"Een beetje babbelen onder elkaar" Verzameling, verwerking en studie van spontane spraak uit het *Corpus Gesproken Nederlands*

door

Hanne KLOOTS[1]

Abstract

This contribution opens with some general information on the registration of speech and the use of corpora. Then the focus is on spontaneous speech: how is "spontaneous speech" defined, why is it so important to study this type of speech and which problems can arise when studying it? Next, we focus on corpora of spontaneous speech: (a) corpora created before the realization of the Spoken Dutch Corpus ("Corpus Gesproken Nederlands" or CGN), (b) the CGN itself and (c) a subcorpus of the CGN, the so-called "teachers'corpus". Finally we discuss some problems which arose in two studies, based on spontaneous speech from the "teachers'corpus": a study on vowel reduction and a study on speech rate. Keywords: Dutch, (corpora of) spontaneous speech, Spoken Dutch Corpus, language variation.

1. REGISTRATIE VAN GESPROKEN TAAL

Een basisvoorwaarde voor een systematische studie van gesproken taal zijn degelijke geluidsopnamen. We kunnen het ons vandaag haast niet meer voorstellen, maar er zijn tijden geweest waarin geluid in het algemeen en gesproken taal in het bijzonder niet kon worden vastgelegd op een geluidsdrager. Wie gesproken taal uit die (oudere) periodes wil bestuderen, beschikt dus – in het beste geval – alleen over geschreven bronnen. Maar hoe nauwkeurig we die geschreven bronnen ook analyseren, elke uitspraak over de precieze klankwaarde van grafemen is en

[1] De auteur is als postdoctoraal onderzoeker van het Fonds voor Wetenschappelijk Onderzoek – Vlaanderen verbonden aan het Centrum voor Nederlandse Taal en Spraak (Universiteit Antwerpen). Met dank aan Griet Depoorter, Steven Gillis en Jacques Van Keymeulen voor hun commentaar bij een eerdere versie van dit artikel.

blijft een reconstructie (Van Bree 1996:69). Ook wat de woordenschat, morfologie en syntaxis van de gesproken taal in oudere taalstadia betreft, zullen we wellicht nooit veel verder komen dan een zgn. *educated guess*, al hangt veel natuurlijk ook af van het *type* schriftelijke bron dat als basis wordt genomen. Zo bevatten getuigenverhoren ongetwijfeld veel meer spreektalige elementen dan wetboeken (zie bv. Vanacker 1963).

Vanaf de 19de eeuw "beginnen de bronnen [...] ruimer te vloeien" (De Vooys 1970:171). In deze periode verschijnen er namelijk voor het eerst een aantal romans waarin gesprekken zo natuurgetrouw mogelijk worden weergegeven, denken we bijvoorbeeld maar aan *Jellen en Mietje* (K. Broeckaert), de *Studententypen* (Klikspaan), de *Camera Obscura* (Hildebrand) en *Woutertje Pieterse* (Multatuli). Deze romans zijn bijzonder interessant vanuit taalkundig perspectief, al moeten we toch ook hier voorzichtig blijven. Omdat de auteurs gebonden waren aan de letters van het gewone alfabet konden ze bijvoorbeeld niet alle klanknuances weergeven. Ook de consistentie in de (klank)weergave laat soms wat te wensen over (zie bv. Van Haeringen 1971). Verder wilden de auteurs hun tekst natuurlijk ook zo leesbaar mogelijk houden.

Pas vanaf het einde van de 19de eeuw wordt (stem)geluid ook vastgelegd op een geluidsdrager, "waarna het op de studeerkamer rustig kan worden nagestudeerd" (Weijnen 1966:168). Eerst was er de fonograaf, een toestel waarbij het geluid werd vastgelegd op wasrollen (cilinders). Later werd gewerkt met schijven uit paraffine (Zwaardemaker & Eijkman 1928). In een volgend stadium kon het geluid worden opgenomen op magnetische banden. Niet iedereen was meteen overtuigd van het belang en de mogelijkheden van de bandrecorder voor taalkundig onderzoek, denken we bijvoorbeeld maar aan directeur Beerta uit *Het Bureau* die – anno 1957 – "die moderne techniek maar griezelig" vond (Voskuil 1997:78). Na de bandrecorder deed de audiocassette haar intrede. Vanaf de tweede helft van de jaren 80 werd het geluid steeds vaker digitaal opgenomen. Eerst kwam de datrecorder in gebruik, later ook de MiniDisk en de solid state recorder. Digitale audiofiles worden achteraf gearchiveerd op cd-roms of een harde schijf, vanaf het einde van de jaren 90 ook op dvd (zie ook Piepenbrock 1999).

2. Corpora

Wie patronen, tendensen en veranderingen in gesproken taal op het spoor wil komen, kan – zoals elke taalwetenschapper – kiezen tussen

een intuïtieve of een empirische benadering. Ruwweg samengevat: in het eerste geval probeert de taalkundige zijn/haar doel te bereiken via introspectie, in het tweede geval gebruikt hij/zij observaties van concreet talig gedrag als uitgangspunt (Sampson 2001). Twee kernbegrippen uit de (recente) empirische taalkunde zijn *corpus* en *computer* (Biber e.a. 1998). Een corpus is een verzameling talige gegevens (i.e. gedrukte teksten, opnamen en/of transcripties van gesproken taal) die als startpunt kan dienen voor een taalkundige beschrijving. Verder kunnen corpora ook uitstekende diensten bewijzen wanneer we bestaande taalkundige hypotheses en/of intuïties willen toetsen (Crystal 2003:112). In de corpuslinguïstiek wordt volop gebruik gemaakt van computers. Hun grootste troef is dat ze op relatief korte termijn grote hoeveelheden data kunnen analyseren. Bovendien zijn computationele analyses doorgaans erg consistent: computers zijn immers nooit verstrooid, raken niet vermoeid en veranderen niet halverwege een analyse van strategie.

Aan het werken met corpora van gesproken taal zijn echter ook enkele problemen verbonden. Als de onderzoeker vertrekt van bestaande transcripties kan hij/zij bijvoorbeeld alleen die variabelen bestuderen waarvan de respectieve varianten effectief weergegeven zijn in (of kunnen worden afgeleid uit) de transcripties (zie ook Bauer 2002). Een ander probleem betreft de betrouwbaarheid van de beschikbare transcripties en annotaties. Zolang die betrouwbaarheid niet geverifieerd is, heeft de corpustaalkundige eigenlijk ook geen goed zicht op de waarde van zijn eigen onderzoeksresultaten (cf. § 4.2). Ten slotte kunnen via deze empirische benadering alleen maar variabelen bestudeerd worden die effectief in het corpus voorkomen. Laagfrequente items zijn dus een stuk moeilijker te bestuderen dan hoogfrequente (cf. § 3.3).

3. Studie van spontane spraak

In deze paragraaf wordt het begrip "spontane spraak" eerst zo nauwkeurig mogelijk omschreven (§ 3.1). Vervolgens kijken we *waarom* het belangrijk is om spontane spraak te bestuderen (§ 3.2) en met welke *problemen* we te maken (kunnen) krijgen als we spontane spraak onderzoeken (§ 3.3).

3.1 Definitie en varianten

Het begrip "spontane spraak" wordt vaak ruwweg omschreven als *niet-voorgelezen* spraak (Fagyal 1995, Beckman 1996, Laan 1997). In de praktijk is er echter sprake van een soort continuüm: we kunnen ver-

schillende types van spontane spraak onderscheiden, variërend van sterk gecontroleerd tot nagenoeg ongecontroleerd. Het verschil tussen de respectieve types van spontane spraak zit hem vooral in de mate waarin de onderzoeker invloed wil c.q. kan uitoefenen op inhoud en vorm van de geproduceerde spraak. Bij matig tot sterk gecontroleerde spontane spraak denken we bijvoorbeeld aan de beschrijving van een plaatje of route, het navertellen van een verhaal, het bevragen van een computerdatabase of het geven van instructies. Bij toespraken, preken, telefoongesprekken of open vragen heeft de onderzoeker al een stuk minder controle over inhoud en vorm van de spraak. Totaal ongecontroleerde spraak wordt verkregen door stiekem geluidsopnamen te maken. Het is echter de vraag in hoeverre deze laatste benadering wel ethisch en juridisch verantwoord is (zie bv. Weijnen 1966, Wodak 1982).

Bij de keuze van de elicitatietechniek kan de onderzoeker ook rekening houden met praktische aspecten zoals geluidskwaliteit en transcribeerbaarheid. Concreet voorbeeld: toespraken worden vaak gehouden in een grote ruimte en in aanwezigheid van een grote groep mensen, wat de opnamekwaliteit niet altijd ten goede komt. Wie doorgedreven akoestische analyses gepland heeft, kiest dan ook beter voor een type spontane spraak dat in een kleine, geluidsarme ruimte kan worden opgenomen. Een tweede voorbeeld: een onvoorbereide telefoondialoog tussen twee bekenden is doorgaans veel moeilijker te transcriberen dan de beschrijving van een (eenvoudig) plaatje. Zo zullen in het telefoongesprek beide gesprekspartners soms tegelijk aan het woord zijn, bijvoorbeeld wanneer de ene beller de andere in de rede valt. Bij de beschrijving van een plaatje door een enkele informant is de kans op zo'n "overlap" per definitie uitgesloten. In het telefoongesprek is de kans ook groter dat er mensen of gebeurtenissen ter sprake worden gebracht die de onderzoeker niet meteen kan thuisbrengen, wat eveneens voor oponthoud kan zorgen bij de transcriptie.

3.2 Belang

De studie van spontane spraak is van groot belang voor zowel taalkundigen als voor taaltechnologen. Veel taalkundige studies zijn gebaseerd op spraak die op een min of meer gestuurde manier ontlokt is. Aan de informanten wordt bijvoorbeeld gevraagd om een aantal woorden of zinnen voor te lezen, een reeks plaatjes te benoemen, zinnen aan te vullen of te vertalen. Daarnaast is echter ook onderzoek nodig van de onvoorbereide, gesproken taal zoals die gebruikt wordt in c.q. voor de dagelijkse communicatie (zie bv. Rischel 1992, Kohler 2000). De stu-

die van zulke "spontane" spraak kan ons niet alleen meer inzicht geven in processen van taalontwikkeling en -verandering, maar ook in de processen die aan de basis liggen van taalproductie en taalperceptie.

Verder kan de studie van spontane spraak bijdragen tot de verbetering van zowel spraakherkenning als spraaksynthese (zie bv. Laan 1997, Strik 2001). Automatische spraakherkenners hebben vaak moeite met de uitspraakvariatie in spontane spraak, in het bijzonder met de insertie, reductie en deletie van klanken. Text-to-speechsystemen, i.e. systemen die een geschreven tekst automatisch omzetten in het gesproken equivalent, leveren vaak spraak op die als voorgelezen klinkt. De oorzaak ligt in beide gevallen bij de manier waarop de betreffende systemen getraind werden: zowel spraakherkenners als text-to-speechsystemen werden tot nu toe namelijk vooral (of zelfs uitsluitend) getraind op voorgelezen spraak. Om de kwaliteit van de systemen te verbeteren zouden trainingssessies met spontane spraak moeten worden georganiseerd. Verder zou bij de ontwikkeling van taaltechnologische toepassingen gebruik kunnen worden gemaakt van inzichten uit het taalkundig onderzoek naar (variatie in) spontane spraak.[2]

3.3 Problemen

Wie aan de slag gaat met spontane spraak krijgt onvermijdelijk te maken met een aantal praktische problemen. Het voornaamste is de geringe controle over de materiaalverzameling. De onderzoeker kan onmogelijk vooraf voorspellen *welke* variabelen (i.e. klanken, woorden, constructies) in de spraak zullen voorkomen, *hoe frequent* ze zullen voorkomen en of de uiteindelijke realisaties ook effectief *bruikbaar* zullen zijn voor de geplande analyse. Dit alles maakt spontane spraak minder geschikt voor de studie van laagfrequente variabelen (zie bv. Milroy & Gordon 2003). Bovendien is het ook totaal onvoorspelbaar *waar* en *wanneer* variabelen zullen voorkomen. Het inventariseren en opsporen van de variabelen is dan ook een bijzonder tijdrovend proces.

Andere factoren die niet onder controle kunnen worden gehouden, zijn bijvoorbeeld intonatie, spreeksnelheid, pauzering en emotionele toestand van de spreker. Ook de kwaliteit van de geluidsopnamen is wisselend. Afhankelijk van de opnamelocatie valt namelijk niet uit te

[2] Een recent voorbeeld van zo'n initiatief was het project *Flexible Large Vocabulary Recognition* (2002-2006), een samenwerkingsproject tussen taalkundigen van de Universiteit Antwerpen en ingenieurs van de KU Leuven. Dit project had precies tot doel om een bestaande spraakherkenner te verbeteren door er taalkundige inzichten in te integreren.

sluiten dat er samen met de spraak ook achtergrondgeluiden worden opgenomen zoals telefoongerinkel of mussengetjilp. Bovendien bevat spontane spraak ook heel wat talige "onvolkomenheden": afgebroken woorden en zinnen, ongrammaticale zinnen, versprekingen en valse starts, onverwachte pauzes en aarzelingen, herhalingen en stopwoorden.

Ten slotte wordt de onderzoeker van spontane spraak ook geconfronteerd met de zogeheten *Observer's Paradox*: de onderzoeker wil observeren hoe de informant klinkt als hij/zij niet geobserveerd wordt, maar de aanwezigheid van een onderzoeker en/of opnameapparatuur maakt de communicatieve situatie – en dus ook de ontlokte spraak – echter per definitie minder natuurlijk (zie bv. Labov 1972, Milroy & Gordon 2003, Crystal 2003). De vrees bestaat dat informanten enigszins geremd zullen zijn als ze weten dat hun spraak wordt opgenomen en/of geanalyseerd. Ook in beschrijvingen van dialectologisch veldwerk werd er al geregeld op gewezen dat bij informanten "microfoonangst" (Floris 1997:88) kan optreden (zie ook Grootaers 1926, Blancquaert 1948).

4. Corpora van spontane spraak

In deze paragraaf focussen we op een aantal corpora van spontane spraak: corpora uit de periode vóór de totstandkoming van het *Corpus Gesproken Nederlands* (§ 4.1), het *Corpus Gesproken Nederlands* zelf (§ 4.2) en een subcorpus van het *Corpus Gesproken Nederlands*, het zgn. "lerarencorpus" (§ 4.3).

4.1 Het pre-CGN-tijdperk

Wie vandaag spontaan Standaardnederlands wil bestuderen, doet meestal een beroep op het *Corpus Gesproken Nederlands* (cf. §4.2). Dat betekent echter niet dat er voorheen nooit een corpus van spontaan gesproken Nederlands zou zijn samengesteld, of dat er nog nooit eerder spontane spraak werd bestudeerd.

Het bekendste spreektaalcorpus uit het pre-CGN-tijdperk is waarschijnlijk dat van De Jong (1979), een verzameling van formele en informele gesprekken met geboren en getogen Amsterdammers, geregistreerd in 1975 en 1976 (zie ook Heikens 1978). Een andere bekende collectie van spontane spraak bevindt zich in Gent. In de jaren 60-70 werd daar door het toenmalige Seminarie voor Nederlandse Taalkunde en Vlaamse Dialectologie (Universiteit Gent) een verzame-

ling aangelegd van gesprekken in het dialect (Van Keymeulen 2002). De Gentse veldwerkers concentreerden zich op de dialecten van Noord-België en Frans-Vlaanderen. Om het materiaal te bewaren voor de toekomst werd in 2001 begonnen met de digitalisering van de opnamen. Het Meertens Instituut in Amsterdam beschikt over een vergelijkbare verzameling voor Nederland. Wat spontane kindertaal betreft, is het Nederlands zelfs een van de best gedocumenteerde talen ter wereld. Nederlandstalige corpora van kindertaal worden al jaren verzameld in de internationale databank CHILDES[3]. Andere voorbeelden van gesproken taalcorpora uit het pre-CGN-tijdperk zijn te vinden in Piepenbrock (1999).

Vóór de totstandkoming van het CGN werden ook al studies uitgevoerd, gebaseerd op spontane spraak. Zo verwezen Vanacker & De Schutter (1967) naar een aantal Gentse licentiaatsverhandelingen, gebaseerd op spontane spraak uit bovengenoemde dialectarchieven. Andere onderzoekers stelden hun eigen corpus van spontane spraak samen, al dan niet gebaseerd op bestaande opnamen. Bekend is bijvoorbeeld het proefschrift van Van de Velde (1996), die uitspraakvariatie onderzocht in de spraak van Vlaamse en Nederlandse radiopresentatoren. Een ander voorbeeld is het proefschrift van Ernestus (2000), die vocaalreductie bestudeerde in de spontane spraak van 16 hoogopgeleide mannen uit het westen van Nederland.

Wie aan de slag wil gaan met corpora uit het pre-CGN-tijdperk kan te maken krijgen met (ten minste) twee praktische problemen. Die problemen hebben betrekking op de beschikbaarheid en de representativiteit van het geluidsmateriaal. Wat de beschikbaarheid betreft, stellen we vast dat een aantal opnamen uit het pre-CGN-tijdperk vandaag eenvoudig niet meer raadpleegbaar zijn, bijvoorbeeld omdat ze niet tijdig gedigitaliseerd zijn en de kwaliteit van de oorspronkelijke bandopnamen intussen te wensen overlaat (zie ook Piepenbrock 1999). Soms is het geluidsmateriaal zelf nog wel beschikbaar, maar is het – bv. om financiële redenen – slechts gedeeltelijk getranscribeerd (zie bv. Van Keymeulen 2002). Dat maakt het een stuk moeilijker om het corpus te raadplegen, zeker voor iemand die niet zelf bij de materiaalverzameling betrokken was. Verder stelt lang niet iedere onderzoeker die een corpus samenstelt zijn/haar materiaal achteraf ook ter beschikking van collega-onderzoekers. Of anders geformuleerd: lang niet elk corpus is ook een *publiek* corpus (Bauer 2002).

[3] Meer informatie over CHILDES is te vinden via <http://childes.psy.cmu.edu>.

Een tweede probleem bij de corpora uit het pre-CGN-tijdperk is hun representativiteit. Bij de samenstelling van de meeste corpora is bijvoorbeeld niet systematisch rekening gehouden met variabelen als sekse, leeftijd en regionale achtergrond. Dat is vooral vervelend voor onderzoekers die geïnteresseerd zijn in taalvariatie: die hebben immers nood aan corpora waarin beide seksen, zo veel mogelijk generaties en zo veel mogelijk regio's vertegenwoordigd zijn. Verder is taal natuurlijk voortdurend in beweging: oudere corpora van spontane spraak hebben een grote historische waarde, maar zijn niet (meer) representatief voor het hedendaagse taalgebruik. Concreet voorbeeld: als we zicht willen krijgen op het hedendaagse taalgebruik in Vlaanderen en Nederland hebben we weinig aan het spreektaalcorpus van De Jong (1979). Deze collectie bevat namelijk alleen spraak van Amsterdammers, verzameld in 1975 en 1976. Ten slotte moeten we nog een onderscheid maken tussen corpora van standaardtaal en dialect. Zo zijn de dialectopnamen van de Universiteit Gent bijzonder geschikt voor dialectologisch onderzoek, maar ze kunnen onmogelijk als basis dienen voor een studie naar (variatie in) de standaardtaal.

Samengevat: ook in het pre-CGN-tijdperk werd al spontane spraak verzameld en bestudeerd, maar wie op het einde van de 20ste eeuw nood had aan een uitgebreide en recente collectie spontaan gesproken Standaardnederlands van volwassenen, een collectie die vlot en in digitale vorm raadpleegbaar was, waarbij transcripties beschikbaar waren en waarin alle leeftijdsgroepen, seksen en regio's vertegenwoordigd waren, vond niet meteen wat hij/zij zocht. Precies deze leemten wilden de samenstellers van het *Corpus Gesproken Nederlands* (§ 4.2) invullen.

4.2 Het Corpus Gesproken Nederlands (CGN)

Het *Corpus Gesproken Nederlands* is een verzameling van ca. negen miljoen woorden gesproken Standaardnederlands, tot stand gekomen in de periode 1999-2003.[4] In totaal gaat het om ongeveer 900 uur spraak. Een derde van die spraak is afkomstig uit Vlaanderen, twee derde uit Nederland. Het CGN wordt gebruikt door taal- en spraaktechnologen, maar het is ook een ware goudmijn voor zowel taalkundigen als (cultuur)historici (Oostdijk 2000). Het corpus geeft namelijk een uitstekend beeld van het taalgebruik van volwassen sprekers in Vlaanderen en Nederland rond de millenniumwisseling.

[4] Informatie over het *Corpus Gesproken Nederlands* is te vinden via de CGN-website <http://lands.let.kun.nl/cgn/home.htm> en via de *Centrale voor Taal- en Spraaktechnologie* <http://www.tst.inl.nl>.

In het CGN zijn verschillende types van spraak vertegenwoordigd. Het corpus bevat zowel monologen (bv. lezingen) als dialogen (bv. telefoongesprekken), zowel opnamen in een privésituatie (zgn. "face-to-face conversaties") als spraak die tot een grote schare toehoorders werd gericht (bv. preken), zowel uitgezonden materiaal (bv. sportcommentaren) als niet-uitgezonden fragmenten (bv. lessen), zowel voorbereide spraak (bv. voorgelezen boeken) als spontane spraak (bv. interviews met leraren Nederlands). In § 4.3 zal het laatstgenoemde type spraak – i.e. de interviews met de leraren Nederlands – nader worden toegelicht.

Alle spraak uit het CGN is ten minste orthografisch getranscribeerd. Die orthografische transcriptie is "een woordelijke neerslag van hetgeen er gezegd is" (Goedertier & Goddijn 2000), waarbij zo veel mogelijk de gewone Nederlandse spellingregels gevolgd werden. Van ongeveer 10% van de spraak bestaat ook een geverifieerde (brede) fonetische transcriptie. Bij (eveneens) 10% is een syntactische annotatie voorhanden. Prosodische annotaties zijn beschikbaar voor ongeveer 2,5% van het materiaal. Helaas is er tot op heden nauwelijks onderzoek gedaan naar de betrouwbaarheid van de CGN-transcripties en -annotaties. Gezien het grote aantal onderzoekers dat vandaag gebruik maakt van het CGN zou het erg zinvol zijn als de consistentie en de kwaliteit van de transcripties en annotaties – ten minste voor een gedeelte van het materiaal – systematisch gecontroleerd zou worden. Een voorzet hiertoe werd gegeven in Schiel (2005) en Coussé & Gillis (2006).

Het CGN wordt momenteel beheerd en onderhouden door de TST-centrale (Centrale voor Taal- en Spraaktechnologie), die op haar beurt deel uitmaakt van het Instituut voor Nederlandse Lexicologie. De rechten zijn in handen van de Nederlandse Taalunie. In het CGN kunnen tegenwoordig gerichte zoekopdrachten worden uitgevoerd met behulp van de exploitatiesoftware COREX.

4.3 Focus: het lerarencorpus

Dat de standaarduitspraak van Vlamingen en Nederlanders in een aantal opzichten verschilt, was genoegzaam bekend. Het onderzoek naar die uitspraakverschillen was (en is) echter vaak uitsluitend gericht op de taal van radio- en televisiepresentatoren (zie bv. Van Oss & Gussenhoven 1984, Cassier & Van de Craen 1986, Van de Velde 1996, Smakman 2006). In het VNC-project *De uitspraak van het Standaardnederlands: variatie en varianten in Vlaanderen en Nederland* (1998-2001), een samenwerkingsproject tussen de Universiteit Antwerpen (UIA) en de Katholieke Universiteit Nijmegen, stond voor het eerst een andere groep van sprekers centraal: leerkrachten

Nederlands. Hun taalgebruik is des te interessanter omdat leraren Nederlands over het algemeen beschouwd worden als prototypische standaardtaalsprekers (Smakman & Van Bezooijen 1997, Van de Velde & Houtermans 1999). Net als nieuwslezers zijn het professionele taalgebruikers die een voorbeeldfunctie vervullen op het vlak van taal. De keuze voor leraren Nederlands had als bijkomend voordeel dat de socio-economische status van de informanten min of meer constant bleef.

Bij de selectie van de proefpersonen werd rekening gehouden met de sociale variabelen *land*, *regio*, *leeftijd* en *sekse*. Er werd spraak verzameld van 80 Vlamingen en 80 Nederlanders. Zowel in Vlaanderen als in Nederland werden vier regio's geselecteerd: een "centrumzone" (VL: Antwerpen/Brabant; NL: Randstad), een zgn. "intermediaire" zone (VL: Oost-Vlaanderen; NL: Gelderland/Utrecht) en twee "perifere" zones (VL: West-Vlaanderen, Limburg; NL: Groningen/Drenthe, Limburg). In elke regio werden twee of meer steden geselecteerd.[5] Alle leraren die in 1998 Nederlands onderwezen in een van deze steden, werden uitgenodigd om deel te nemen aan het uitspraakonderzoek. Bij de selectie van regio's en steden werd rekening gehouden met criteria als dialectbasis en verzorgingsfunctie[6].

In Vlaanderen werden uiteindelijk alleen leraren geselecteerd die hun hele leven in dezelfde regio hadden gewoond. In Nederland werd als voorwaarde gesteld dat de leraren voor hun achtste verjaardag in het gebied waren komen wonen, en er voor hun 18de minstens acht jaar gewoond hadden. In elke regio werden tien "oudere" en tien "jongere" leerkrachten gezocht, van wie telkens vijf mannen en vijf vrouwen. Wie geboren was voor 1955 werd tot de "oudere" generatie gerekend. De "jongeren" zijn geboren na 1960. Voor meer informatie over de opzet van het uitspraakproject, de onderzochte variabelen en de criteria bij de selectie van regio's, steden en leraren, zie Van Hout e.a. (1999) en Kloots (2005).

In de loop van 1999 werd van de leerkrachten Nederlands een sociolinguïstisch gefundeerd interview afgenomen. Het interview bestond uit een gestuurd deel (vnl. voorleestaken), een semi-gestuurd

[5] In Vlaanderen werden alle middelbare scholen gecontacteerd van Lier en Heist-op-den-Berg (Antwerpen/Brabant), Tongeren en Bilzen (Limburg), Ieper en Poperinge (West-Vlaanderen) en Oudenaarde en Zottegem (Oost-Vlaanderen). In Nederland werd gewerkt via scholen uit Alphen a/d Rijn en Gouda (Randstad), Veenendaal, Ede, Tiel, Culemborg en Elst (Gelderland/Utrecht), Assen, Veendam en Winschoten (Groningen/Drenthe) en Geleen, Sittard en Roermond (Limburg).

[6] De term "verzorgingsfunctie" verwijst naar de aanwezigheid van (o.a.) winkelcentra, ziekenhuizen, scholen, centra voor dienstverlening en bioscopen.

gedeelte (bv. plaatjes benoemen) en een spontaan deel (gesprek tussen leerkracht en interviewer). De spontane spraak – of zoals een van de Vlaamse proefpersonen het noemde: "een beetje babbelen onder elkaar" – is later opgenomen in het *Corpus Gesproken Nederlands*. Deze spraak wordt hier kortweg het "lerarencorpus" genoemd.

De interviews werden afgenomen door twee projectmedewerkers: in Nederland door een Nederlander (drs. Ton van Hoek), in Vlaanderen door een Belgische (de auteur). Beiden hadden een taalkundige opleiding genoten, waren nagenoeg even oud en spraken tijdens het interview consequent Standaardnederlands. De opnamen werden gemaakt met behulp van een Tascam DA-P1 draagbare datrecorder en AKG-C420 headsets met een condensatormicrofoon. De interviews vonden plaats in een rustige ruimte op de universiteit, op school of bij de leraar thuis.

De interviewers deden zoveel mogelijk aan "participerende observatie" (zie bv. Fasold 2001, Wodak 1982). Ze probeerden zelf zo weinig mogelijk te zeggen en namen pas het woord als de leerkracht uitgepraat was. De vragenlijst die de interviewers hadden voorbereid, fungeerde louter als inspiratiebron. In de praktijk probeerden de interviewers zo veel mogelijk in te haken op onderwerpen die de proefpersonen zelf aanbrachten. Onderwerpen die geregeld aan bod kwamen, waren bijvoorbeeld literatuur, theater, vakantie(plannen), (onderwijs)actualiteit en hobby's. Van eventuele "microfoonangst" (Floris 1997:88) leken onze leraren nauwelijks last te hebben.

Een factor die bij de samenstelling van toekomstige corpora zeker extra aandacht verdient, is de invloed van de sekse van de gesprekspartners. De Nederlandse interviewer had het gevoel dat de gesprekken met de vrouwelijke informanten iets vlotter verliepen, terwijl de Vlaamse interviewster juist het idee had dat de gesprekken met de mannelijke proefpersonen wat soepeler liepen. Eigenaardig genoeg is over deze factor nauwelijks informatie te vinden in de taalkundige literatuur, al lijken sommige onderzoekers toch ook niet helemaal uit te sluiten dat een gesprek tussen een man en een vrouw op een of andere manier verschilt van een conversatie tussen twee seksegenoten (zie bv. Heikens 1978, De Jong 1979). Ook bij de samenstelling van het CGN werd geen rekening gehouden met deze variabele. Wel zijn over dit onderwerp een aantal boeiende sociologische studies verschenen (zie bv. Catania e.a. 1996, Hansen & Schuldt 1982).[7]

[7] Met dank aan Sonja Spee, voormalig medewerkster van het Centrum voor Vrouwenstudies (Universiteit Antwerpen), die me op het spoor bracht van literatuur in verband met de invloed van de sekse van de interviewer.

5. Aan de slag met spontane spraak

In deze paragraaf bespreken we twee studies, gebaseerd op de spontane spraak uit het "lerarencorpus": een onderzoek naar vocaalreductie (§ 5.1) en een onderzoek naar spreeksnelheid (§ 5.2). Meer specifiek zullen we nagaan met welke onverwachte praktische problemen – verbonden met het gebruik van spontane spraak – we in deze studies te maken kregen.

5.1 Casestudy 1: onderzoek naar reductieverschijnselen

Vocaalreductie kan zowel fonetisch als fonologisch geïnterpreteerd worden. Voor fonetici gaat het om een *gradueel* proces dat kan worden omschreven als "making the pronunciation of a vowel shorter, less loud, lower in pitch and more central in quality" (Laver 1995:157). Fonologen benaderen het verschijnsel eerder categorisch: voor hen gaat het om *substitutie* van de ene klank door de andere (zie bv. Van Bergem 1995). In de literatuur worden drie types van klinkerreductie onderscheiden: verkorting, verdoffing en deletie. Bij verkorting wordt een (fonologisch) lange klinker vervangen door een (fonologisch) korte, bijvoorbeeld *m*[o]*ment* > *m*[ɔ]*ment*.[8] Bij verdoffing wordt een volle vocaal vervangen door sjwa (bv. *moment* > *m*[ə]*ment*). Bij deletie valt de klinker helemaal weg (bv. *moment* > '*ment*).

In het proefschrift van Kloots (2005) werd de uitspraak onderzocht van beklemtoonde en onbeklemtoonde vocalen in open syllaben uit bisyllabische woorden met twee volle vocalen, bijvoorbeeld *mo̲ment*, *dicte̲e*, *the̲ma̲* en *ni̲veau̲*. Alle woorden van dit type moesten een voor een worden opgespoord in het lerarencorpus. Omdat niet voorspelbaar was welke woorden we konden verwachten, kon die opsporing niet geautomatiseerd worden. Er zat dus weinig anders op dan de orthografische transcripties, beschikbaar via het CGN, herhaaldelijk door te lezen, op zoek naar woorden die aan de bovengenoemde criteria voldeden. Vervolgens werd in het geluidssignaal voor en achter elk woord een grens geplaatst met behulp van de software *Praat*. Ook dat was geen sinecure: in spontane spraak vloeien woorden namelijk nogal eens in elkaar over. Ten slotte werden alle stimuli opgeslagen als afzonderlijke geluidsfiles met een unieke bestandsnaam.

[8] We spreken van *fonologisch* lange en korte klinkers omdat de begrippen "lang" en "kort" hier niet geïnterpreteerd mogen worden in termen van vocaalduur. Een aantal klinkers wordt namelijk tot de fonologisch lange klinkers gerekend, terwijl ze in het Standaardnederlands toch een korte duur hebben (m.n. /i/, /u/ en /y/).

Vervolgens werden de woorden een voor een beluisterd en beoordeeld door drie luisteraars[9] via een internetapplicatie. De luisteraars maakten een keuze uit de volgende labels: "lang" (= /a/, /o/, /i/, /e/, /y/), "kort" (= /ɑ/, /ɔ/, /ɪ/, /ɛ/, /ʏ/), "sjwa" (= verdofte vocaal), "zero" (= gedeleerde vocaal) en hun respectieve tussenvormen ("lang/kort", "kort/sjwa", "sjwa/zero"). Alleen stimuli waarover de drie beoordelaars het eens waren, werden verder verwerkt. Voor meer details over de wijze van dataverwerking, zie Kloots (2005).

Bij de voorbereiding van het luisterexperiment bleek al snel dat er nog een extra label nodig was, eentje dat niet kon worden voorzien op basis van de bestaande literatuur. Woorden, afkomstig uit spontane spraak, die uit hun oorspronkelijke context worden geknipt, bleken soms namelijk totaal onherkenbaar te worden. Het spreekt vanzelf dat het bij de betreffende stimuli ook eenvoudig onmogelijk was om individuele klinkers uit het woord nauwkeurig te beoordelen. Deze stimuli kregen dan ook het label "onherkenbaar" en werden niet verwerkt in het onderzoek. Het Nederlandse materiaal bleek iets meer onverstaanbare stimuli te bevatten dan het Vlaamse. Van de Vlaamse stimuli was 4,9% onherkenbaar (122 op 2485), van het Nederlandse materiaal 10,9% (210 op 1927).

De belangrijkste vaststelling uit dit onderzoek was dat heel wat klinkers het label "kort" kregen. Vanuit fonologisch perspectief is dat erg opmerkelijk. Fonologen gaan er namelijk van uit dat aan het syllabeeinde nooit "korte" vocalen voorkomen (cf. *Minimal Rhyme Constraint* – zie bv. Booij 1995). Van de beklemtoonde klinkers werd 5,2% als "kort" geclassificeerd (68 op 1316), van de onbeklemtoonde vocalen 55,5% (881 op 1588).

5.2 Casestudy 2: onderzoek naar spreeksnelheid

Verhoeven e.a. (2004) gebruikten het lerarencorpus als uitgangspunt voor een onderzoek naar spreeksnelheid in Vlaanderen en Nederland. Deze studie vulde een belangrijke leemte. Naar spreeksnelheid was voor het Nederlands namelijk nog nauwelijks onderzoek gedaan. Over de spreeksnelheid in Vlaanderen was – voor zover we konden nagaan – zelfs nog helemaal niets bekend.

Helaas was slechts voor een deel van het lerarencorpus een geverifieerde (brede) fonetische transcriptie beschikbaar in het CGN. Basis

[9] Het ging om drie taalkundigen die ervaring hadden met de beoordeling van klankverschijnselen in het Nederlands. Hun dialectachtergrond is constant gehouden: ze zijn alle drie opgegroeid in de provincie Antwerpen.

voor het onderzoek vormden daarom ook hier *orthografische* transcripties: die waren namelijk wel beschikbaar voor alle 160 opnamen. Eerst werden de transcripties verdeeld in syllaben met behulp van een computerscript. Uitgangspunt daarbij was het fonologische principe dat de Nederlandse syllabe een vocaal als kern heeft, oftewel: 1 vocaal = 1 syllabe (zie bv. Booij 1995). Vervolgens werd het aantal syllaben per seconde geteld met behulp van een (ander) computerscript. We berekenden zowel *articulatiesnelheid* en *spreeksnelheid*. In het eerste geval wordt de duur van stiltes niet meegerekend bij de berekening van de totale spreektijd, in het tweede geval wel.

Ook in dit onderzoek werden we met enkele onverwachte praktische problemen geconfronteerd. Zo bleek de spontane spraak niet alleen stiltes, maar ook zgn. *filled pauses* te bevatten (bv. *uh, mmm*). In navolging van Laver (1995: 158) werden deze *filled pauses* steeds meegerekend bij de berekening van de totale spreektijd. Andere moeilijkheden hingen samen met de onvoorspelbare inhoud van spontane spraak. Bij de voorbereiding van een computationele analyse probeert de onderzoeker via een grondige (voor)studie van het materiaal zo veel mogelijk problemen te voorzien, maar achteraf blijkt datzelfde materiaal toch altijd nog een aantal verrassingen in petto te hebben. Zo traden er bijvoorbeeld moeilijkheden op bij de verwerking van acroniemen. De woorden *VDAB* en *AIDS* bestaan allebei uit vier grafemen, maar terwijl *VDAB* ook vier syllaben telt, bevat *AIDS* er maar één. Omdat het computerscript de acroniemen niet altijd correct syllabificeerde, bleek een manuele controle noodzakelijk. Een vergelijkbaar probleem trad op bij ambigue vocaalcombinaties, i.e. combinaties van twee vocalen die de ene keer afzonderlijk worden uitgesproken en de andere keer voor een enkele klank staan (bv. *reactie* vs. *league*). Ook hier was een manuele controle nodig. Ten slotte moest er ook extra aandacht worden besteed aan getallen. In de orthografische transcripties werden die soms in cijfers weergegeven. Een moedertaalspreker weet dat de gesproken variant van "12" één syllabe telt en die van "14" twee, een computerscript heeft op dit punt wat sturing nodig.

Het opmerkelijkste onderzoeksresultaat was het verschil tussen Vlaanderen en Nederland. De Nederlandse leraren bleken namelijk significant sneller te spreken dan hun Vlaamse collega's. Articulatiesnelheid en spreeksnelheid bedroegen respectievelijk 5,05 en 4,23 syllaben per seconde in Nederland, 4,23 en 4,00 syllaben per seconde in Vlaanderen.

6. Conclusie

Wie inzicht wil verwerven in de onvoorbereide, gesproken taal zoals die gebruikt wordt in c.q. voor de dagelijkse communicatie, bestudeert het best spontane spraak. Ook in de taaltechnologie bestaat er een groeiende belangstelling voor spontane spraak. Wie aan de slag gaat met dit type van gesproken taal moet echter bereid zijn om zich te laten verrassen door zowel spraak als corpus. Beide bevatten namelijk onvermijdelijk een aantal onvolkomenheden en leemten. Het eerste grote, publieke corpus dat zowel spontaan Standaardnederlands van Vlaamse als Nederlandse volwassenen bevat, is het *Corpus Gesproken Nederlands*. Een van de interessantste componenten van het CGN is het zgn. "lerarencorpus", een verzameling interviews met 80 Vlaamse en 80 Nederlandse leraren Nederlands. Wat dit subcorpus zo aantrekkelijk maakt, is dat bij de samenstelling van de steekproef systematisch rekening gehouden werd met de variabelen land, sekse, leeftijd en regio. De spontane spraak uit het "lerarencorpus" vormde al de basis voor een onderzoek naar reductieverschijnselen en een studie naar spreeksnelheid. Een combinatie van computationele technieken en voorwereldlijk geduld bleek uiteindelijk wel degelijk te leiden tot meer inzicht in "the most basic type of communicative use of language" (Rischel 1992: 380).

Bibliografie

Bauer, L. (2002): Inferring Variation and Change from Public Corpora. In: J. Chambers, P. Trudgill & N. Schilling-Estes [eds.], *The Handbook of Language Variation and Change*. Malden/Oxford, Blackwell Publishers, 97-114

Beckman, M. (1996): A Typology of Spontaneous Speech. In: Y. Sagisaka, N. Campbell & N. Higuchi [eds.], *Computing Prosody. Computational Models for Processing Spontaneous Speech*. New York e.a., Springer, 7-26

Biber, D., S. Conrad & R. Reppen (1998): *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge e.a., Cambridge University Press

Blancquaert, E. (1948): *Na meer dan 25 jaar Dialect-onderzoek op het Terrein*. Tongeren, Michiels

Booij, G. (1995): *The Phonology of Dutch*. Oxford, Clarendon Press

Cassier, L. & P. Van de Craen (1986): Vijftig jaar evolutie van het Nederlands. In: J. Creten, G. Geerts & K. Jaspaert [red.], *Werk-in-uitvoering. Momentopnamen van de sociolinguïstiek in België en Nederland*. Leuven/Amersfoort, Acco, 59-73

Catania, J., D. Binson, J. Canchola, L. Pollack, W. Hauck & T. Coates (1996): Effects of interviewer gender, interviewer choice, and item

wording on responses to questions concerning sexual behaviour. In: *Public Opinion Quarterly*, 60, 345-375

COUSSÉ, E. & S. GILLIS (2006): Regional Bias in the Broad Phonetic Transcriptions of the Spoken Dutch Corpus. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Parijs, ELDA, 2080-2083

CRYSTAL, D. (2003): *A Dictionary of Linguistics and Phonetics*. Malden e.a., Blackwell Publishing, 5th ed.

DE JONG, E. [red.] (1979): *Spreektaal. Woordfrequenties in gesproken Nederlands*. Utrecht, Bohn, Scheltema & Holkema

DE VOOYS, C. (1970): *Geschiedenis van de Nederlandse taal*. Groningen, Wolters-Noordhoff

ERNESTUS, M. (2000): *Voice Assimilation and Segment Reduction in Casual Dutch. A Corpus-Based Study of the Phonology-Phonetics Interface*. Proefschrift Vrije Universiteit Amsterdam. Utrecht, Landelijke Onderzoeksschool Taalkunde

FAGYAL, Z. (1995): *Aspects phonostylistiques de la parole mediatisée lue et spontanée. Age, prestige, situation, style et rythme de parole de l'écrivain M. Duras*. Proefschrift Université de la Sorbonne Nouvelle Paris III

FASOLD, R. (2001): *The Sociolinguistics of Language*. Oxford, Blackwell

FLORIS, R. (1997): 'Een dialect moet je horen'. In: H. van de Wijngaard & R. Belemans [red.], *Nooit verloren werk. Terugblik op de Reeks Nederlandse Dialectatlassen (1925-1982)*. Groesbeek, Stichting Nederlandse Dialecten, 87-90

GOEDERTIER, W. & S. GODDIJN (2000): *Protocol voor orthografische transcriptie*. Te vinden via de oorspronkelijke CGN-website <http://lands.let.kun.nl/cgn/doc_Dutch/topics/version_1.0/annot/orthography/info.htm> en de website van de Centrale voor Taal- en Spraaktechnologie <http://ww2.tst.inl.nl/index.php?option=com_content&task=view&id=257&Itemid=265>

GROOTAERS, L. (1926): Geschiedenis van het Zuid-Nederlandsch dialectonderzoek. In: L. Grootaers & G. Kloeke, *Handleiding bij het Noord- en Zuid-Nederlandsch dialectonderzoek*. 's-Gravenhage, Martinus Nijhoff, 27-56

HANSEN, J. & W. SCHULDT (1982): Physical distance, sex and intimacy in self-disclosure. In: *Psychological Reports*, 51, 3-6

HEIKENS, H. (1978): Het sociolinguïstisch opgebouwd corpus Amsterdamse spreektaal. In: *Taal en Tongval*, 30, 36-49

KLOOTS, H. (2005): *Vocaalreductie in het Standaardnederlands in Vlaanderen en Nederland*. Proefschrift Universiteit Antwerpen

KOHLER, K. (2000): Investigating Unscripted Speech: Implications for Phonetics and Phonology. In: *Phonetica*, 57, 85-94

LAAN, G. (1997): The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. In: *Speech Communication*, 22, 43-65

LABOV, W. (1972): *Sociolinguistic Patterns.* Philadelphia, University of Pennsylvania Press

LAVER, J. (1995): *Principles of phonetics.* Cambridge e.a., Cambridge University Press, herdr.

MILROY, L. & M. GORDON (2003): *Sociolinguistics. Method and Interpretation.* Malden/Oxford, Blackwell Publishing

OOSTDIJK, N. (2000): Het Corpus Gesproken Nederlands. In: *Nederlandse Taalkunde*, 5, 280-284

PIEPENBROCK, R. (1999): *Nederlandse gesproken corpora: een inventarisatie.* Intern CGN-document, te vinden via <http://lands.let.ru.nl/cgn/publs/1999_01.pdf>

RISCHEL, J. (1992): Formal linguistics and real speech. In: *Speech Communication*, 11, 379-392

SAMPSON, G. (2001): *Empirical Linguistics.* London/New York, Continuum

SCHIEL, F. (2005): *CGN 1.0 – Validation Report.* München, Universität München, Institut für Phonetik, BAS Services. Beschikbaar via de website van de *Centrale voor Taal- en Spraaktechnologie* <http://www.tst.inl.nl>.

SMAKMAN, D. (2006): *Standard Dutch in the Netherlands. A Sociolinguistic and Phonetic Description.* Proefschrift Radboud Universiteit Nijmegen. Utrecht, Landelijke Onderzoeksschool Taalkunde

SMAKMAN, D. & R. VAN BEZOOIJEN (1997): Een verkenning van populaire ideeën over de standaardtaal in Nederland. In: R. van Bezooijen, J. Stroop & J. Taeldeman [red.], *Standaardisering in Noord en Zuid* [= *Taal & Tongval*, themanummer 10], 126-139

STRIK, H. (2001): 'Dat heb ik helemaal niet gezegd!' De prestaties van de spraakherkenner. In: *Onze Taal*, 70, 284-286

VANACKER, V. (1963): *Syntaxis van gesproken taal te Aalst en in het Land van Aalst in de XVde, de XVIde en de XVIIde eeuw.* s.l., Belgisch Interuniversitair Centrum voor Neerlandistiek

VANACKER, V. & G. DE SCHUTTER (1967): Zuidnederlandse dialekten op de band. In: *Taal & Tongval*, 19, 35-51

VAN BERGEM, D. (1995): *Acoustic and Lexical Vowel Reduction.* Proefschrift Universiteit van Amsterdam. Amsterdam, IFOTT. Dordrecht, ICG-Printing

VAN BREE, C. (1996): *Historische taalkunde.* Leuven/Amersfoort, Acco, 2de, herz. dr.

VAN DE VELDE, H. (1996): *Variatie en verandering in het gesproken Standaard-Nederlands (1935-1993).* Proefschrift Katholieke Universiteit Nijmegen

VAN DE VELDE, H. & M. HOUTERMANS (1999): Vlamingen en Nederlanders over de uitspraak van nieuwslezers. In: E. Huls & B. Weltens [red.], *Artikelen van de Derde Sociolinguïstische Conferentie.* Delft, Eburon, 451-462

VAN HAERINGEN, C. (1971): Amsterdams van Multatuli. In: *De nieuwe taalgids*, 65, 370-376

VAN HOUT, R., G. DE SCHUTTER, E. DE CROM, W. HUINCK, H. KLOOTS & H. VAN DE VELDE (1999): De uitspraak van het Standaard-Nederlands:

variatie en varianten in Vlaanderen en Nederland. In: E. Huls & E. Weltens [red.], *Artikelen van de Derde Sociolinguïstische Conferentie*. Delft, Eburon, 183-196.

VAN KEYMEULEN J. [red.] (2002): *Taalkamer*. Themanummer van *Oost-Vlaamse Zanten*, 77, nr. 3-4 (de volledige Taalkamer van het Huis van Alijn in Gent is beschikbaar op cd-rom en op <http://www.huisvanalijn.be/taalkamer/index.html>)

VAN OSS, F. & C. GUSSENHOVEN (1984): De Nederlandse slot-n in het nieuws. In: *Gramma*, 8, 37-45

VERHOEVEN, J., G. DE PAUW & H. KLOOTS (2004): Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. In: *Language and Speech*, 47, 299-310

VOSKUIL, J. (1997): *Het Bureau I: Meneer Beerta*. Amsterdam, G. van Oorschot, 8ste dr.

WEIJNEN, A. (1966): *Nederlandse dialectkunde*. Assen, Van Gorcum & comp / H. Prakke & H. Prakke, 2de dr.

WODAK, R. (1982): Erhebung von Sprachdaten in natürlicher oder simuliert-natürlicher Sprechsituation. In: W. Besch, U. Knoop, W. Putschke & H. Wiegand [Hrsg.], *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung*. Erster Halbband. Berlin/New York, Walter de Gruyter, 539-544

ZWAARDEMAKER, H. & L. EIJKMAN (1928): *Leerboek der phonetiek*. Haarlem, De Erven F. Bohn