Over het Corpus Gesproken Nederlands

Laura van Eerten[1]

1 Inleiding

Sinds jaar en dag worden grote verzamelingen data gebruikt voor wetenschappelijk onderzoek. Deze verzamelingen, veelal omvangrijke tekstbestanden, worden corpora genoemd. De ontwikkeling van het Corpus Gesproken Nederlands (CGN) was daarom niet direct een vernieuwend project, maar wel uniek: nog niet eerder was er zo een grote hoeveelheid gesproken Nederlandse taaldata bij elkaar gebracht. Enkele Nederlandse en Vlaamse universiteiten hebben vijf jaar lang gewerkt aan het verzamelen, catalogiseren en annoteren van de ongeveer negenhonderd uur spraak die het corpus bevat. Het CGN-project, dat gefinancierd werd door NWO en de overheid van Nederland en Vlaanderen, is in 2004 succesvol afgerond.

In dit artikel wordt een overzicht gegeven van het Corpus Gesproken Nederlands vanaf de start van het project tot nu. Eerst wordt in paragraaf 2 de achtergrond – motivatie en projectorganisatie – van het project uiteengezet. Paragraaf 3 behandelt de opbouw en de inhoud van het corpus. In paragraaf 4 wordt een overzicht gegeven van een aantal onderzoeken waarin gebruikgemaakt is van het CGN. In paragraaf 5 wordt ten slotte praktische informatie verstrekt over onderhoud en distributie bij de TST-centrale, gevolgd door toekomstige plannen zoals een online webversie van het CGN en het aanvullende project JASMIN-CGN.

2 Achtergrond

2.1 Motivatie

De voornaamste reden voor de aanvang van het project in 1998, was het versterken van de positie van het Nederlands in de taal- en spraaktechnologie. Voor het Engels waren er al veel taalbronnen beschikbaar zoals bijvoorbeeld het *British National Corpus* en het *Wall Street Journal text corpus.* De beschikbaarheid van deze bronnen heeft de Engelse taal- en spraaktechnologie in een leidende positie geplaatst. Voor de ontwikkeling van de Nederlandse taal- en spraaktechnologie ontbrak dergelijk materiaal. Het project Corpus Gesproken Nederlands werd opgezet om dit hiaat op te vullen.

Naast de taaltechnologische motivatie was er ook vanuit het linguïstische onderzoeksveld vraag naar een Nederlands spraakcorpus. Tot voor de start van het project waren er voor taalkundig onderzoek alleen tekstcorpora beschikbaar. Zoals iedereen die zich met taal bezighoudt weet, is gesproken taal in bepaalde opzichten veel complexer dan geschreven taal: onvolledige zinnen, versprekingen, aarzelingen, interjecties en wederzijdse beïnvloeding van klanken zijn veelvoorkomende fenomenen. Er was nog geen corpus beschikbaar waarin dit soort eigenschappen van gesproken Nederlands onderzocht kon worden.

Een derde argument voor de ontwikkeling van het corpus had betrekking op het educatieve belang. Toepassing van gesproken taalmateriaal in het onderwijs biedt nieuwe perspectieven. In het onderwijs van het Nederlands wordt vaak alleen gebruik gemaakt van geschreven Nederlands, terwijl in het onderwijs van bijvoorbeeld Frans of Duits

[1] Een uitgebreide versie van dit artikel is verschenen in *Nederlandse Taalkunde* 12 (3) (Themanummer: Corpus Gesproken Nederlands), 194 – 215. Met dank aan Griet Depoorter voor het kritisch lezen en becommentariëren van de concepttekst, het uitvoeren van zoekacties in Corex en het maken van schermafbeeldingen. Tevens dank aan de redactie van Nederlandse Taalkunde voor suggesties en commentaar bij een eerdere versie van dit artikel.

luistervaardigheid een prominent onderdeel vormt van het programma. De ontwikkeling van het CGN zou voor spraaktoepassingen binnen het Nederlandse taalonderwijs een eerste stap in de goede richting zijn.

2.2 Projectorganisatie

Het project ging in 1998 van start onder leiding van een bestuur bestaande uit zes leden, met een evenredige vertegenwoordiging uit Vlaanderen en Nederland. Het bestuur werd benoemd door de Vlaamse en Nederlandse financiers en bestond uit vertegenwoordigers van de twee overheden, Nederlandse en Vlaamse onderzoeksinstituten en de Landelijke Onderzoeksschool Taalwetenschap (LOT). Daarnaast nam een afgevaardigde van de Nederlandse Taalunie als waarnemer deel aan de bestuursvergaderingen. De voorzitter van het bestuur was Professor W. Levelt van het Max Planck Instituut te Nijmegen. Het bestuur stelde een stuurgroep in die verantwoordelijk was voor de daadwerkelijke uitvoering van het project.

De dagelijkse coördinatie was in handen van twee projectleiders: W. Goedertier (Rijksuniversiteit Gent) voor Vlaanderen en N. Oostdijk (Universiteit van Nijmegen) voor Nederland. Door de projectleiders werden drie werkgroepen aangestuurd: corpusopbouw, signaalanalyse en corpusannotatie. De werkgroep corpusopbouw was verantwoordelijk voor het ontwerp en de samenstelling van het corpus, de werving van sprekers en de acquisitie van opnames. De werkgroep signaalanalyse ontwikkelde de protocollen en de procedures voor orthografische transcriptie, woordsegmentatie, fonetische transcriptie en prosodische annotatie. De verantwoording voor de daadwerkelijke uitvoering van de annotaties lag bij de werkgroep corpusannotatie (zie ook de website van het CGN http://www.tst.inl.nl/cgndocs/doc_Dutch/start.htm).

3 De inhoud van het corpus

3.1 Spraakfragmenten en annotaties

Het Corpus Gesproken Nederlands bevat bijna dertienduizend spraakfragmenten van Nederlandse en Vlaamse sprekers in verschillende taalgebruiksituaties. Naast voorgelezen teksten en fragmenten van radio en televisie, zoals nieuwsuitzendingen, (sport-) commentaren actualiteitenrubrieken en reportages, zijn er ook spontane conversaties, telefoondialogen, interviews, debatten, lezingen en Nederlandse lessen opgenomen. Deze categorieën zijn in het CGN nauwkeurig opgedeeld in vijftien verschillende componenten:

(1)
- a spontane conversaties (*face-to-face*)
- b interviews met leraren Nederlands
- c telefoondialogen opgenomen met behulp van een telefooncentrale
- d telefoondialogen opgenomen met behulp van een minidiskrecorder
- e zakelijke onderhandelingen
- f interviews en discussies uitgezonden op radio en televisie
- g discussies, debatten, vergaderingen (met name politieke)
- h lessen (middelbare school met focus op leerkracht)
- i spontane commentaren (onder andere sport) uitgezonden op radio en televisie
- j actualiteitenrubrieken en reportages uitgezonden op radio en televisie
- k nieuwsbulletins uitgezonden op radio en televisie
- l beschouwingen en commentaren uitgezonden op radio en televisie
- m missen, lezingen, plechtige toespraken
- n colleges, voordrachten, lezingen
- o voorgelezen teksten

Bij de samenstelling van het corpus is zo veel mogelijk rekening gehouden met wensen en behoeftes van diverse groepen potentiële gebruikers, in plaats van één specifieke doelgroep te

bedienen zoals dat bij eerder samengestelde corpora vaak het geval was. Er bleek vooral vraag te zijn naar spontane spraakdata waarin interactie als belangrijk bestanddeel gezien werd. In het corpus zijn spontane dialogen en multilogen daarom in ruime mate vertegenwoordigd. In eerste instantie was het de bedoeling om een corpus samen te stellen dat uitsluitend spontane spraak bevatte. Door de grote behoefte die er onder spraaktechnologen bestaat aan voorgelezen teksten voor de training van spraakherkenners, is ervoor gekozen om deze vorm van gesproken Nederlands ook op te nemen in de data (Oostdijk 2000).

De uiteindelijke structuur van het corpus is het resultaat van een 'getrapte sampling'. Tabel 1 laat zien hoe de componenten tot stand gekomen zijn door enerzijds *dialoog/multiloog* en anderzijds *monoloog* uit te splitsen naar steeds specifiekere componenten.

Sampling					**Component**
dialoog / multiloog	privé		spontaan	direct	**a.** spontane conversaties
					b. interviews leraren Nederlands
				indirect	**c.** telefoondialogen (centrale)
					d. telefoondialogen (minidisk)
					e. zakelijke onderhandelingen
	publiek	uitgezonden	min of meer voorbereid		**f.** interviews en discussies
		niet uitgezonden	spontaan		**g.** discussies, debatten, vergaderingen
					h. lessen
monoloog	privé		min of meer voorbereid		*
	publiek	uitgezonden	spontaan		**i.** spontaan commentaar
			min of meer voorbereid		**j.** actualiteitenrubrieken, reportages
					k. nieuwsbulletins
					l. beschouwingen, commentaren
		niet uitgezonden	min of meer voorbereid		**m.** lezingen, toespraken
					n. colleges, voordrachten
					o. voorgelezen tekst

Tabel 1: Het definitieve corpusontwerp.

In het oorspronkelijke ontwerp was er nog een component voorbereid op de plaats van de asterisk: beschrijvingen van route of plaatjes. De realisatie van het corpus liep echter vertraging op waardoor de omvang en samenstelling van sommige componenten bijgesteld moest worden. Deze component kwam daarmee te vervallen. De tweedeling in de component telefoondialogen is ook later ingevoerd door onvoorziene technische problemen met de telefooncentrale.

De omvang van de fragmenten loopt per component tamelijk uiteen. Omdat het moeilijk bleek de optimale lengte voor een fragment te bepalen, hebben vooral intuïtieve factoren een rol gespeeld. De 'natuurlijke' lengte van een bepaald soort spraak speelde mee: een nieuwsbericht op de radio is bijvoorbeeld van kortere duur dan een gesproken tekst bij een documentaire (Oostdijk 2000).

Aan de spraakfragmenten is een aanzienlijke hoeveelheid taalkundige informatie toegevoegd in diverse annotatielagen. Echter niet elk fragment is even 'rijk' geannoteerd. Op basis van de verschillende annotaties werd onderscheid gemaakt tussen het basiscorpus en het kerncorpus.

De term *basiscorpus* verwijst naar het volledige corpus en omvat alle fragmenten van het CGN voorzien van orthografische en automatisch gegenereerde fonetische transcripties, woordsoortinformatie, lemma-informatie en de automatische verbinding van transcripties met het geluidssignaal (signaalkoppeling). Onder het *kerncorpus* verstaat men een uitvoeriger verrijkte selectie van het basismateriaal, waaronder de handmatig geverifieerde fonetische transcripties en signaalkoppeling, de syntactische annotaties en de prosodische informatie. In figuur 1 is de indeling van het complete CGN schematisch weergegeven.

Figuur 1: Schematische weergave van het CGN.

Het basiscorpus bevat alle orthografisch getranscribeerde spraakfragmenten. Alle spraak is per spreker uitgeschreven inclusief herhalingen, versprekingen en afgebroken woorden. Voor onder andere dialectwoorden, niet-Nederlandse woorden en sprekergeluiden zoals lachen of hoesten, is een afzonderlijke markering toegepast. De transcripties zijn vervaardigd met behulp van het programma Praat (Boersma & Weenink 1992-2007), en kunnen hierin – onafhankelijk van de exploitatiesoftware Corex (zie paragraaf 4) – samen met het geluidsbestand geopend worden.

Op basis van de orthografische transcripties is het volledige corpus automatisch verrijkt met lemmata en woordsoortinformatie (POS-tags). Binnen het project werd daarvoor een eigen CGN-tagset gedefinieerd die driehonderdzestien verschillende tags omvat, en aansluit bij de praktijk van de ANS (Haeseryn en anderen 1997). De automatisch toegekende tags zijn naderhand nagekeken en waar nodig gecorrigeerd.

De fonetische transcripties zijn eveneens automatisch gegenereerd voor het gehele corpus. Een deel ervan, ongeveer tien procent, is handmatig geverifieerd. De gebruikte foneemset vertoont veel overeenkomsten met gebruikelijke fonetische alfabetten zoals IPA en SAMPA. De fonetische transcripties zijn evenals de orthografie te bekijken in het programma Praat.

Syntactische annotaties zijn vervaardigd voor ongeveer tien procent van het opgenomen materiaal. Alle afhankelijkheidsrelaties binnen een geannoteerde zin zijn hiërarchisch in kaart gebracht in boomstructuren. De boomstructuren zijn te visualiseren met behulp van het programma *Tiger* (zie paragraaf 4) dat geïntegreerd is in het exploitatieprogramma Corex.

Voor ongeveer tweeënhalf procent van het corpus zijn prosodische verschijnselen gemarkeerd. De markering betreft prominente lettergrepen, prosodische grenzen en abnormale klankverlengingen en is aangebracht in de orthografische transcriptielaag.

De annotaties zijn op woordniveau gekoppeld aan het geluidsmateriaal. Deze automatische segmentatie is nadien voor tien procent van de data gecontroleerd en handmatig gecorrigeerd. Het doel van de signaalkoppeling is om woorden in verbonden spraak van elkaar te scheiden, door grenzen of markeringen te plaatsen in het spraaksignaal. Hierdoor is het mogelijk om het spraaksignaal direct terug te vinden bij een bepaalde annotatie of andersom.

3.2 Lexica, frequentielijsten en metadata

Zoals uit figuur 1 blijkt behoren naast de spraakdata met annotaties ook een lexicon, frequentielijsten en bestanden met metadata tot het CGN. Het CGN-lexicon is een woordenlijst die vrijwel alle unieke woordvormen uit het corpus bevat. Het ontwerp is gebaseerd op bestaande elektronische bronnen zoals CELEX (Baayen en anderen 1993), het RBN (Referentiebestand Nederlands 1998) en het WNT (Woordenlijst Nederlandse Taal 1995). De *entries* in het lexicon zijn voorzien van uitgebreide lexicale informatie zoals woordsoort, lemma, syntactisch complementatiepatroon, uitspraak en morfologische segmentatie.

Het lexicon is onderverdeeld in een enkelwoordslexicon en een meerwoordslexicon. Het enkelwoordslexicon (ook wel standaardlexicon genoemd) bevat uitsluitend aaneengeschreven woordvormen. Het meerwoordslexicon bestaat uit meerledige expressies zoals scheidbaar samengestelde werkwoorden (*nemen op, halen adem*), ingeburgerde vreemdtalige uitdrukkingen (*et cetera, wishful thinking*), eigennamen en titels (*Berg En Dal, De Pfaffs*). Door middel van een lexicologische koppeling zijn in het CGN de los geschreven delen van meerledige uitdrukkingen met elkaar verbonden. Via de meerwoordslemma's zijn verwijzingen naar het lexicon opgenomen. Hierdoor zijn zoekacties op de afzonderlijke delen van een expressie mogelijk.

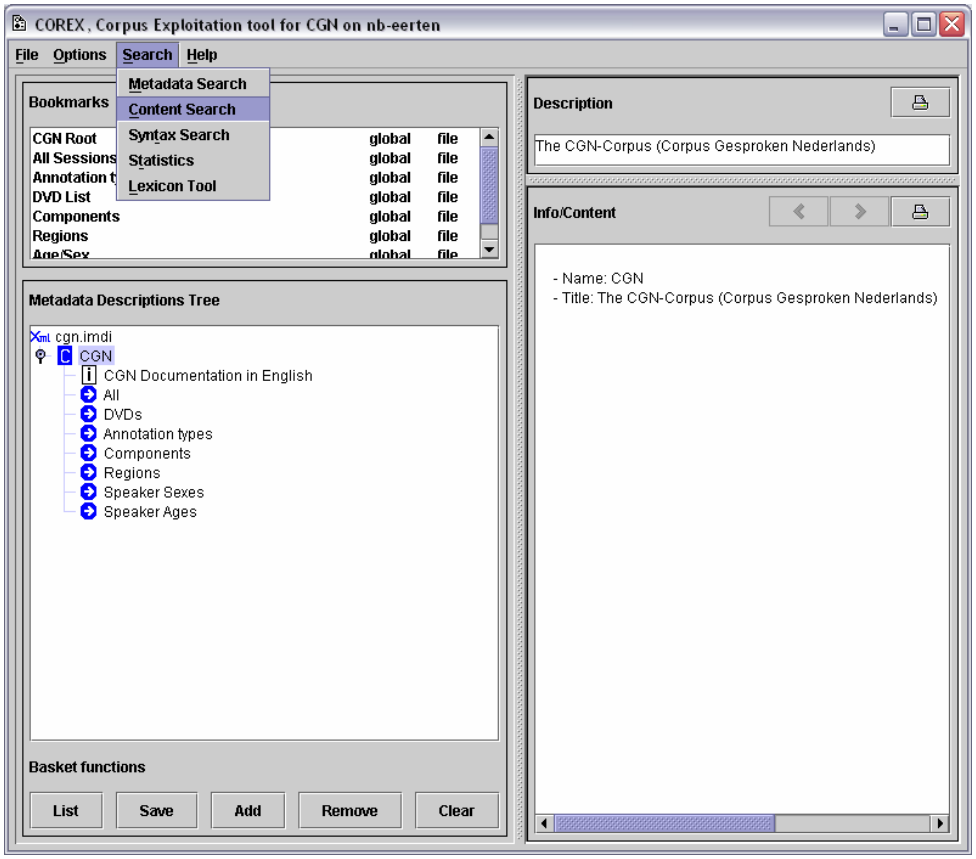
In verschillende lijsten zijn frequentiegegevens uit het CGN opgenomen. Naast een complete lijst van alle woordvormen die in het CGN voorkomen ('totrank'), zijn er aparte frequentielijsten waarbij onderscheid gemaakt wordt tussen Vlaamse en Nederlandse data. Daarnaast zijn frequentielijsten beschikbaar waarbij een uitsplitsing is gemaakt naar de vijftien componenten die in het corpus worden onderscheiden. Al deze frequentielijsten zijn op alfabet of frequentievolgorde gesorteerd. Van de woordsoortinformatie (POS-tags), de lemmata en woordvormen met bijbehorende fonetische transcripties zijn tevens afzonderlijke frequentielijsten beschikbaar.

Figuur 2: Top tien van de frequentielijst 'totrank'.

Het onderdeel metadata geeft nadere informatie over de sprekers en de fragmenten die in het corpus voorkomen: van zo veel mogelijk sprekers is sociolinguïstisch relevante informatie verzameld zoals leeftijd, geboorteplaats of geboorteregio en opleidingsniveau. Met betrekking tot de opnames wordt onder andere informatie verschaft over de locatie, het gemiddelde spreektempo en de datum van opname. Dit soort aanvullende gegevens maakt het mogelijk om diverse variabelen te selecteren voor een specifiek onderzoek. Denk hierbij bijvoorbeeld aan spontane conversaties tussen Nederlandse vrouwen ouder dan 25 jaar, waarin een hoog spreektempo gehanteerd wordt. Meer voorbeelden van onderzoeksvragen en zoekacties worden aan de hand van Corex behandeld in de volgende paragraaf.

4 Het zoekprogramma Corex[2]

Om op een eenvoudige en efficiënte manier door de grote hoeveelheid data te kunnen navigeren, is parallel aan het CGN-project een speciaal zoekprogramma ontwikkeld: de *cor*pus*ex*ploitatiesoftware Corex. In figuur 3 wordt het openingsvenster van Corex afgebeeld van waaruit alle zoekacties starten.

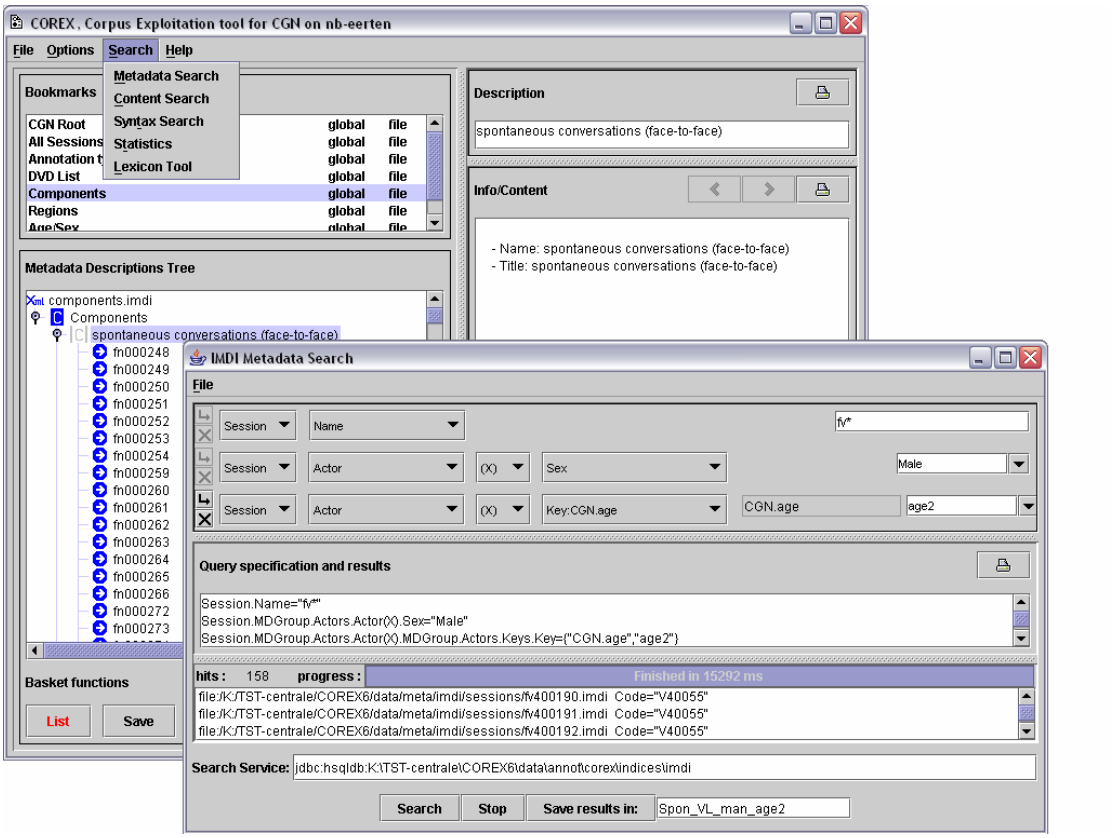


Figuur 3: Het hoofdvenster van Corex met uitgeklapt zoekmenu.

Het programma biedt veel mogelijkheden doordat er gewerkt kan worden met subcorpora, gebaseerd op voorgedefinieerde of eigen criteria, zoals het geslacht en de leeftijd van de spreker en diverse andere metadata. In bijvoorbeeld een onderzoek naar spontane conversaties tussen uitsluitend Vlaams sprekende mannen boven de vijfentwintig, kan er met behulp van Corex een selectie (subcorpus) gemaakt worden van deze groep alvorens de meer specifieke zoekacties uit te voeren. Figuur 4 illustreert hoe de samenstelling van een dergelijk subcorpus in zijn werk gaat.

---

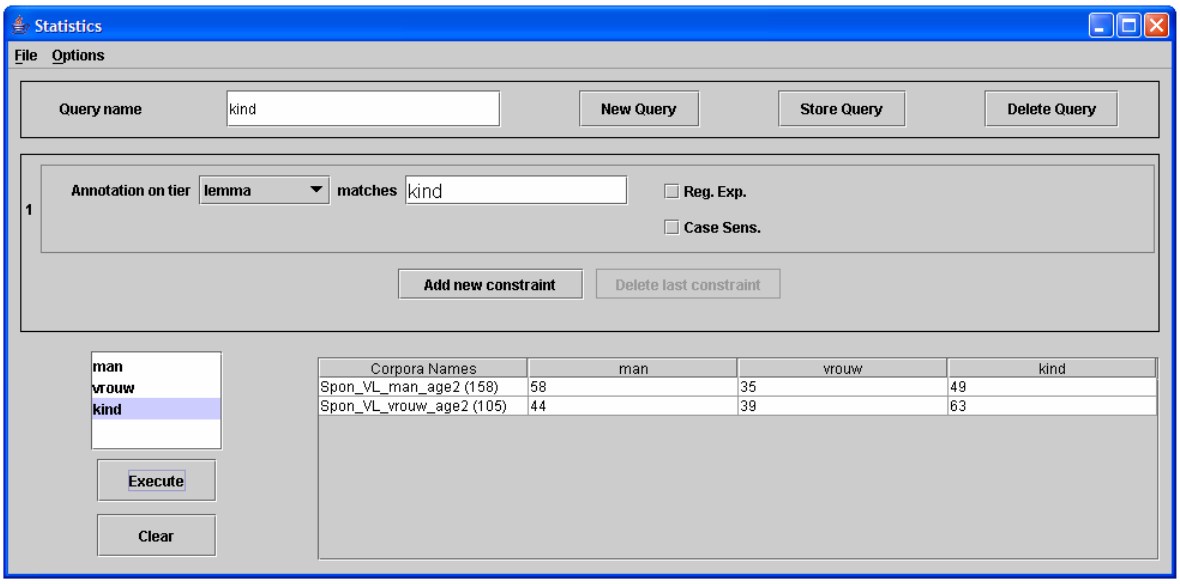
[2] In dit artikel is gebruikgemaakt van versie 6.1 van Corex. Het is mogelijk dat de schermen en zoekacties zoals geïllustreerd in dit artikel er anders uitzien in een oudere of nieuwere versie van het programma.



Figuur 4: Samenstelling van een subcorpus van spontane conversaties tussen Vlaamse mannen in de leeftijdscategorie 25 - 34 jaar.

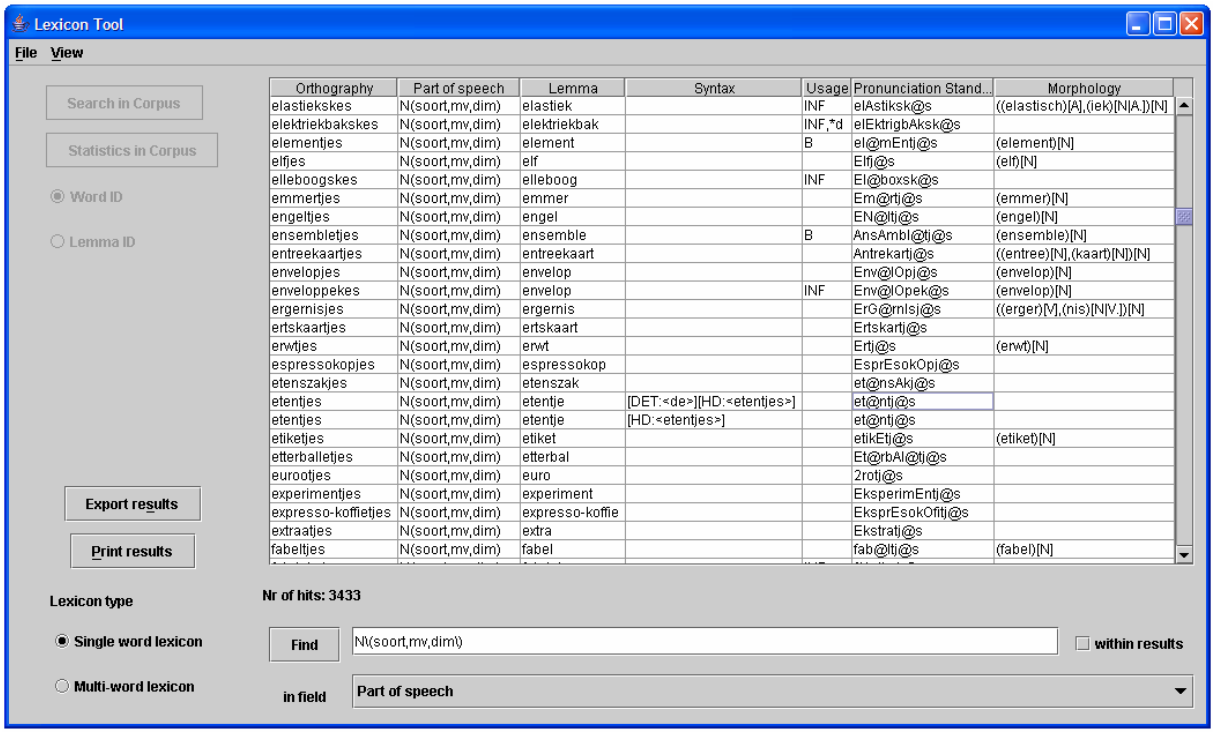
In het hoofdvenster is een eerste selectie gemaakt op basis van de component met spontane conversaties. Binnen de *Metadata Search* wordt het subcorpus verder gespecificeerd door alleen Vlaamse fragmenten ('fv*'), mannelijke sprekers en de gewenste leeftijdscategorie ('age2') te selecteren (Corex maakt hier gebruik van zogenaamde *keywords* waarvan verklarende lijsten opgenomen zijn in de data; *age2* betreft de leeftijdscategorie 25 - 34). De zoekresultaten kunnen vervolgens binnen Corex opgeslagen worden als subcorpus.

Met de zoekoptie *Statistics* kan een grote variatie aan frequentiegegevens opgevraagd worden van woordvormen, lemma's, tags, prosodische en fonetische fenomenen of een combinatie daarvan. In figuur 5 is een voorbeeld weergegeven van frequentieresultaten van de lemma's *man*, *vrouw* en *kind*, onder andere binnen het subcorpus uit figuur 4. De *queries* worden eerst één voor één opgeslagen en vervolgens tegelijkertijd uitgevoerd. Let er bij een zoekactie in *Statistics* wel op dat de frequenties absolute getallen zijn en daardoor een vertekend beeld kunnen geven.



Figuur 5: Frequentiegegevens van de lemma's 'man' 'vrouw' en 'kind' in spontane conversaties van Vlaamse mannen tussen de 25 en 34 jaar tegenover spontane conversaties van Vlaamse vrouwen tussen de 25 en 34 jaar.

Naast de *Metadata Search* en de *Statistics Tool* is er binnen Corex een speciale *Lexicon Tool* om in het lexicon te kunnen zoeken. In figuur 6 wordt een schermafbeelding getoond van een lijst diminutieven in de *Lexicon Tool*. Onder de menuoptie *View* kan er een keuze gemaakt worden in het aantal kolommen met woordinformatie dat getoond wordt.



Figuur 6: Weergave van een lijst diminutieven in de Lexicon Tool.

Zoals uit bovenstaande voorbeelden blijkt, benadert iedere zoekfunctie de data op een eigen, zo efficiënt mogelijke wijze. Het verschilt daarom per zoekfunctie welke informatie ingevoerd moet worden om het optimale resultaat te verkrijgen. Er zijn vaak meerdere manieren om resultaten te vinden. Daarvoor moet een tussenweg gevonden worden tussen de

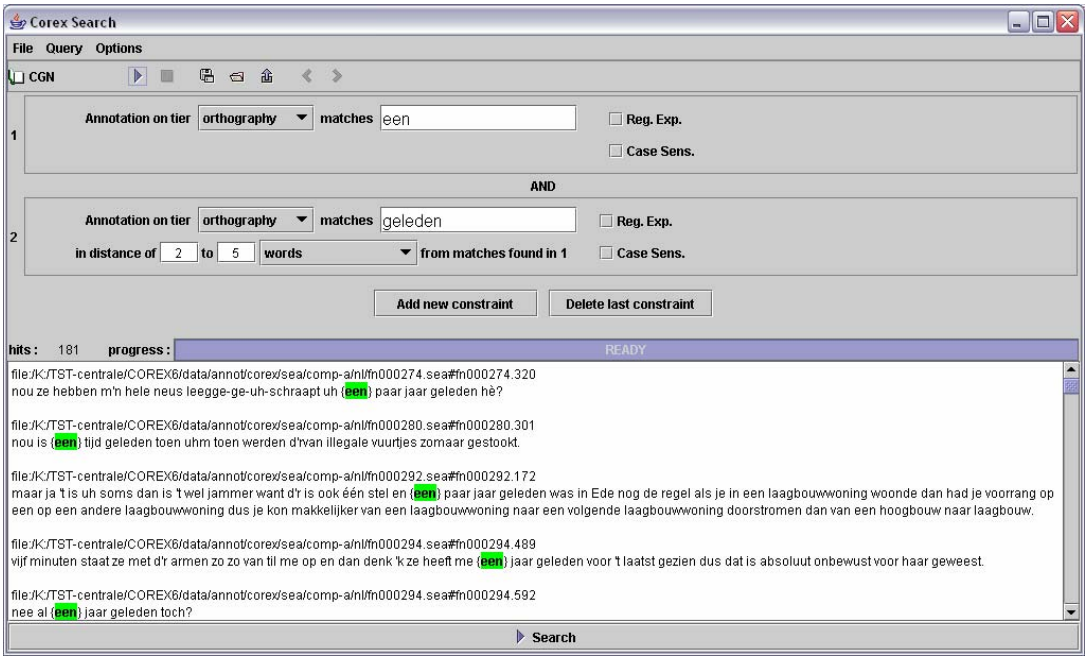
hoeveelheid ruis die optreedt bij een zoekactie en de hoeveelheid data die doorzocht wordt. Aangezien de syntactische annotaties toegevoegd zijn aan maar tien procent van de data, wordt bij een syntactische zoekactie alleen dit deel van het corpus doorzocht. Hierover meer in de uitgewerkte onderzoeksvraag in 4.3.

Corex kan met veel onderzoeksvragen uit de voeten, maar uiteraard zijn er beperkingen. Het is niet mogelijk om negatieve zoekacties uit te voeren, zoals zoeken op woorden die eindigen op *–ig* en die geen bijvoeglijke naamwoorden zijn. De frequentiegegevens zijn ook maar beperkt toegankelijk (zie figuur 5); voor het samenstellen of herschikken van een bepaalde frequentielijst moet er buiten Corex gewerkt worden in een teksteditor of in Excel.

Hierboven zijn drie zoekfuncties van Corex kort behandeld: *Metadata Search*, *Statistics* en *Lexicon Tool.* Twee andere belangrijke zoekfuncties die nog niet eerder genoemd zijn – *Content Search* en *Syntax Search* – worden in 4.1 en 4.2 uitgebreider toegelicht.

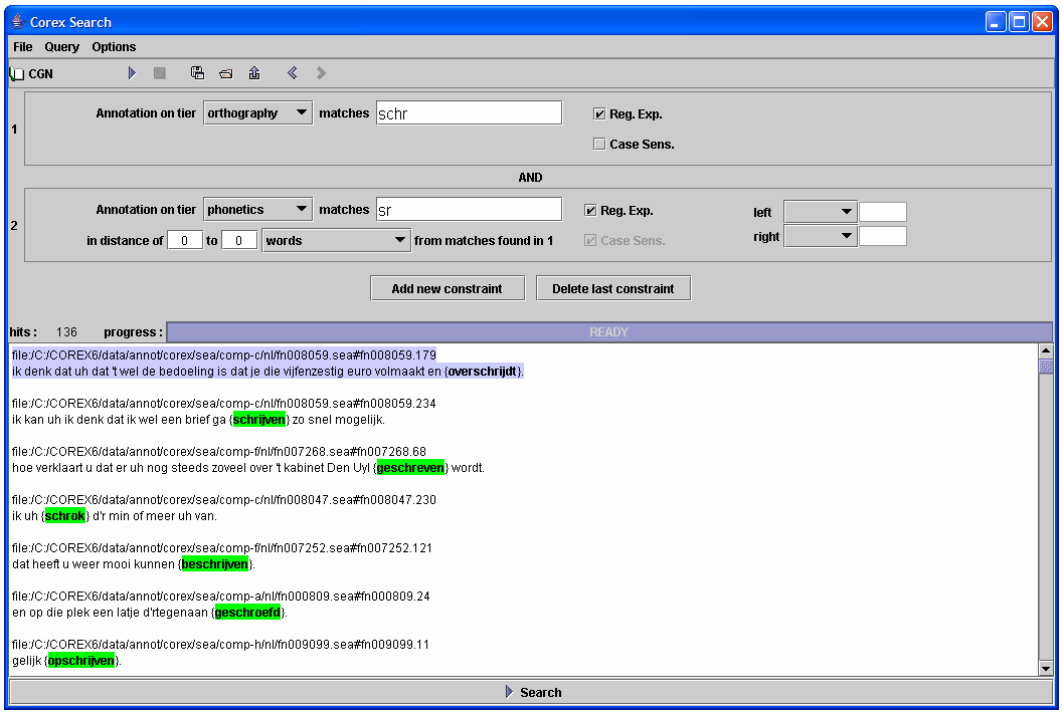
4.1 Content Search

Eén van de meest gebruikte en veelomvattende zoekfuncties binnen Corex, is de *Content Search.* Met uitzondering van de syntactische annotaties, kan met deze functie in elke annotatielaag gezocht worden op specifieke woordvormen, lemma's, POS-tags, prosodische of fonetische fenomenen of apart gemarkeerde spraak zoals dialectwoorden of versprekingen. Daarnaast kunnen meerdere parameters tegelijkertijd gedefinieerd worden: in de orthografie kan het hele corpus doorzocht worden op een enkel woord of op woordcombinaties. Neem bijvoorbeeld de woordcombinatie *een … geleden.* De afstand tussen de woorden *een* en *geleden* kan variabel zijn afhankelijk van de zin waarin de combinatie voorkomt (*een jaar geleden, een hele tijd geleden, een maand of zes geleden*). Corex biedt in dit geval de mogelijkheid om het minimale en/of maximale aantal woorden tussen *een* en *geleden* in te voeren in het zoekscherm. Hoe groter de ingestelde afstand tussen de twee woorden wordt, hoe meer ruis er ontstaat. In dat geval moet er handmatig meer filtering plaatsvinden. In figuur 7 is gekozen voor een afstand van twee tot vijf woorden tussen de twee doelwoorden.



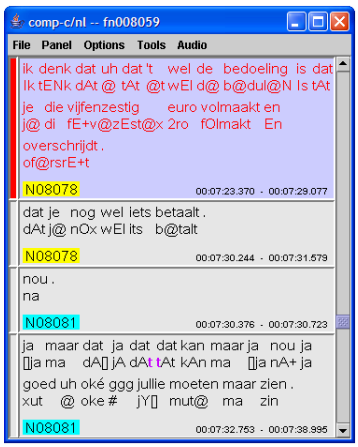
Figuur 7: Het zoekscherm van de Content Search waarin gezocht wordt op de woordcombinatie ' een ... geleden'.

In de *Content Search* zijn meerdere annotatielagen met elkaar te combineren. In de orthografische annotatielaag kunnen we bijvoorbeeld zoeken op woorden beginnend met de cluster <schr>. In de fonetische transcriptielaag kan vervolgens parallel gezocht worden op de uitspraakvariant /sr/. De zoekresultaten zoals weergegeven in figuur 8, geven nu alle 'schr'-clusters weer die uitgesproken worden als /sr/. Let hierbij wel op dat alleen de handmatig gecontroleerde fonetische transcripties toegankelijk zijn via Corex; slechts tien procent van het corpus wordt doorzocht.



Figuur 8: Het zoekscherm van de Content Search waarin gezocht wordt op uitspraakvariant /sr/ van de cluster <schr> (de optie Reg. Exp. zorgt ervoor dat er woordintern gezocht wordt).

Doordat de annotaties door middel van een signaalkoppeling verbonden zijn aan de spraakfragmenten, zijn de zoekresultaten van de *Content Search* direct te bekijken in combinatie met de geluidsbestanden. In de zogenaamde *Corex-viewer* (zie figuur 9) kan het spraaksignaal synchroon worden afgespeeld met de gewenste annotaties. Deze *viewer* kan direct worden geopend vanuit het zoekvenster met de resultaten.

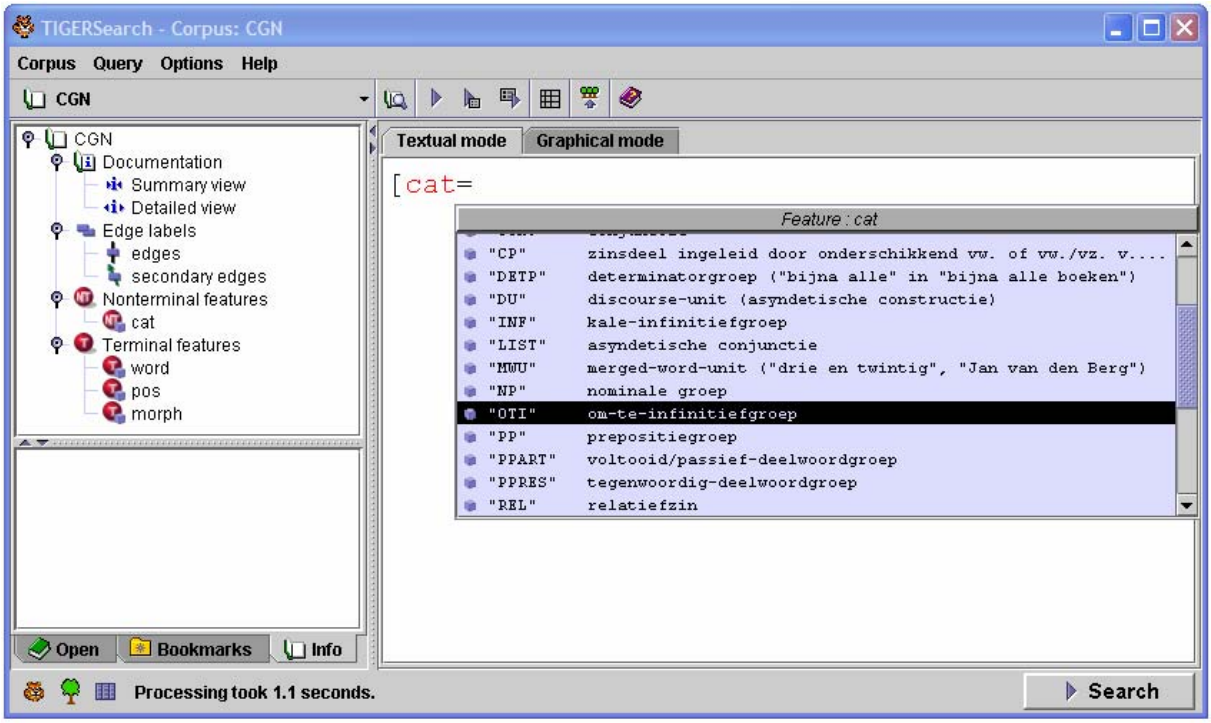


Figuur 9: De corex-viewer, gekoppeld aan het eerste resultaat in figuur 8.

4.2 Syntax Search

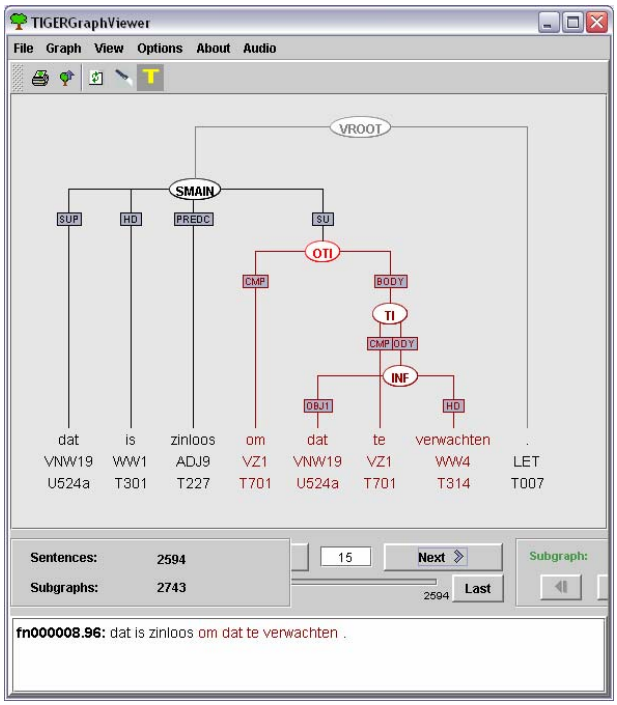
Om de syntactische annotaties te bekijken en te analyseren, is er een speciale applicatie in Corex ingebouwd, genaamd *TigerSearch*. De applicatie wordt via het hoofdscherm geopend door in het menu *Syntax Search* te kiezen. Binnen *Tiger* is alleen dát deel van het corpus toegankelijk, waarvan een syntactische transcriptie beschikbaar is. In *Tiger* kan gezocht worden op woorden, woordsoort en morfologische informatie. Complexere zoekopdrachten in verband met adjacentie, dominantie of grammaticale functie behoren ook tot de mogelijkheden. In 4.3 worden deze begrippen verder verduidelijkt aan de hand van een concrete onderzoeksvraag,

Er zijn twee manieren om in *TigerSearch* te zoeken. De tekstuele modus wordt ten opzichte van de grafische modus het meeste gebruikt. In de tekstuele modus kunnen syntactische zoekvragen gedefinieerd worden met behulp van de *TigerSearch query language* (zie ook paragraaf 4.3). Met behulp van deze taal kan er gezocht worden op dependentiestructuren en syntactische labels. In *TigerSearch* is een beschrijving van alle syntactische labels geïntegreerd en de mogelijke opties verschijnen automatisch wanneer het begin van een zoekvraag in het scherm ingetikt wordt. De syntactische constructie *om te + infinitief* is bijvoorbeeld standaard opgenomen in *Tiger* als categorie ‘OTI’.



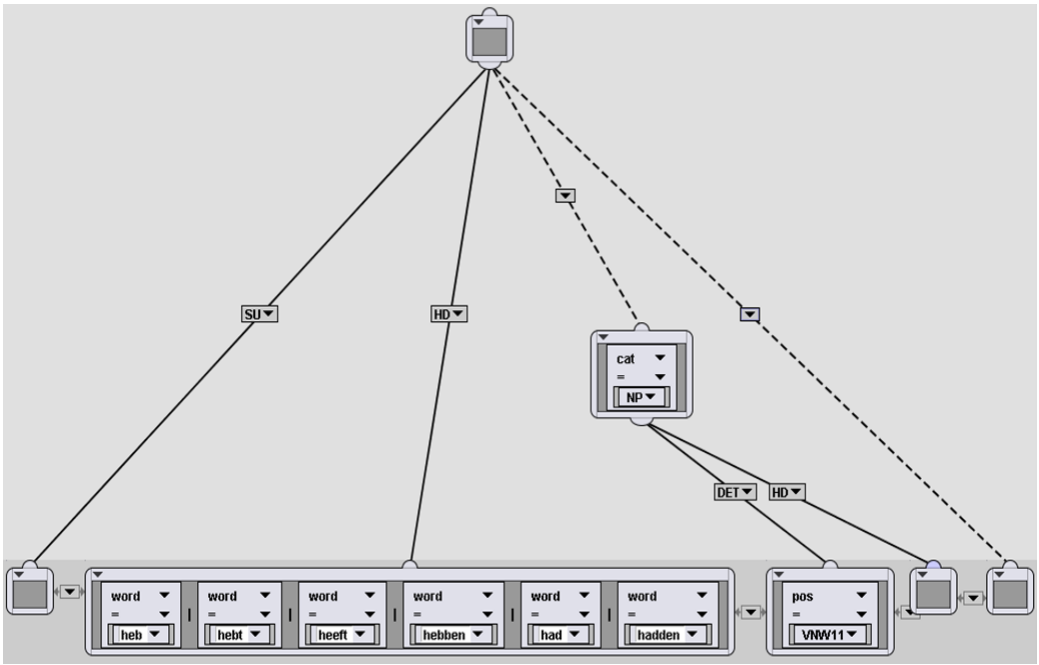
Figuur 10: Zoekvraag in de tekstuele modus van Tiger; om te+infinitiefgroep.

Na uitvoering van de zoekvraag [cat=“OTI”] (figuur 10) verschijnt de *Tigergraph-viewer* waarin de boomstructuren te zien zijn. In figuur 11 wordt één van de resultaten getoond.



Figuur 11: De Tigergraph-viewer; om te+ infinitiefgroep.

In de grafische modus kunnen de afhankelijkheidsrelaties en hiërarchie van een syntactische zoekvraag visueel in kaart gebracht worden. In figuur 12 is de grafische modus afgebeeld waarin gezocht wordt op constructies vergelijkbaar met *ik heb mijn band lek.* De afbeelding uit de grafische modus kan direct omgezet worden naar de tekstuele modus. Andersom is dit niet mogelijk.



Figuur 12: Zoekvraag in de grafische modus van Tiger; SU - hebben - *VNMW11 (bezittelijk voornaamwoord) - HD.*

4 Onderzoeken met het CGN

Het Corpus Gesproken Nederlands wordt vandaag de dag ingezet bij uiteenlopende onderzoeken. De geluidsbestanden worden door verschillende taal- en spraaktechnologiebedrijven gebruikt voor het trainen van spraakherkenners. In het wetenschappelijke veld wordt onderzoek gedaan naar de verbetering van huidige spraakherkenners. In Van Dalen, Wiggers & Rothkrantz (2006) wordt beschreven hoe de automatische herkenning van lexicale klemtoon in continue spraak onderzocht wordt. Een bijzonder aspect aan het onderzoek is dat niet alleen van klinkers, maar ook van consonanten aangenomen wordt dat deze klanken een indicatie van klemtoon met zich meedragen. In het experiment krijgen consonanten daarom in sommige gevallen ook lexicale klemtoon toegewezen. Een dergelijk experiment kan tot aanzienlijke verbeteringen van de Nederlandse spraakherkenning leiden.

Ook binnen de taaltheoretische gebieden van de wetenschap wordt het corpus frequent gebruikt. Van Son, Wesseling & Pols (2006) voerden een experiment uit met betrekking tot beurtwisselingen in spontane gesprekken. De hypothese werd getest, dat de aanwezigheid van geaccentueerde woorden vlak voor een beurtwisseling tot snellere reacties leidt van de gesprekspartner. Uit het CGN zijn hiervoor de spontane dialogen gebruikt. De sprekers werden van elkaar gescheiden met behulp van de woordsegmentatie. In de experimentopstelling kon nu een dialoogsituatie gecreëerd worden, door de combinatie van een opname van één spreker uit het corpus en de proefpersoon. Ook de prosodische annotatie uit het CGN werd gebruikt om de accenten te bepalen.

Naast deze praktische experimenten zijn er ook voorbeelden van meer theoretisch gerichte onderzoeken, zoals dat van Kloots, Gillis & Swerts (2006). Hierin wordt crosslinguïstisch onderzoek gedaan naar onbeklemtoonde vocalen aan het syllabe-einde in het Nederlands van Vlamingen en Nederlanders. In de Nederlandse fonologie wordt in tegenstelling tot de Vlaamse onderscheid gemaakt tussen zogenaamde ‘lange’ en ‘korte’ vocalen. De intuïtie bestaat dat een open syllabe in het Nederlands steeds eindigt op een fonologisch ‘lange’ vocaal. Een perceptie-experiment werd uitgevoerd aan de hand van spontane gesprekken uit het corpus, waaronder interviews met leraren Nederlands.

Bovenstaande beschrijvingen zijn slechts een greep uit de verschillende soorten onderzoek die uitgevoerd zijn met het CGN. Uitgave nummer 12 (2007) van Nederlandse Taalkunde bevat verdere illustraties van gebruiksmogelijkheden van het CGN: op het terrein van de semantiek (zie de bijdrage van Oosterhof en Coussé), in onderzoek naar gebruiksverschillen tussen Belgisch Nederlands en Nederlands Nederlands (zie het artikel van Boogaart e.a.), en bij het vaststellen van ‘Extended Lexical Units’ (Van der Wouden).

5 Het CGN bij de TST-centrale

Na afronding van het project werd het CGN aanvankelijk gedistribueerd door ELDA (Evaluations and Language resources Distribution Agency) maar sinds 2004 is het corpus beschikbaar via de centrale voor Taal- en Spraaktechnologie (TST-centrale). De TST-centrale is op initiatief van de Nederlandse Taalunie opgericht omdat er behoefte was aan een kennis- en distributiecentrum voor digitale taalmaterialen zoals het CGN. De TST-centrale is op dit moment als afdeling ondergebracht bij het Instituut voor Nederlandse Lexicologie met een vestiging in Leiden en Antwerpen. De taalmaterialen worden bij de centrale niet alleen gedistribueerd maar ook onderhouden. Wat het CGN betreft, houdt dat in dat er onder andere talloze fouten in de annotaties, de software en de documentatie verbeterd werden. Deze verbeteringen hebben in 2006 geleid tot de lancering van versie 2.0 van het corpus.

Naast distributie en onderhoud, stimuleert de TST-centrale het gebruik van het CGN en fungeert de centrale tevens als helpdesk. Zo worden er aan universiteiten regelmatig workshops en gastcolleges verzorgd, over de gebruiksmogelijkheden van het CGN en Corex. Momenteel wordt er ook gewerkt aan een online CGN-webcursus, een nieuwe en verbeterde versie van Corex (door Polderland, Nijmegen) en uiteindelijk zal er een online webversie van het CGN beschikbaar komen. Dit alles wordt naar verwachting binnen een jaar gerealiseerd.

De webversie van het CGN is op dit moment in ontwikkeling en er wordt naar gestreefd om zo veel mogelijk functionaliteit van Corex in deze versie te herbergen. Tot in hoeverre álle functies en achterliggende data gratis online beschikbaar komen, moet nog onderzocht worden. Het complete CGN wordt op dit moment voor onderzoek en commerciële doeleinden aangeboden op drieëndertig dvd's: een annotatie-dvd en tweeëndertig geluids-dvd's. De annotatie-dvd bevat onder andere alle transcripties, de metadata, lexica en de frequentiegegevens. Ook de exploitatiesoftware Corex en alle documentatie en handleidingen zijn aanwezig op de annotatie-dvd. Op de tweeëndertig geluids-dvd's staan alle 12.780 geluidsbestanden. De annotatie-dvd is ook te gebruiken zonder de dvd's met spraakbestanden.

De Nederlandse Taalunie heeft een commissie aangesteld die zich bezig gaat houden met het vaststellen van marktconforme prijzen voor taalmaterialen. Voor individuele onderzoekers is de aanschaf van het gehele CGN nu nog een kostbare aangelegenheid. De 'prijzencommissie' zal onderzoeken of zo veel mogelijk materialen voor onderzoeksdoeleinden voor een minimale prijs beschikbaar gesteld kunnen worden. Voor commercieel gebruik zullen de (marktconforme) prijzen echter wel gehandhaafd worden.

En wat staat er verder nog te gebeuren? Het Corpus Gesproken Nederlands is een afgerond project en zal in zijn huidige staat waarschijnlijk niet verder worden aangevuld. Er zijn wel projecten die in lijn van het CGN ontwikkeld worden, zoals bijvoorbeeld het JASMIN-CGN (Cucchiarini en anderen 2006). De afkorting JASMIN staat voor Jongeren, Anderstaligen, Senioren, mens-Machine-Interactie voor het Nederlands. Dit corpus wordt onder leiding van de Radboud Universiteit Nijmegen ontwikkeld in het kader van STEVIN (Spraak- een Taaltechnologische Essentiële Voorzieningen In het Nederlands), een onderzoek- en stimuleringsprogramma voor de Nederlandse en Vlaamse taal- en spraaktechnologie. Het doel van het project is om het CGN uit te breiden, door een nieuw corpus samen te stellen van hedendaags Nederlands zoals gesproken door kinderen van verschillende leeftijdsgroepen, niet-moedertaalsprekers en ouderen. Een bijzondere taalgebruiksituatie die onderdeel vormt van het project JASMIN-CGN, is communicatie tussen mens en computer, oftewel mens-machine-interactie. Een dialoog met een spraakcomputer leidt tot typische fenomenen zoals hyperarticulatie, syllabeverlenging, stemverheffing of klemtoonverschuiving. Deze kunstmatige dialogen vormen samen met de andere kenmerkende spraak uit het JASMIN-CGN een interessante bron voor toekomstig onderzoek naar gesproken Nederlands. Het Corpus Gesproken Nederlands is nog maar het begin.

Bibliografie

Baayen, R.H., R. Piepenbrock & Rijn, H. van (1993). *The CELEX Lexical Database*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Boersma, P.& D. Weenink, (1992-2007). Praat: doing phonetics by computer. <http://www.praat.org/>.

Cucchiarini, **C., H. van Hamme**, **O. van Herwijnen**, **& F. Smits (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*. Genoa, Italy, 135-138.

Dalen, R.C. van, P. Wiggers & L.J.M. Rothkrantz (2006). Lexical Stress in Continuous Speech Recognition. In: *Proceedings Interspeech 2006*, 2382–2385.

Kloots, H., S. Gillis & M. Swerts (2006). Onbeklemtoonde vocalen aan het syllabe-einde in het Standaardnederlands van Vlamingen en Nederlanders. In: *Artikelen van de Vijfde Sociolinguïstische Conferentie* / Koole Tom [edit.], e.a., Delft: Eburon, 296-307.

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

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and first Evaluation. In: *Proceedings LREC 2000*, Genoa, Italy.

Oostdijk, N. (2004). De website van het Corpus Gesproken Nederlands. <http://www.tst.inl.nl/cgndocs/doc_Dutch/start.htm>

***Referentiebestand Nederlands (RBN)* (1998).** Samengesteld onder auspiciën van de Commissie Lexicografische Vertaalvoorzieningen (CLVV). Beschikbaar bij het Instituut voor Nederlandse Lexicologie.

Son, R.J.J.H. van, W. Wesseling & L.C.W. Pols (2006). Prominent Words as Anchors for TRP Projection. In: *Proceedings Interspeech 2006*, paper 1235.

***Woordenlijst Nederlandse taal* (1995).** Samengesteld door het Instituut voor Nederlandse Lexicologie in opdracht van de Nederlandse Taalunie. Met een Leidraad van Jan Renkema. Den Haag/Antwerpen: Sdu Uitgever/Standaard Uitgeverij.