Digitaal
Het Corpus gesproken Nederlands

In juni 1998 is het project Corpus Gesproken Nederlands (CGN) van start gegaan. Dit vijfjarig project is gericht op de aanleg van een databank van het hedendaags Standaardnederlands zoals dat wordt gesproken door volwassenen in Nederland en Vlaanderen. De beoogde omvang van het corpus is circa tien miljoen woorden, waarvan tweederde deel afkomstig is uit Neder- land, en eenderde uit Vlaanderen. Het CGN bestaat uit een verzameling van een groot aan- tal fragmenten van (opnames van) gesproken tekst. In totaal gaat het hierbij om een duizend- tal uren spraak. Al het materiaal wordt orthografisch getranscribeerd. Daarbij worden om de twee à drie seconden ankerpunten aangebracht die in een later stadium worden gebruikt om de transcriptie te koppelen aan het spraaksignaal. De orthografische transcriptie vormt het uitgangspunt voor de lemmatisering en de verrijking van het materiaal met woordsoortinformatie. Verder is er voor een selectie van één miljoen woorden voorzien dat er een fonetische transcriptie wordt vervaardigd, er een geverifieerde koppeling van het spraaksignaal met de transcriptie op woordniveau beschikbaar komt en dat het materiaal door middel van een syntactische analyse wordt verrijkt. Tenslotte wordt een bescheiden deel van het corpus, circa 250.000 woorden, van een prosodische annotatie voorzien, waarbij de belangrijkste grenzen van woordgroepen (frasegrenzen) alsmede de één of twee belangrijkste woorden (zinsaccen- ten) van elke frase worden aangeduid.

Het CGN is van groot belang voor de verdere ontwikkeling van de Nederlandstalige taal- en spraaktechnologie. Daarbij valt te denken aan toepassingen op het gebied van de mens-machine-communicatie zoals bv. automatische spraakherkenners en dialoogsystemen. Daarnaast biedt het Corpus interessant materiaal voor allerhand taalkundig en cultuur-historisch onderzoek. Met zijn authentieke opnames geeft het CGN immers een uniek beeld van het Nederlands zoals Nederlanders en Vlamingen dat spreken aan het begin van de 21e eeuw in het leven van alledag.

Het CGN-project wordt gefinancierd door de Vlaamse en Nederlandse Regering en door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). Het totale budget bedraagt circa 4,6 miljoen euro. Alle rechten zijn in handen van de Nederlandse Taalunie. Van het materiaal mag derhalve niets verveelvoudigd en/of openbaar gemaakt worden op welke wijze dan ook zonder voorafgaande schriftelijke toestemming van de Nederlandse Taalunie. Het project wordt gecoördineerd vanuit twee locaties: Gent voor Vlaanderen en Nijmegen voor Nederland.

Corpusontwerp en -opbouw

Het uitgangspunt bij het ontwerp van het CGN is geweest dat het corpus bij voorkeur zo wordt samengesteld en een zodanige omvang heeft dat het optimaal bruikbaar is voor het onderzoek in de diverse onderzoekdisciplines en toepassingsgebieden. Het probleem dat zich daarbij voordoet is dat het gesproken Nederlands wordt gekenmerkt door een grote mate aan diversiteit, terwijl ook de interessen van de verschillende gebruikersgroepen en de daaruit voortvloeiende vereisten ten aanzien van het corpus op een aantal punten nogal uiteenlopend blijken te zijn. Daarnaast zijn er een aantal beper-

kende factoren die van invloed zijn op het ontwerp. Onder die factoren zijn de volgende: (1) de beschikbare tijd en middelen; (2) de technische mogelijkheden waarover men kan beschikken voor het maken van de opnames en de verdere bewerking van de data; en (3) de juridische regelgeving waaraan men gehouden is met betrekking tot het verzamelen en openbaar maken van data. Er is derhalve gekozen voor een ontwerp zodanig dat het resulterende corpus beschouwd kan worden als een noodzakelijkerwijs beperkte doch plausibele steekproef van het hedendaags Standaardnederlands, waarbij tevens zoveel mogelijk tegemoet gekomen wordt aan de wensen en behoeften van de verschillende groepen potentiële gebruikers. Ook is rekening gehouden met de databestanden die voor het Nederlands reeds beschikbaar zijn, dit om duplicatie te voorkomen en de beschikbare middelen optimaal in te zetten.

Het CGN omvat opnames van alledaagse gesprekken die Nederlanders c.q. Vlamingen met elkaar voeren over allerhande onderwerpen en in uiteenlopende situaties. Ook worden er in het corpus (fragmenten van) radio- en televisieprogramma's, interviews, lezingen en telefoongesprekken opgenomen. De samenstelling van het corpus met daarbij de omvang van de componenten in aantallen woorden wordt schematisch weergegeven in Figuur 1. Voor alle opnames is informatie beschikbaar over de sprekers. Het gaat daarbij o.a. om gegevens over geslacht, leeftijd, regio, opleiding en beroep.

Figuur 1. Samenstelling Corpus Gesproken Nederlands

<table>
<tr><td rowspan="8">dialoog / multiloog
8.110.000</td><td rowspan="5">privé
6.635.000</td><td rowspan="5"></td><td rowspan="5">spontaan
6.635.000</td><td rowspan="3">direct 'face- to-face'
3.460.000</td><td>conversaties
3.000.000</td></tr>
<tr><td>interviews
460.000</td></tr>
<tr><td rowspan="3">'distanced
3.175.000</td><td>'telefoondialogen
3.000.000</td></tr>
<tr><td>zakelijke onderhandelingen
175.000</td></tr>
<tr><td></td></tr>
<tr><td rowspan="2">publiek
1.475.000</td><td>uitgezonden
750.000</td><td colspan="2">min of meer voorbereid
750.000</td><td>interviews en discussies
750.000</td></tr>
<tr><td>niet uitgezonden
725.000</td><td colspan="2">spontaan
725.000</td><td>discussies, debatten, vergaderingen
375.000</td></tr>
<tr><td colspan="4"></td><td>lessen
350.000</td></tr>
<tr><td rowspan="7">monoloog
1.890.000</td><td colspan="2">privé
40.000</td><td colspan="2">min of meer voorbereid
40.000</td><td>beschrijving van route of plaatjes
40.000</td></tr>
<tr><td rowspan="6">publiek
1.850.000</td><td rowspan="4">uitgezonden
950.000</td><td colspan="2">spontaan
250.000</td><td>spontaan commentaar
250.000</td></tr>
<tr><td colspan="2">min of meer voorbereid
700.000</td><td>actualiteitenrubrieken, reportages
250.000</td></tr>
<tr><td colspan="2"></td><td>nieuwsbulletins
250.000</td></tr>
<tr><td colspan="2"></td><td>beschouwingen, commentaren
200.000</td></tr>
<tr><td>niet uitgezonden
900.000</td><td colspan="2">min of meer voorbereid
900.000</td><td>lezingen, toespraken
275.000</td></tr>
<tr><td colspan="4"></td><td>voorgelezen tekst
625.000 (+ 375.000</td></tr>
</table>

Opname en digitalisering

Voor een deel worden opnames in eigen beheer gemaakt; daarnaast worden opnames verkregen uit samenwerkingsverbanden met andere projecten, bedrijven, organisaties en instellingen. We noemen hier o.a. het VNC project 'De uitspraak van het Standaardnederlands', de Blindenbibliotheken in Vlaanderen en Nederland, de VRT, de lokale omroepen en het Transferbureau van de KUN. Materiaal wordt zoveel mogelijk aan de basis digitaal opgenomen. Wanneer gebruik gemaakt wordt van bestaand materiaal zijn digitale opnames echter niet altijd beschikbaar. Alle opnames worden - voor zover ze niet al in electronische vorm zijn binnengekomen - via een geluidskaart in een pc ingelezen. Met uitzondering van telefoonopnames, wordt het materiaal opgeslagen in een ongecomprimeerd 16 bits, 16 kHz wav-formaat. Informatie over de opnameomstandigheden, de gebruikte apparatuur e.d. is beschikbaar als onderdeel van de meta-data.

Orthografische transcriptie

Al het opgenomen materiaal wordt orthografisch getranscribeerd. Het orthografisch transcript is een woordelijk neerslag van wat er gezegd werd. Het transcript is in overeenstemming met de regels die daarvoor zijn vastgelegd in een protocol (Goedertier & Goddijn 2000). Daarbij worden herhalingen, versprekingen, aarzelingen en dergelijke uitgeschreven; achtergrondgeluiden daarentegen worden niet in het transcript weergegeven. Op alle transcripten wordt een spellingcontrole uitgevoerd. Voor het transcriberen wordt gebruik gemaakt van het programma PRAAT.[1] Dit stelt de transcribenten in staat om het spraaksignaal te beluisteren en te bekijken, en tegelijkertijd een transcriptie in te voeren. Voor iedere spreker is een aparte regel beschikbaar. Tijdens het transcriptieproces worden tevens tijdsmarkeringen aangebracht waarbij korte stukjes van twee à drie seconden in het signaal worden aangeduid. Deze markeringen worden in een later stadium gebruikt voor de automatische koppeling van het transcript en het spraaksignaal.

Lemmatisering en verrijking met woordsoortinformatie

Het volledige corpus wordt verrijkt met woordsoortinformatie. Binnen het project is daarvoor een tagset gedefinieerd (Van Eynde 2000) die ca. 300 tags omvat en die aansluit bij de praktijk van de *Algemene Nederlandse Spraakkunst* (ANS; Haeseryn et al. 1997). De tagset is conform de EAGLES richtlijnen die daarvoor opgesteld zijn in het kader van de internationale standaardisering (Gibbon et al. 1998).[2] Voor het aanbrengen van de tags wordt gebruik gemaakt van daartoe ontwikkelde software die aan elk woord de meest waarschijnlijke tag toekent. Het resultaat van deze automatische bewerking wordt gecontroleerd en waar nodig handmatig gecorrigeerd. Voor het lemmatiseren wordt eveneens gebruik gemaakt van een computerprogramma en ook hiervan wordt de uitvoer handmatig gecorrigeerd.

Lexicologische koppeling

Binnen het project wordt een CGN-lexicon ontwikkeld. Het lexicon is van belang voor de verschillende vormen van transcriptie en annotatie, maar vervult daarnaast een belangrijke rol in de ontsluiting van de data. Door middel van een lexicologische koppeling wordt het mogelijk een verder doorgedreven lemmatisering te realiseren waarbij onder meer scheidbare werkwoorden en preposities gerelateerd worden aan de juiste lemmata.

Fonetische transcriptie

Voor ongeveer één miljoen woorden zal een

(geverifieerde) fonetische transcriptie worden vervaardigd. Daarbij wordt gebruik gemaakt van de SAMPA-symbolenset (Gibbon et al. 1998).[3] Om te bepalen welke procedure het meest geschikt is voor het maken van een fonetische transcriptie zijn er enkele experimenten uitgevoerd. Daarbij is onder meer gekeken naar het effect dat een voorgegeven, automatisch gegenereerde transcriptie heeft op de snelheid van werken en accuratesse. Inmiddels is een begin gemaakt met de transcriptiewerkzaamheden, waarbij de transcribent een automatisch gegenereerde transcriptie krijgt voorgegeven en die waar nodig corrigeert.

Signaalkoppeling

Voor het materiaal waarvoor een geverifieerde fonetische transcriptie beschikbaar is zal het spraaksignaal op woordniveau worden gekoppeld aan het orthografisch transcript en zal het resultaat van deze oplijning handmatig worden geverifieerd. Voor het overige materiaal is voorzien dat het signaal en het orthografisch transcript weliswaar gekoppeld zullen worden (automatisch), maar zal er geen verificatie plaatsvinden.

Prosodische annotatie

Ongeveer 250.000 woorden zullen worden voorzien van een prosodische annotatie. Welke invulling hieraan gegeven wordt staat op dit moment (juni 2000) nog niet vast. Het ligt in de bedoeling in ieder geval de belangrijkste grenzen van woordgroepen (frasegrenzen) alsmede de één of twee belangrijkste woorden (zinsaccenten) van elke frase aan te duiden. Een besluit over de exacte invulling van deze taak zal worden genomen op basis van het advies dat daarover wordt uitgebracht door een commissie waarin verschillende experts zitting hebben.

Syntactische annotatie

Ten behoeve van de syntactische annotatie die is voorzien voor één miljoen woorden wordt een annotatieschema ontwikkeld (Moortgat & Schuurman in voorbereiding). Zodra op basis van ervaringen met het annoteren van een reeks van proeffragmenten het annotatieschema kan worden vastgesteld, zal daadwerkelijk met het uitvoeren van de annotatietaak worden begonnen. Daarbij zal gebruik gemaakt worden van het in Saarbrücken ontwikkelde programma Annotate.[4]

Verspreiding van de resultaten

De resultaten van het project komen beschikbaar voor wetenschappelijk onderzoek en voor de ontwikkeling van commerciële producten. In deze producten mogen de bijdragen van individuele personen niet op een herkenbare manier aanwezig zijn.
Delen van het corpus worden al tijdens de looptijd van het project (ongeveer om de zes maanden) beschikbaar gesteld. Sedert 1 maart jl. is de eerste tranche (met een omvang van 615.000 woorden) beschikbaar. De release van de tweede tranche is voorzien voor oktober 2000. Het volledige corpus zal naar verwachting medio 2003 beschikbaar zijn. De distributie ervan - inclusief de geluidsopnames - zal waarschijnlijk worden verzorgd door de European Language Resources Association (ELRA)[5].

Nadere informatie

Meer informatie over het project is te vinden op http://lands.let.kun.nl/cgn/. Ook verspreidt het CGN-project met enige regelmaat een nieuwsbrief. Heeft u hierin interesse of heeft u andere vragen, dan kunt u zich wenden tot het CGN-bureau:

Bureau Corpus Gesproken Nederlands
NWO, Geesteswetenschappen
Mw. drs. A. Dijkstra
Postbus 93120
2509 AC Den Haag
e-mail: dijkstra@nwo.nl

4 Voor meer informatie zie http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html.
5 Zie ook http://www.icp.grenet.fr/ELRA/home.html.

Nelleke Oostdijk
Afdeling Taal en Spraak, Katholieke Universiteit
Nijmegen
N.Oostdijk@let.kun.nl

Bibliografie

Goedertier, W. & S. Goddijn (2000). *Protocol voor orthografische transcriptie*. Interne publi- catie CGN-project.
Zie http://lands.let.kun.nl/cgn/.

Gibbon, D., R. Moore & R. Winski, (red.) (1998). *Handbook of standards and resources for spoken language systems*. Berlijn, New York: Mouton de Gruyter.

Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij & M.C. van den Toorn (1997). *Algemene Nederlandse Spraakkunst*. Gronin- gen: Martinus Nijhoff.

Moortgat, M. & I. Schuurman (in voorbe- reiding). *Syntactische annotatie*.

Van Eynde, F. (2000). *Part-of-speech tagging en lemmatisering*. Interne publicatie CGN-pro- ject. Zie http://lands.let.kun.nl/cgn/.

1 Voor meer informatie zie http://www.fon.hum. uva.nl/praat/.

2 EAGLES staat voor Expert Advisory Group for Lan- guage Engineering Standards. Zie ook http://www.ilc. pi.cnr.it/EAGLES96/home.html.

3 SAMPA is een ASCII codering van de fonemen van bepaalde talen, waaronder het Nederlands, en is gebaseerd op het International Phonetic Alphabet (IPA). Zie ook http://coral.lili.unibielefeld.de/ Documents/ sampa.html.