaven maken het mogelijk om op een bepaalde uitspraak van een woord te zoeken, om uitspraakvarianten van woorden te kunnen analyseren. De a in *tram* en *flat* bijvoorbeeld wordt door Vlamingen en Nederlanders anders uitgesproken. Naast grammaticale en klankinformatie zijn er nog annotatielagen toegevoegd met woordsoortinformatie, lemma's en intonatiepatronen.

Ja, ik uh...

Het CGN is ook uitgebreid met statistische informatie, zoals woordfrequentie. Alle woorden in het corpus zijn automatisch geteld en daarvan zijn zogeheten frequentielijsten gemaakt. Er is een frequentielijst van het complete corpus, maar er zijn ook specifiekere lijsten waarin onderscheid gemaakt wordt tussen de Vlaamse en Nederlandse data, en de verschillende spreeksituaties in het corpus. In de lijsten kun je bijvoorbeeld goed zien dat in spontane gesprekken tussenwerpsels als *ja* en *uh* het vaakst gebruikt worden, terwijl in voorgelezen teksten vooral functiewoorden als *de, het* en *een* hoogfrequent zijn. Woordfrequentie is belangrijk voor vergelijkend onderzoek, maar wordt ook gebruikt als hulpmiddel bij het maken van woordspelletjes of taallessen.

Spontane gesprekken	Voorgelezen teksten
ja	de
dat	en
en	een
ik	het
uh	van
die	in
maar	ik
een	dat
de	ze
´t	hij

Top tien meest voorkomende woorden in spontane gesprekken en voorgelezen teksten



Voorbeelden van de fonetische (links) en syntactische (rechts) annotaties

Poldernederlands

Voor onderzoek naar spraak is het vaak ook belangrijk dat je wat weet over de achtergrond van de personen die aan het woord zijn. In het CGN is hier ook rekening mee gehouden. Van zo veel mogelijk sprekers is informatie verzameld over leeftijd, geboorteplaats, regio waar iemand is opgegroeid, opleidingsniveau, beroep, enzovoort. Dat soort aanvullende gegevens maakt het mogelijk om een onderzoek te richten op een specifieke groep sprekers, bijvoorbeeld sprekers van het Poldernederlands. Dat is een variatie op het Nederlands waarin de tweeklanken ei, ui en ou met een wijdere mond gearticuleerd worden en daardoor klinken als aai, ou en aau. Die uitspraak wordt voornamelijk toegedicht aan hoogopgeleide vrouwen van middelbare leeftijd. Met de informatie in het CGN kunnen de personen die aan die voorwaarden voldoen, eenvoudig gevonden worden.

Het bos door de bomen

Spraakopnames, taalkundige annotaties, woordfrequentie, sprekergegevens ... Alles bij elkaar zit er in het CGN een enorme hoeveelheid data. Het is niet eenvoudig om er je weg in te vinden. Daarom is er binnen het project een speciaal zoekprogramma ontwikkeld: Corex. Met behulp van Corex kun je alle annotaties bekijken, doorzoeken en gelijktijdig de opnames beluisteren. Het is ook mogelijk om slechts een deel van de data te doorzoeken, een zogeheten deelverzameling of subcorpus. Zo'n subcorpus kun je zelf samenstellen op basis van spreker- en/of fragmentkenmerken, bijvoorbeeld van alle telefoongesprekken tussen hoogopgeleide vrouwen van middelbare leeftijd, de sprekers van het Poldernederlands. Corex is een vast onderdeel van het CGN en wordt altijd samen met de data aangeboden.

Winkel voor corpora

Het CGN is voor wetenschappers en bedrijven verkrijgbaar bij de TST-Centrale: het centrale punt voor opslag, onderhoud en distributie van digitale Nederlandstalige taalmaterialen. De Nederlandse Taalunie richtte de TST-Centrale in 2004 op om ervoor te zorgen dat kostbare dataverzamelingen als het CGN goed bewaard worden en beschikbaar blijven voor onderzoek. Projectsubsidies voor het aanleggen van een corpus dekken namelijk over het algemeen alleen de kosten voor de ontwikkeling, maar niet voor het beheer en de distributie van de uiteindelijke resultaten. Als een project afgerond is, verdwijnen de kostbare data meestal in een la. Zonde. De oprichting van de TST-Centrale heeft ervoor gezorgd dat kostbare dataverzamelingen herbruikbaar zijn. De TST-Centrale bewaart en distribueert het CGN niet alleen, maar geeft ook advies over het gebruik en biedt introductiecolleges en workshops aan. Zo gaat ook de kennis over het CGN niet verloren.

Het goede voorbeeld

Dat het CGN een belangrijk corpus is, blijkt wel uit het feit dat het corpus als voorbeeld dient voor nieuwe corpusbouwprojecten. Het formaat van de annotaties, aangeduid als CGN-formaat, wordt bijvoorbeeld vaak hergebruikt en de samenwerking tussen Nederland en Vlaanderen wordt bewaakt. De ontwikkeling van het JASMIN-spraakcorpus[1] is zo'n project in lijn van het CGN. Dat spraakcorpus is een verzameling van hedendaagse Nederlandse spraak van kinderen, niet-moedertaalsprekers en ouderen. Een belangrijk deel van het corpus bestaat uit opnames van mens-machine-interactie. Een gesprek met een spraakcomputer lokt bij mensen vaak overdreven uitspraak uit of mensen gaan harder praten. Die sprekergroepen en taalgebruiksituaties ontbreken in het CGN en vormen daarmee een waardevolle aanvulling.

Op een bepaald moment zal de spraak in het CGN niet meer hedendaags, Standaardnederlands zijn. Levende talen veranderen en verouderen, zo ook het Nederlands. Maar dat betekent niet dat het corpus zijn waarde verliest. Het CGN zal nog steeds gebruikt kunnen worden om de Nederlandse taal uit de periode 1991-2003 te bestuderen, al dan niet in vergelijking met 'modernere' spraakcorpora. Daarnaast is het een goed voorbeeld voor nieuwe corpusbouwprojecten. Kortom: het CGN is al bijna tien jaar een begrip en zal dat de komende tien jaar zeker nog blijven.

Literatuur

- Cucchiarini, C., Van Hamme, H., Van Herwijnen, O. & Smits, F. (2006). JASMIN-CGN: Extension of the Spoken Dutch Corpus with Speech of Elderly People, Children and Non-natives in the Human-Machine Interaction Modality. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation* (pp. 135-138). Genua.
- Eerten, L van (2007). Over het Corpus Gesproken Nederlands. *Nederlandse Taalkunde*, 12 (3), pp. 194-215.
- Oostdijk, N. (2000). The spoken Dutch Corpus: Overview and first Evaluation. In: *Proceedings LREC* 2000. Genua.

De auteurs zijn als taalkundigen verbonden aan de TST-Centrale/het INL, Matthias de Vrieshof 2-3, 2311 BZ Leiden. E-mail: laura.vaneerten@inl.nl, griet.depoorter@inl.nl.
Uitgebreide informatie en documentatie over het Corpus Gesproken Nederlands is te vinden op de website van de TST-Centrale: www.inl.nl/tst-centrale. Vragen of opmerkingen over het CGN kunt u sturen naar servicedesk@inl.nl.

[1] JASMIN is een acroniem voor Jongeren, Anderstaligen, Senioren en Machine Interactie. Het corpus is gemaakt in het kader van STEVIN: een meerjarig subsidieprogramma dat de ontwikkeling van bouwstenen voor Nederlandstalige taal- en spraaktechnologie (zoals grote spraakcorpora) stimuleert. STEVIN is een acroniem voor Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands.