Het Nederlands in taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen

Walter Daelemans & Helmer Strik (red.)

Een rapport in opdracht van de Nederlandse Taalunie

01/07/2002

Projectuitvoerders:

Diana Binnenpoorte
Masja Kempen
Janienke Sturm
Folkert de Vriend

Deskundigen:

Gosse Bouma
Dirk Van Compernolle
Walter Daelemans
Arthur Dirksen
Frank Van Eynde
Dirk Heylen
Franciska de Jong
Jean-Pierre Martens
Anton Nijholt
Helmer Strik
Raymond Veldhuis

Inhoudsopgave

Deel I

Werkwijze en aanbevelingen

1 Inleiding

Om de status van het Nederlands te vrijwaren in een snel evoluerende multilinguale informatiemaatschappij moeten ICT toepassingen snel en adequaat in een Nederlandstalige versie ontwikkeld worden. Een voorwaarde hiervoor is de beschikbaarheid van een digitale taalinfrastructuur, bestaande uit kwalitatief hoogstaande data (lexica, corpora) en softwaremodules die kunnen worden gebruikt in de ontwikkeling van applicaties[1]. Op dit moment zijn er ernstige lacunes in de beschikbare basismaterialen voor het Nederlands [23].

In dit document wordt een prioriteitenlijst voorgesteld voor de ontwikkeling van dergelijke basismaterialen voor de Taal- en Spraaktechnologie (TST) voor het Nederlands. Deze prioritering is gebaseerd op (i) de definitie van een basis TST-voorziening voor het Nederlands (BATAVO) en (ii) een inventarisatie en evaluatie van de bestaande en op korte termijn te verwachten basisvoorzieningen.

Opdrachtgever van dit project was het Nederlands-Vlaams Platform voor het Nederlands in Taal- en Spraaktechnologie. Het project werd uitgevoerd in het kader van de actielijnen B en C uit het Actieplan voor het Nederlands in Taal- en Spraaktechnologie. Naast deze twee actielijnen bevat het actieplan de actielijnen A en D. Actielijn A wil een betere samenwerking tussen de diverse partijen (industrie, academia en beleidsinstanties) stimuleren en de zichtbaarheid van het veld verbeteren door publiciteit te geven aan de resultaten van onderzoek in taal- en spraaktechnologie. Actielijn D wil een blauwdruk [107] definiëren voor management, onderhoud en distributie van door de overheid gefinancierde digitale materialen. Het project waarvan we in dit document verslag doen zal verder het platform-BC project genoemd worden. Om te beginnen zullen we een korte chronologische beschrijving van het verloop van dit project geven.

Vooraf was besloten dat er een stuurgroep moest komen van acht deskundigen op het gebied van taal- en spraaktechnologie, uit Nederland (NL) en Vlaanderen (VL). Begin 2000 werd deze stuurgroep samengesteld. Voor taaltechnologie waren dat (in alfabetische volgorde): Gosse Bouma (NL), Walter Daelemans (VL), Frank van Eynde (VL) en Anton Nijholt (NL). Voor spraaktechnologie waren de deskundigen (in alfabetische volgorde): Dirk van Compernolle (VL), Jean-Pierre Martens (VL), Helmer Strik (NL) en Raymond Veldhuis (NL). Anton Nijholt werd regelmatig vervangen door Dirk Heylen of Franciska de Jong. In het laatste jaar van het project werd Raymond Veldhuis vervangen door Arthur Dirksen. Ook vanuit de beleidsinstanties waren een aantal mensen bij het platform-BC project betrokken. In het begin waren dit Jeannine Beeken (NTU), Elisabeth D'Halleweyn (NTU) en Alice Dijkstra (NWO). Later werden Jeannine Beeken en Elisabeth D'Halleweyn vervangen door Catia Cucchiarini (NTU) en Lisanne Teunissen (NTU).

In 2000 is de stuurgroep een aantal keren bij elkaar geweest op verschillende plaatsen, om te discussiëren over de manier waarop de opdracht het beste uitgevoerd kon worden. De coördinatie voor het taalgedeelte werd verzorgd door Walter Daelemans en de coördinatie van het spraakgedeelte door Helmer Strik. Gekozen werd voor de volgende werkwijze, die in essentie bestaat uit het maken van drie lijsten:

1. BATAVO: Een Basis-Taal&Spraak-Voorziening (BATAVO) die definieert welke middelen (data en modules) aanwezig moeten zijn in een basisinfrastructuur voor taal- en spraaktechnologie, en wat de noodzaak van aanwezigheid is (prioritering van basisvoorzieningen).

[1] Actieplan voor het Nederlands in Taal- en Spraaktechnologie, NTU, 2000.

2. Inventarisatie & Evaluatie: Een overzicht van de middelen (data en modules) die aanwezig en van voldoende kwaliteit zijn.

3. Prioriteitenlijst: Op basis van 1. BATAVO en 2. Inventarisatielijst is het dan mogelijk om een lijst te maken van de middelen die ontwikkeld moeten worden en hierin een prioritering aan te brengen.

Een eerste versie van de drie bovengenoemde lijsten werd gemaakt tijdens de bijeenkomsten van het platform-BC project in 2000, op basis van de expertise van de aanwezigen. Vernieuwde en verbeterde versies van deze lijsten zijn daarna gemaakt op basis van de resultaten van een veldonderzoek. Dit veldonderzoek vond plaats in 2001 en werd uitgevoerd door Masja Kempen en Folkert de Vriend voor taaltechnologie, onder begeleiding van Walter Daelemans en Gosse Bouma, en door Diana Binnenpoorte en Janienke Sturm voor spraaktechnologie, onder begeleiding van Helmer Strik. Het veldonderzoek zelf bestond uit twee fases:

1. Door te zoeken in de literatuur en op internet, en door het raadplegen van experts (zowel uit de stuurgroep als andere experts) werd een tweede versie gemaakt van de bovengenoemde drie lijsten. De resultaten werden beschreven in een eerste versie van het rapport [52].

2. Een belangrijk aspect van het platform-BC project is, dat een zo groot mogelijke consensus bereikt moest worden binnen het taal- en spraaktechnologieveld. Daarom werd in de tweede fase van het veldonderzoek geprobeerd om het complete veld zo goed mogelijk te informeren en de gelegenheid te geven om te reageren. Daartoe werd een brief gestuurd naar zo'n 2000 actoren in het taal- en spraaktechnologieveld. Hiervoor werd gebruik gemaakt van de lijst die samengesteld was binnen actielijn A van het eerdergenoemde platform voor het Nederlands in Taal- en Spraaktechnologie. In deze brief stond o.a. de (voorlopige) prioriteitenlijst, een link naar de voorlopige (eerste) versie van het rapport: `http://www.taalunieversum.org/tst/actieplan/batavo-v1.pdf`, en de mededeling dat er op 15 november 2001 een bijeenkomst zou zijn waarop over de voorlopige (eerste) versie van het rapport gediscussieerd kon worden. In de brief werd ook expliciet gevraagd om te reageren, door aanwezig te zijn op de bijeenkomst op 15 november, of per post of e-mail. Op de bijeenkomst van 15 november waren ongeveer 100 deelnemers aanwezig. De reacties zijn verwerkt in de finale versie van het rapport, dwz. dit document.

In Sectie 2 geven we een beschrijving van de toepassingen, modules en data die we onderscheiden binnen de TST. Sectie 3 geeft een inschatting van de beschikbaarheid en het belang van de verschillende modules en data voor verschillende toepassingscategorieën. Deze inschatting is gebaseerd op een inventarisatie en evaluatie die u terugvindt in Deel II van dit rapport. Dat deel geeft een overzicht van de stand van zaken op het gebied van software modules en dataverzamelingen voor het Nederlands. Sectie 4 vertaalt deze informatie naar een eerste voorstel voor een BATAVO voor het Nederlands en een bijhorende prioriteitenlijst. Sectie 6 beschrijft de algemene criteria die we hanteren voor de evaluatie van alle hulpmiddelen. Sectie 5 tenslotte formuleert aanbevelingen en prioriteringen voor de te ontwikkelen basisinfrastructuur voor de TST voor het Nederlands.

In onze prioritering geven we de voorkeur aan data of modules die (i) direct of indirect belangrijk zijn voor relatief veel toepassingen, (ii) op dit moment niet of onvoldoende

beschikbaar zijn, en (iii) waarvan de ontwikkeling haalbaar is met de huidige stand van zaken in wetenschap en technologie.

Een belangrijke categorie van hulpmiddelen die in deze studie niet verder aan bod komt zijn de taalonafhankelijke hulpmiddelen. Bijv. omgevingen voor annotatie van corpora en voor opbouw van lexicale databanken, grammatica-ontwikkelomgevingen, tools voor statistische modellering of machine learning, ... De motivatie hiervoor is dat deze tools nuttig zijn voor TST in alle talen en dus niet het onderwerp hoeven uit te maken van een initiatief voor de opbouw van basisvoorzieningen specifiek voor het Nederlands. Dat neemt niet weg dat deze tools en recente methodes voor gerichte selectie van data om te annoteren (bijv. *active learning*) een belangrijke rol zullen spelen in de TST als taalonafhankelijke basisinfrastructuur.

2 Toepassingen, modules en data

Eindgebruikers komen alleen in aanraking met complete toepassingen van TST zoals automatische vertaling op internet, dicteerpakketten, spraakinterfaces naar informatiesystemen via de telefoon, etc. Dit zijn complexe systemen die gebruik maken van heel wat verschillende componenten (vaak lingware genoemd) als lemmatisering, woordsoortdisambiguering, enz., en van dataverzamelingen (resources) als lexica en corpora. Verschillende van deze 'halffabrikaten' (op zichzelf hebben ze beperkt commercieel belang) spelen een meer of minder belangrijke rol in de ontwikkeling van verschillende toepassingen. In een gezonde TST-infrastructuur zouden de belangrijkste van deze halffabrikaten beschikbaar moeten zijn zonder beperkingen zodat zowel de taalindustrie als het onderzoek ze als uitgangspunt kan nemen –de eerste voor de ontwikkeling van toepassingen, de tweede voor verdere research en ontwikkeling– zonder ze zelf steeds opnieuw te moeten ontwikkelen. De achtergrond van deze studie is onze overtuiging dat in het Nederlands taalgebied, in vergelijking met de ons omringende taalgebieden, een sterke inhaalbeweging nodig is in de ontwikkeling en het beschikbaar stellen van deze halffabrikaten.

We zullen hier kort een beschrijving geven van de verschillende toepassingsgebieden en halffabrikaten (modules en dataverzamelingen) die we onderscheiden in dit rapport. Voor een meer diepgaande inleiding tot het domein van de TST verwijzen we naar [75]. Het is onmogelijk om een indeling te maken waarover een volledige consensus mogelijk is, of die het probleem vanuit alle kanten adequaat belicht, maar wat volgt volstaat voor de doelstellingen van dit rapport.

2.1 Toepassingen

Gemakshalve maken we hier een onderscheid tussen spraaktechnologie, waarbij gesproken taal de invoer of uitvoer van een toepassing is, en taaltechnologie waarbij dat tekst is. We onderscheiden acht belangrijke categorieën van toepassingen. In realiteit is de situatie complexer omdat tekst en spraak en verschillende types van toepassingen gecombineerd kunnen worden (bijv. e-mail voorlezers, multimodale informatie-extractie, meertalige spraakindexering etc.).

1. *CALL* (Computer Assisted Language Learning) systemen worden al langer gebruikt voor training van lees-, schrijf- en luistervaardigheid. De laatste tijd is er een toenemende belangstelling van het gebruik van CALL systemen voor het trainen van spreekvaardigheid. Voor dit laatste is zeer specialistische spraaktechnologie nodig die nog grotendeels ontwikkeld moet worden.

2. Met *toegangscontrole* als toepassing van taal- en spraaktechnologie wordt bedoeld dat met behulp van fysieke kenmerken, in dit geval een spraakgeluid, toegang kan worden verschaft tot systemen of gebouwen, of personen geverifieerd of geïdentificeerd kunnen worden.

3. Systemen waarbij *spraakinvoer* wordt geanalyseerd (spraakherkenning) produceren als uitvoer tekst (spraak-naar-tekst), systeemcommando's (command & control), of domeinconcepten (spraak-naar-concept). Bij spraak-naar-tekst behoren dicteertoepassingen voor specifieke beroepsgroepen als radiologen, advocaten, enz., maar ook voor thuisgebruik door particulieren, en automatische transcriptie (bijv. voor indexering of voor samenvatting).

4. Systemen voor *spraakuitvoer* worden gebruikt in toepassingen waarbij tekst moet worden voorgelezen (gesproken e-mail, voorleesmachine voor blinden, uitspraakwoordenboeken, ...), of waarbij concepten of data moeten worden omgezet naar spraak (concept-naar-spraak).

5. *Dialoogsystemen* vormen een natuurlijke taal interface tot databanken, expertsystemen en informatiesystemen (bijv. reisinformatiesystemen). Hoewel de interactie kan gebeuren in geschreven taal, zijn deze systemen meestal gebaseerd op spraak in- en uitvoer. Ook toepassingen van Virtuele Realiteit waarin spraakinteractie een rol speelt horen hier bij.

6. Onder *documentproductie* groeperen we een reeks toepassingen die te maken hebben met de productie van tekst, van spelling- en grammatica-correctie, woordafbreking en style checking tot tekstgeneratie.

7. Toepassingen waarin tekst- en spraakanalyse een rol speelt bij het lokaliseren van informatie en het extraheren van kennis uit tekst en spraak noemen we hier *informatie-extractie*. Hiertoe behoren de natuurlijke taal verwerkingsaspecten van Information Retrieval, multilinguale en multimediale information retrieval en text mining. Text mining omvat toepassingen voor de classificatie van documenten (bijv. om ze te kunnen beheren, 'routen' of filteren), de automatische extractie van informatievelden (om grote hoeveelheden tekstdata om te kunnen zetten in een vast databaseformaat), *question answering*, en automatische samenvatting.

8. *Multilinguale toepassingen* omvatten naast systemen voor automatische vertaling (domeinspecifiek of domeinbreed) ook specifieke vertaalhulpmiddelen (woordenboeken, grammar checkers), en technieken voor de ontwikkeling van taalkundig verrijkte vertaalgeheugens.

2.2 Halffabrikaten

Lingware modules maken gebruik van een model dat een invoerrepresentatie afbeeldt op een uitvoerrepresentatie (bijv. van tekst naar spraak, of van een woord naar een reeks morfemen). Het domein van deze afbeeldingen is een woord, een zin, een tekst, een spraakfragment of een interne representatie. Op dit moment zijn twee methodes (in toenemende mate in combinatie) gangbaar in de TST voor de constructie van lingware modules. Zo'n module kan met de hand worden gemaakt (gebruik makend van taalkundig inzicht), maar kan ook automatisch afgeleid

worden door middel van inductie op basis van grote hoeveelheden data, al dan niet voorzien van taalkundige verrijking. In het eerste geval spreken we van een deductieve of kennisgebaseerde benadering, in het tweede geval van een inductieve of statistische benadering. Beide benaderingen vullen elkaar aan: een kennisgebaseerde aanpak laat de incorporatie van een taalkundig gesofistikeerde analyse in het computationele model toe, gebaseerd op een lange traditie van taalkundig onderzoek, waar de inductieve benadering meestal niet toelaat deze kennis te verwerken. Aan de andere kant laat een corpusgebaseerde inductieve benadering toe modellen af te leiden uit data met een grotere dekking van taalfenomenen, een hogere voorspellende accuraatheid, en meer robuustheid ten opzichte van 'werkelijk' (soms ongrammaticaal) taalgebruik dan deductieve benaderingen. Zeker in de spraaktechnologie, maar de laatste tijd ook in de taaltechnologie domineert de statistische benadering. We hechten in onze indeling dan ook voldoende belang aan de dataverzamelingen die nodig zijn om modules te 'trainen' volgens deze aanpak, waarvoor nog steeds geldt dat meer (data) beter is.

2.2.1 Modules

Taalmodules:

1. *Grafeem-foneemomzetting.* Deze module staat tussen spraak- en taaltechnologie. Gegeven de spelling van een (nieuw) woord moet een acceptabele uitspraakrepresentatie van dat woord opgeleverd worden, eventueel met voorspelling van woordgrenseffecten (assimilatie tussen het woord en de woorden in de onmiddellijke context). Twee deelproblemen hierbij zijn de toekenning van woordklemtoon en de opsplitsing van een woord in syllaben (in benaderingen die gebruik maken van syllabenstructuur).

2. *Spraak- en tekstvoorverwerking.* Lingware modules gaan er van uit dat het domein van de afbeelding die ze implementeren beschikbaar is (bijv. dat een inputzin of een inputwoord beschikbaar is). Dit is op zichzelf echter geen triviaal probleem (bijv. disambiguering tussen punt van een afkorting en aan einde van een zin, interpunctie bij titels, ondertitels etc. Speciale lingware kan worden ontwikkeld die zinsgrenzen detecteert in tekst (uitingen segmenteert in spraak), en voorkomens van woorden (tokens) detecteert. Een analysestap verder is naamherkenning (named entity recognition). Voorkomens van speciale types van namen (eigennamen, bedrijfsnamen, locaties, datums, tijdstippen, e.d.m.) kunnen vooraf gedetecteerd en gelabeld worden.

3. *Spellingcontrole.* In alle toepassingen is de correcte herkenning van woordvormen noodzakelijk (voor lexicale toegang of verdere verwerking). Om te verhinderen dat elke spelfout wordt gedetecteerd als een nieuw, onbekend, woord is spellingcontrole een nuttige aparte component, eventueel in combinatie met tokenherkenning.

4. *Morfologie.* In verschillende toepassingen moet voor een onbekend woord (bijv. een nieuwe samenstelling) toch een betrouwbare analyse kunnen worden gegeven. In *lemmatisering* worden het lemma (stam) en de woordsoort achterhaald van een verbogen, vervoegd of samengesteld woord. *Morfologische analyse* gaat een stap verder en bepaalt eveneens de interne structuur van complexe woorden. In beide gevallen moet rekening worden gehouden met spellingveranderingen. Bij *morfologische synthese* worden alle verbogen of vervoegde vormen berekend van een woordstam.

5. *Woordsoortdisambiguering* (part of speech tagging). Een eerste stap in de disambiguering van taal is de toekenning van de woordsoort die een woord heeft in de context waarin het voorkomt (morfo-syntactische disambiguering). Afhankelijk van de toepassing kan een eenvoudige (tiental hoofdwoordsoorten), een intermediaire (enkele tientallen 'tags'), of een complexe tag set (tot enkele honderden tags) nodig zijn.

6. *Syntactische analyse.* Om de syntactische structuur van zinnen te achterhalen wordt traditioneel grammaticagebaseerd parsing gebruikt (CFGs of op basis van constraints). Grammaticaregels kunnen ook probabilistisch (PCFG) of contextgevoelig worden gemaakt (DOP). Een alternatief voor volledige syntactische analyse is *shallow parsing*. Voor verschillende toepassingen is het niet noodzakelijk om een volledige syntactische analyse beschikbaar te hebben. Herkenning van de (hoofden van de) belangrijkste nominale constituenten (via een proces dat *chunking* wordt genoemd) en hun onderlinge relaties (*relation finding*) is vaak al voldoende.

7. *Semantische en pragmatische analyse.* Semantische analyse moet in staat zijn van alle zinnen in een tekst (in een specifiek domein) een letterlijke betekenis op te leveren, bijv. in een variant van eerste-orde predikatenlogica. Op het gebied van meer robuuste semantische analyse wordt vooral gewerkt aan *woordbetekenisdisambiguering*, een proces waarbij de contextueel juiste betekenis van een woord wordt gezocht. Gegeven de letterlijke betekenis(sen) van zinnen moet pragmatische analyse deze verder disambigueren door gebruik te maken van context (zowel linguïstische als niet-linguïstische). Voor een deel van deze taak, *referentresolutie* (het in verband brengen van tekstitems die naar hetzelfde concept verwijzen), wordt ook vaak aparte software ontwikkeld.

8. *Tekstgeneratie.* In modules voor tekstgeneratie moet op basis van een interne representatie (die ook niet-verbaal kan zijn, bijv. tabellen) tekst gegenereerd worden waarvan de betekenis correspondeert met die van de interne representatie. Meestal wordt voor specifieke toepassingen templaatgebaseerde generatie gebruikt, maar ook diepe generatie (gebaseerd op tekst- en zinsgrammatica's) wordt onderzocht.

9. *Taalpaarafhankelijke vertaalmodules.* Net zoals volledige spraakanalyse en spraaksynthesecomponenten een nuttige module kunnen zijn in allerlei toepassingen, is dat ook het geval voor volledige vertaalmodules (die uiteraard zelf uit een groot aantal modules kunnen bestaan).

Spraakmodules:

1. *Prosodiegeneratie en -herkenning.* Bij spraakherkenning en gespreksanalyse wil men uit het akoestisch signaal prosodische elementen halen die kunnen bijdragen tot een betere herkenning, of tot een betere identificatie van de sprekerintentie. De extractie van die prosodische elementen noemt men prosodieherkenning. *Prosodiegeneratie* is voor spraaksynthese belangrijk om correcte intonatiepatronen te kunnen berekenen (pauzes en zinsaccenten).

2. *Volledige spraaksynthese en spraakherkenning.* Hoewel spraaksystemen op zich uit verschillende componenten bestaan, is het soms nuttig ze eveneens te beschouwen als potentiële componenten in TST-toepassingen. De beschikbaarheid van open en modulaire spraaksynthese- en analysesystemen zou de ontwikkeling van bijv. natuurlijke taal interfaces aanzienlijk vooruithelpen.

3. *Synthesemodules: allofoonsynthese, difoonsynthese, unit selection.* Spraaksynthese is het omzetten van geschreven tekst in gesproken taal. Traditioneel werd hier vooral allofoonsynthese voor gebruikt, maar langzamerhand is de nadruk verschoven naar concatenatieve synthese. Hierbij worden segmenten achter elkaar geplakt die geknipt zijn uit opgenomen spraak. Voor het selecteren van deze segmenten is unitselectie nodig.

4. *Spraakherkenningsmodules: akoestische modellen, taalmodellen, uitspraaklexicon.* Automatische spraakherkenning is het automatisch omzetten van menselijke spraak in woorden. Er zijn veel verschillende spraakherkenningsystemen, maar de meeste systemen hebben dezelfde basisarchitectuur waarin drie componenten kunnen worden onderscheiden: akoestische decodering, lexicon en taalmodellering.

5. *Foonstringbewerkingen: transcriptie, segmentering, oplijning, afstandberekening.* Foonstrings zijn reeksen van fonen, en fonen is de verzamelnaam van de binnen spraaktechnologie vaak gebruikte basiseenheden die gerelateerd zijn aan fonemen, allofonen, etc. De foonstringbewerkingen die besproken worden zijn transcriptie, segmentering, het oplijnen van twee foonstrings, en het berekenen van de afstand tussen twee foonstrings.

6. *Robuuste spraakherkenning.* Automatische spraakherkenning is het automatisch omzetten van spraak in woorden. Behalve de doelspraak (die herkend moet worden) zullen er meestal andere geluiden aanwezig zijn (ruis, spraak van andere mensen, radio of TV, etc.). Deze andere geluiden zullen leiden tot fouten in de herkenresultaten. Met zgn. 'robuuste technieken' wordt geprobeerd om de fouten veroorzaakt door achtergrondgeluiden te reduceren.

7. *Non-native spraakherkenning.* Sommige spraakherkenningsapplicaties zullen (ook) door non-natives gebruikt worden. Dit is bijv. het geval bij CALL systemen die vrijwel alleen door non-natives gebruikt worden. Als spraakherkenners die getraind zijn voor natives gebruikt worden om non-natives te herkennen, is het aantal gemaakte herkenfouten veel te groot. Er zullen speciale systemen en technieken ontwikkeld moeten worden om dit aantal fouten sterk te reduceren.

8. *Sprekerherkenning: verificatie, identificatie, tracking.* Met sprekerherkenning wordt bedoeld het automatisch herkennen van wie er aan het woord is op basis van specifieke individuele kenmerken in het spraakgeluid. Sprekerherkenning kan worden opgedeeld in: sprekeridentificatie, sprekerverificatie en spreker-tracking. Het doel van sprekerverificatie is vast te stellen of de spreker daadwerkelijk degene is die hij claimt te zijn. Bij sprekeridentificatie is de taak de identiteit van de spreker uit een eindige set sprekers vast te stellen. Spreker-tracking betekent vaststellen welke spreker wanneer aan het woord is.

9. *Taal- en dialectidentificatie.* Taal- en dialectidentificatie is het proces dat gebruik maakt van specifieke kenmerken in het geluidssignaal om vast te stellen welke taal, dialect of accent wordt gesproken (kortweg gesproken taalidentificatie). Er zijn twee soorten toepassingen denkbaar voor taalidentificatiesystemen: 1) als voorbewerking voor automatische systemen, en 2) als voorbewerking voor menselijke operators.

10. *Adaptatie.* Adaptatie wordt in de spraaktechnologie toegepast om het verschil tussen train- en testcondities van een spraakherkenner te verkleinen. Twee soorten adaptatie zijn sprekeradaptatie en lexiconadaptatie. Sprekeradaptatietechnieken proberen

de sprekeronafhankelijke spraakherkenner af te stemmen op de karakteristieken van de betreffende spreker, om op deze manier de performantie van het systeem te verbeteren. In gevallen waar de uitspraak zeer sterk afwijkt van de uitspraken in de train set, kan sprekermodellering door middel van het lexicon uitkomst bieden, lexiconadaptatie. De term lexiconadaptatie wordt ook gebruikt voor het toevoegen van woorden aan het lexicon om het aantal OOV (out of vocabulary) woorden te verkleinen.

11. *Betrouwbaarheidsmaten en uitingverificatie.* Bij uitingverificatie wordt ernaar gestreefd om onjuist herkende (deel)uitingen vroegtijdig te verwerpen. De beslissing om een oplossing te verwerpen is veelal gebaseerd op de toetsing van een kansscore aan een vooraf gestelde drempelwaarde. Betrouwbaarheidsmaten zijn schattingen van de waarschijnlijkheid dat een oplossing correct is.

2.2.2 Data

Dataverzamelingen op zich zijn zelden noodzakelijk voor de ontwikkeling van concrete toepassingen. Ze spelen echter een essentiële rol in de ontwikkeling van de verschillende modules die noodzakelijk zijn voor toepassingen.

1. *Lexica.* Computationele lexica bevatten informatie over de woordvoorraad van een taal op verschillende taalkundige niveaus (van fonetiek tot pragmatiek). Ze kunnen lemma's bevatten, of woordvormen, of *multiword items* (frases of zelfs collocaties). In tegenstelling tot monolinguale lexica bevatten multilinguale lexica ook vertaalverbanden tussen de woorden van twee of meer talen.

2. *Thesauri.* Thesauri of *wordnets* bevatten woorden met hun betekenisrelaties, meestal in een hiërarchische of een netwerkstructuur. Thesauri voor specifieke domeinen worden *ontologieën* genoemd.

3. *Corpora.* De beschikbaarheid van grote verzamelingen tekst of spraakopnamen zijn essentieel geworden voor de ontwikkeling van taal- en spraaktechnologie. Deze corpora worden liefst gebalanceerd geconstrueerd (met alle relevante types van taalgebruik). *Geannoteerde corpora* zijn verrijkt met taalkundige beschrijvingen (bijv. fonetische transcriptie, morfologische structuur en woordsoort van woorden, syntactische structuur van zinnen, ...). *Niet-geannoteerde corpora* bevatten uitsluitend spraakopnamen en een orthografische transcriptie (in het geval van spraakcorpora), of uitsluitend tekst (in het geval van tekstcorpora). *Multilinguale* corpora bevatten spraak of tekst en bijbehorende transcriptie van meerdere talen. In *multimodale* corpora zijn verschillende (communicatie-)modi tussen mens/machine opgenomen, naast bijvoorbeeld spraak en tekst ook muisklikken. *Multimediale* corpora representeren de media waarin de verschillende modi kunnen voorkomen, zoals audio, video, plaatjes.

4. *Test suites en testcorpora.* Evaluatie en 'benchmarking' van spraak- en taaltechnologie toepassingen, modules en data is van uitzonderlijk belang voor de vooruitgang van de technologie. Allerlei testcorpora en verzamelingen voorbeelden (test suites) kunnen worden ontwikkeld om systemen en hulpmiddelen op een objectieve manier te vergelijken en evalueren.

Nu we een indeling hebben van de TST realiteit in drie dimensies (toepassingen, modules en data) proberen we in Sectie 3 een inschatting te maken van beschikbaarheid en belang van modules en data voor de ontwikkeling van commerciële toepassingen in TST. Dit zal leiden in Sectie 4 tot een inschatting van de noodzakelijke basisvoorzieningen voor de Nederlandstalige TST, en een intitiële prioritering, gebaseerd op de inventarisatie en evaluatie in Deel II.

3 Beschikbaarheid en belang van data en modules

We onderzoeken respectievelijk de beschikbaarheid van verschillende halffabrikaten (tabellen 1 en 2), het belang van dataverzamelingen voor modules (tabel 3), en het belang van modules voor toepassingen (tabel 4).

3.1 Beschikbaarheid van modules en data

In tabellen 1 en 2 staan de verschillende modules en dataverzamelingen met een inschatting van de beschikbaarheid ervan met een minimale kwaliteit voor commerciële toepassingsontwikkelaars. Samengevat komt onze inschatting er op neer dat behalve modules voor spraaksynthese, lexicale databanken, niet-geannoteerde corpora, en lingware op woordanalyseniveau niet veel "off-the-shelf" beschikbaar is voor de TST voor het Nederlands.

3.2 Belang van data voor modules en modules voor toepassingen

Uit onze analyse van welke dataverzamelingen relevant zijn voor de ontwikkeling van welke software modules (tabel 3) komt naar voor dat het meeste belang moet worden toegekend aan monolinguale lexicale databanken en geannoteerde corpora. Deze zijn immers van belang voor een brede waaier van lingware modules. Omwille van hun speciale aard en status zijn testcorpora niet opgenomen in deze tabel. Testcorpora zijn van belang voor alle modules en toepassingen.

Vanuit onze analyse van het belang van verschillende modules voor concrete toepassingen in tabel 4 komt vooral naar voor dat syntactische analyse (grammatica's, parsers, shallow parsing, POS tagging, morfologische analyse), tekstvoorverwerking (zinsgrenzendetectie, named entity recognition) en semantische analyse nuttige componenten zijn in een brede waaier toepassingen. Voor specifieke spraaktoepassingen is dat uiteraard volledige spraakherkennings of -synthese modules, maar ook prosodievoorspelling, unit selection en difoonsynthese.

4 BATAVO

Basis TST-voorzieningen zijn componenten die met voldoende kwaliteit beschikbaar zouden moeten zijn voor onderzoek en taal- en spraakindustrie als uitgangspunt voor verdere ontwikkeling zowel in onderzoek als in de ontwikkeling van toepassingen. De criteria om een component op te nemen in de BATAVO zijn: (i) beschikbaarheid van de technologie om de component met voldoende kwaliteit te construeren, en (ii) belang van de component in een ruime scala van toepassingen. De architectuur van deze modules en de codering van de dataverzamelingen moet zo open mogelijk zijn, en gebruik maken van bestaande Europese of wereldstandaarden waar mogelijk (bijv. TEI, Eagles, etc.)

Als we de aandacht beperken tot tabellen 3 en 4 komt hieruit een volgende BATAVO naar voor.

4.1 Voor taaltechnologie

1. Modules:
   (a) Robuuste modulaire tekstvoorverwerking: tokenisation (herkenning van voorkomens van woorden), aanbrengen zinsgrenzen in tekst en getranscribeerde spraak, *Named Entity Recognition* voor verschillende teksttypes.
   (b) Morfologische analyse en morfosyntactische disambiguering (POS tagging). Toekenning van contextueel relevante woordsoort van woorden (inclusief onbekende woorden) aan de hand van een voldoende gedetailleerde 'tagset'.
   (c) Syntactische analyse. Robuuste herkenning van de structuur van zinnen in tekst. De te gebruiken technologie hiervoor kan klassiek zijn (grammatica + parser), gebaseerd op statistische modellen, of op *shallow parsing* (constituentendetectie en toekennen van grammaticale relaties).
   (d) Aspecten van semantische analyse (met name woordbetekenisdisambiguering en oplossen van referentieproblemen).
2. Data:
   (a) Monolinguaal lexicon. Uitgebreide informatie over de woordvoorraad van het Nederlands (inclusief vertaalequivalenten in de belangrijkste talen).
   (b) Geannoteerd corpus geschreven Nederlands (een treebank met syntactische, morfologische en semantische structuren).
   (c) Benchmarks voor evaluatie van bovengenoemde modules en data, en voor de belangrijkste categorieën van toepassingen (test suites, testcorpora).
   (d) Opgelijnd parallel vertaalcorpus (Nederlands-Engels in elk geval, liefst meerdere talen).

4.2 Voor spraaktechnologie

1. Modules:
   (a) Automatische spraakherkenning, inclusief robuuste herkenning, herkenning van non-natives, adaptatie van de spraakherkenner, en prosodieherkenning.
   (b) Spraaksynthese, inclusief tools voor unit selection.
   (c) Tools voor het berekenen van betrouwbaarheidsmaten en het uitvoeren van uitingverificatie.
   (d) Tools voor identificatie, zowel sprekeridentificatie als taal- en dialectidentificatie.
   (e) Tools voor (semi-)automatische annotatie van spraakcorpora (labels op verschillende niveaus: segmenteel, prosodisch, syntactisch, semantisch, pragmatisch).
2. Data:
   (a) Spraakcorpora voor specifieke, belangrijke applicaties zoals bijv. directory assistance, customer care, CALL.
   (b) Multimodale spraakcorpora: corpora die naast spraak ook gegevens van andere modaliteiten bevatten.

(c) Multimedia spraakcorpora: corpora die naast spraak van radio en TV ook informatie van andere media bevatten (bijv. teksten en figuren van WWW, kranten, tijdschriften e.d.).

(d) Multilinguale spraakcorpora: corpora die naast Nederlandse spraak ook spraak van andere talen bevatten.

(e) Benchmarks voor evaluatie (test suites, test corpora).

4.3 Prioriteitenlijst

Wanneer we de voorgaande BATAVO combineren met de inschatting van beschikbaarheid van de verschillende halffabrikaten in tabel 1 en 2 (componenten en dataverzamelingen die al redelijk goed beschikbaar zijn komen niet in aanmerking voor een hoge prioritering), komen we tot de volgende prioritering.

Voor taaltechnologie:

1. Geannoteerd corpus geschreven Nederlands (een treebank met syntactische, eventueel morfologische structuren).
2. Robuuste modulaire tekstvoorverwerking: tokenisation (herkenning van voorkomens van woorden), indeling van tekst in zinnen, *Named Entity Recognition* voor verschillende teksttypes.
3. Syntactische analyse. Robuuste herkenning van de structuur van zinnen in tekst. De te gebruiken technologie hiervoor kan klassiek zijn (grammatica + parser), gebaseerd op statistische modellen, of op *shallow parsing* (constituentendetectie en toekennen van grammaticale relaties).
4. Semantische annotaties voor de hogergenoemde treebank.
5. Vertaalequivalenten in belangrijkste talen voor basislexicon.
6. Benchmarks voor evaluatie (test suites, test corpora).

Voor spraaktechnologie:

1. Automatische spraakherkenning, inclusief robuuste herkenning, herkenning van non-natives, adaptatie van de spraakherkenner, en prosodieherkenning.
2. Spraakcorpora voor specifieke, belangrijke applicaties zoals bijv. directory assistance, customer care, CALL.
3. Multimedia spraakcorpora: corpora die naast spraak van radio en TV ook informatie van andere media bevatten (bijv. teksten en figuren van WWW, kranten, tijdschriften e.d.).
4. Tools voor (semi-)automatische annotatie van spraakcorpora (labels op verschillende niveaus: segmenteel, prosodisch, syntactisch, semantisch, pragmatisch).
5. Spraaksynthese inclusief tools voor unit selection.

6. Benchmarks voor evaluatie van spraaktechnologische toepassingen (test suites, testcorpora).

Hierbij moet worden opgemerkt dat voor de spraaktechnologie meer dan voor de taaltechnologie geldt dat nog veel generisch (niet-taalspecifiek) onderzoek nodig is om de doelstellingen voor het Nederlands te kunnen realiseren.

In de volgende sectie formuleren we de uiteindelijke aanbevelingen van onze studie.

5 Aanbevelingen

De Taal- en Spraaktechnologie (TST) is voor het Nederlandse taalgebied een sector van uitzonderlijk economisch en cultureel belang. Economisch belang omdat TST als ingebedde technologie een belangrijke rol gaat spelen in alle vormen van ICT. Cultureel belang omdat het voortbestaan van het Nederlands ernstig in gevaar dreigt te komen wanneer deze ICT toepassingen niet 'gelokaliseerd' worden voor ons taalgebied.

Een gezonde TST-industrie heeft niet alleen behoefte aan een goed functionerende kennisinfrastructuur (academische opleiding en onderzoek) en investeringen, maar ook aan de beschikbaarheid van een voldoende TST-infrastructuur in termen van basale software modules, lexica, corpora, benchmarks en andere hulpmiddelen. Te vaak opnieuw moeten dezelfde hulpmiddelen worden ontwikkeld door applicatiebouwers waardoor te veel tijd en middelen verloren gaan.

In onze studie hebben we gevonden dat de huidige beschikbare TST-infrastructuur verspreid, onvolledig, slecht beschikbaar, en vaak van onvoldoende kwaliteit is om bruikbaar te zijn voor de taal- en spraakindustrie.

Een andere opvallende conclusie van ons onderzoek is dat er nauwelijks sprake is van werk aan de objectieve en methodologisch verantwoorde vergelijking en benchmarking van TST-modules en -data.

Vanuit een analyse van het belang van verschillende onderdelen van een TST-infrastructuur voor commercieel interessante toepassingen zijn we gekomen tot de definitie van een BATAVO, de basis TST-voorzieningen die *zo veel mogelijk* vrij (of goedkoop) beschikbaar zouden moeten zijn voor onderzoek en commerciële ontwikkeling (sectie 4).

Vanuit een analyse van de beschikbaarheid van de onderdelen van die BATAVO zijn we gekomen tot een lijst van prioriteiten voor toekomstig toepassingsgericht TST-onderzoek met openbare middelen (sectie 4.3).

Meer specifiek willen we de volgende aanbevelingen richten aan de opdrachtgevers van onze studie.

- Creatie en financiering van een organisatievorm die een verzamelpunt vormt voor Nederlandse TST-infrastructuur[2], met de volgende opdrachten:
  - Verzameling, documentatie en onderhoud van de al bestaande onderdelen van de BATAVO voor het Nederlands. Eventueel kan dit ook door aankoop van al bestaande software of data. Er moet vermeden worden dat overheidsgeld wordt gebruikt voor de ontwikkeling van al bestaande software of data.

[2]Het was niet onze opdracht na te gaan wat hiervoor de meest geschikte organisatievorm is, maar willen er toch de aandacht op vestigen dat zo'n organisatie niet noodzakelijk fysiek op één locatie gevestigd moet zijn. Om er voor te zorgen dat het initiatief voldoende steun en inzet van de volledige kennisinfrastructuur krijgt, zou het ook gedistribueerd ondergebracht kunnen worden bij meerdere groepen actief in TST.

  - Vervollediging van de BATAVO door stimulering van de reguliere fondsenverstrekkers tot financiering van projecten op het terrein van de door ons voorgestelde prioriteiten of door eigen initiatieven op dat gebied.
  - Aanbieden van de BATAVO aan onderzoek en taalindustrie, *zoveel mogelijk* onder de filosofie van *open source* ontwikkeling: het volledig vrij maken van de ontwikkelde software en data voor alle geïnteresseerden voor alle toepassingen ter stimulering van verdere ontwikkeling en uitbreiding ervan. Wellicht zal dit niet meteen voor alle onderdelen van de BATAVO op dezelfde manier mogelijk zijn (bijv. voor al bestaande, in een commerciële context ontwikkelde componenten), maar als doelstelling op langere termijn verdient een open source BATAVO de voorkeur[3]. In elk geval zouden met overheidsmiddelen ontwikkelde componenten en data *in elk geval* vrij beschikbaar moeten zijn; iets wat in het verleden niet altijd het geval was. Een haalbare initiële vorm lijkt een BATAVO met twee types van componenten: (i) volledig open source (door aankoop of ontwikkeling met overheidsmiddelen), en (ii) op basis van commerciële licentie (waarbij overheid bemiddelt en eventueel tussenkomt bij aanschaf van de licentie ten behoeve van specifieke gebruikers). Op die manier kan samenwerking met de overheid bij de ontwikkeling van de BATAVO ook voor de bedrijven aantrekkelijk worden gemaakt.
  - Opstellen van benchmarks en testcorpora en een methodologie voor de objectieve vergelijking, evaluatie en validatie van onderdelen van de BATAVO. Door verschillende contactpersonen (zowel in bedrijven als onderzoek) wordt dit laatste een absolute topprioriteit genoemd.

- Daarnaast is er behoefte in de TST-industrie aan meer en beter opgeleide TST-onderzoekers. Om hieraan tegemoet te komen moet de huidige afbouw van academisch TST-onderzoek en -opleiding een halt toegeroepen worden en waar nodig moeten integendeel relevante opleidingen structureel versterkt worden. De kennisinfrastructuur in Vlaanderen en Nederland zou zelf meer moeten openstaan voor uitwisseling van ideeën en software.

- We willen er tenslotte op wijzen dat naast de stimulering van toepassingsgericht onderzoek ter vervollediging van een BATAVO voor het Nederlands, eveneens voldoende middelen moeten worden vrijgemaakt voor *fundamenteel* onderzoek in taal- en spraaktechnologie. De belangrijkste problemen van de TST (bijv. contextueel bepaalde ambiguïteit, het begrijpen van tekst en spraak) die toepassingen als robuuste spraakherkenning en automatisch vertalen in de weg staan, zijn nog steeds onopgelost. Voor de oplossing ervan zullen nieuwe modellen en technieken nodig zijn, aangebracht vanuit multidisciplinair fundamenteel onderzoek.

In het vervolg van dit document onderbouwen we onze aanbevelingen met recente informatie over de beschikbare halffabrikaten voor de Nederlandse TST. We bouwen hierbij voort op een bestaande studie [23] die we up-to-date maken en uitbreiden waar nodig. Graag willen we op deze plaats onze dank betuigen aan iedereen (te veel mensen om op te noemen) die ons heeft geholpen met informatie, commentaar en suggesties.

[3]De geschiktheid van verschillende types van open source licenties voor de BATAVO zou verder moeten worden onderzocht. Zie bijv. `http://www.opensource.org` voor voorbeeldlicenties.

De hier genoemde personen vormen een belangrijke deelverzameling van deze groep:

Louis ten Bosch	L&H
Gies Bouwman	KUN
Jesse de Does	INL
Hans van Halteren	KUN
Henk van den Heuvel	SPEX
Joop Kerkhoff	KUN
Esther Klabbers	TUE
Kees Koster	KUN
Erwin Marsi	KUB
Roeland Ordelman	UTwente
Richard Piepenbrock	KUN
Rob van Son	UvA
Marc Swerts	TUE
Mariët Theune	UTwente
Vincent Vandeghinste	KU Leuven
Johan de Veth	KUN
John van der Voort van de Kleij	INL
Stichting NOTAS	

MODULES	Beschikbaarheid
Grafeem-foneemomzetting	8
Tekstvoorverwerking	4
Tokenisering	9
Zinsgrenzendetectie	3
Naamherkenning	4
Spellingcontrole	8
Morfologie	8
Lemmatisering	9
Analyse	7
Synthese	9
Woordsoortdisambiguering	7
Syntactische analyse	3
parsers en grammatica's	3
Shallow Parsing	2
Constituentenherkenning	5
Semantische en pragmatische analyse	2
Referentresolutie	2
Woordbetekenisdisambiguering	2
Pragmatische analyse	1
Tekstgeneratie	3
Taalpaarafh. vertaalmodules	3
Volledige spraakherkenning	4
Akoestische modellen	8
Taalmodellen	3
Uitspraaklexicon	5
Robuuste spraakherkenning	2
Non-native spraakherkenning	2
Sprekeradaptatie	2
Lexicon-adaptatie	2
Prosodieherkenning	2
Volledige spraaksynthese	6
Allofoonsynthese	7
Difoonsynthese	6
Unit selection	1
Prosodievoorspelling	3
Foonstringbewerkingen	?
Automatische fonetische transcriptie	3
Automatische fonetische segmentering	5
Oplijnen fonemen	8
Afstandsberekening fonemen	8
Sprekerherkenning	2
Sprekerverificatie	2
Sprekeridentificatie	2
Spreker tracking	2
Taalidentificatie	2
Dialectidentificatie	2
Betrouwbaarheidsmaten	2
Uitingverificatie	2

Tabel 1: Geschatte beschikbaarheid van TST-modules. Zie 2.2.1 voor een beschrijving van

DATA	Beschikbaarheid
Corpora	
Niet-geannoteerd	9
Geannoteerd	5
Spraak	4
Multilinguaal	3
Multimodaal	1
Multimediaal	1
Testcorpora	1
Lexica	
Monolinguaal	8
Multilinguaal	6
Thesauri	4

Tabel 2: Geschatte beschikbaarheid van data. Zie 2.2.2 voor een beschrijving van de verschillende types data.

MODULES	monol lex	multil lex	thes	anno corp	unanno corp	spraak corp	multil corp	multimo corp	multime corp
Grafeem-foneemomzetting	++			++					
Tokendetectie	++			+	++				
Zinsgrenzendetectie	+			++	++				
Naamherkenning	+	+	+	++	++	++			
Spellingcorrectie									
Lemmatisering	++			++	+				
Morfologische analyse	++			++	+				
Morfologische synthese	++			++	+				
Woordsoortdisambiguering	++			++	+				
Parsers en grammatica's	++			++					
Shallow Parsing	++			++	++				
Constituentenherkenning	++			++	+				
Semantische analyse	++		++	++				++	++
Referentresolutie	+		++	++	+				
Woordbetekenisdesambig.	+		++	++	+				
Pragmatische analyse	+		+	++				++	++
Tekstgeneratie	++		++	++				++	++
Taalpaarafh. Vertaalmodules		++	++	++			++		
Volledige spraakherkenning	++	+		++	+	++	+	++	++
Akoestische modellen	++	+		++	+	++	+	+	+
Taalmodellen	+			++	+	+	+	+	+
Uitspraaklexicon	++	+		+		++	+	+	+
Robuuste spraakherkenning	+			+	+	+	+	+	++
Non-native spraakherkenning	+	++		+		++	++	+	+
Sprekeradaptatie	+			+	+	++	+	+	++
Lexiconadaptatie	++	+		+		++	+	+	+
Prosodieherkenning	+	+		++	+	++	+	+	+
Volledige spraaksynthese	++	+		+		+		+	
Allofoonsynthese	+	+		+		+		+	
Difoonsynthese	++	+		+		+		+	
Unit selection	++	+		+		+		+	
Prosodievoorspelling	++	+		+		+		+	+
Autom. fon. transcriptie	++	++		+	+	++	+	+	+
Autom. fon. segmentering	++	++		+	+	++	+	+	+
Oplijnen fonemen	+	+		+		++	+	+	+
Afstandsberekening fonemen	+	+		+		++	+	+	+
Sprekeridentificatie	+			++	++	++	+	++	+
Sprekerverificatie	+			++	++	++	+	++	
Spreker-tracking	+			++		++			++
Taalidentificatie	+	++		+	+	++	++	+	+
Dialectidentificatie	+	++		+	+	++	++	+	+
Confidence measures	+			+	+	++	+	++	+
Utterance verification	+			+	+	++	+	+	+

Tabel 3: Belang van data voor de modules. Deze tabel geeft aan welke dataverzamelingen meer of minder noodzakelijk zijn voor het ontwikkelen van de verschillende modules: heel noodzakelijk (++), in zekere mate noodzakelijk (+), en weinig noodzakelijk (geen symbool).

MODULES	CALL	toegangs controle	invoer	uitvoer	interf	doc prod	info toegang	vertalen
Grafeem-foneemomzetting	+			++	++	+	+	
Tokendetectie	+		+		+	+	+	+
Zinsgrenzendetectie	+		++	++	+	++	++	++
Naamherkenning	+		++	++	+	++	++	++
Spellingcorrectie	+							
Lemmatisering	+		+	+	+	+	+	+
Morfologische analyse	+		+	++	+	++	++	++
Morfologische synthese	+			++	+	++		++
Woordsoortdisambiguering	+		++	x	++	++	++	++
Parsers en grammatica's	+		++	++	++	++	++	++
Shallow Parsing	+		++	++	++	++	++	++
Constituentenherkenning	+		++	++	++	++	++	++
Semantische analyse	+		++	++	++		++	++
Referentresolutie	+		++		++	++	++	++
Woordbetekenisdesambig.	+		++	+	+	+	++	++
Pragmatische analyse	+		++	++	++		+	++
Tekstgeneratie	+			++	++	++		++
Taalpaarafh. vertaalmodules	+						++	++
Volledige spraakherkenning	++	++	++		++	++	++	++
Akoestische modellen	++	+	++		++	+	+	+
Taalmodellen	++	+	++		++	++	++	++
Uitspraaklexicon	++	+	++	+	++	+	++	++
Robuuste spraakherkenning	+	+	++		++	+	+	+
Non-native spraakherkenning	++	+	+		+		+	+
Sprekeradaptatie	+	+	++		+	+	++	+
Lexiconadaptatie	++	+	++	+	++	+	++	++
Prosodieherkenning	++	+	++		++	++	++	++
Volledige spraaksynthese	+			++	++	+	+	++
Allofoonsynthese	+			+		+	+	+
Difoonsynthese	+			++	++	+	+	+
Unit selection	+			++	++	+	+	+
Prosodievoorspelling	++			++	++		+	++
Autom. fonet. transcriptie	++	+	+	+	+	+	+	+
Autom. fonet. segmentering	++	+	+	+	+	+	+	+
Oplijning fonemen	++	+	+		+			+
Afstandsberekening fonemen	++	+	+		+			+
Sprekeridentificatie	+	++	+		+		+	+
Sprekerverificatie	+	++	+		+		+	+
Spreker-tracking	+	++	+		+	+	+	+
Taalidentificatie	+	+	+		+		+	+
Dialectidentificatie	+	+	+		+		+	+
Confidence measures	++	++	++		++	+	+	+
Utterance verification	+	+	++		++	+	+	+

Tabel 4: Belang van modules voor klassen van toepassingen. Deze tabel geeft aan welke modules (vertikaal) meer of minder noodzakelijk zijn voor het ontwikkelen van de verschillende toepassingen (horizontaal): heel noodzakelijk (++), in zekere mate noodzakelijk (+), en weinig noodzakelijk (geen symbool).

Deel II
Inventarisatie en evaluatie

6 Algemene criteria

In onze studie veronderstellen we strenge criteria voor de beoordeling van bestaande hulpmiddelen. Dat is noodzakelijk omdat het uitgangspunt van onze opdracht de identificatie en evaluatie is van hulpmiddelen die bruikbaar zijn voor de ontwikkeling van commerciële toepassingen door TST-bedrijven. Dat betekent minimaal dat de kosten noodzakelijk om een bestaande module of dataverzameling bruikbaar te maken voor toepassingen significant lager moeten zijn dan de kosten voor het helemaal opnieuw ontwikkelen ervan.

Voor alle hulpmiddelen gaan we uit van de volgende minimale vereisten.

- Beschikbaarheid. Het hulpmiddel moet verkregen kunnen worden (onder licentie of voor een eenmalige kostprijs) voor een kost die acceptabel is. Idealiter is het vrij beschikbaar.
- Bruikbaarheid. Het hulpmiddel moet aantoonbaar integreerbaar en uitbreidbaar zijn in toepassingen.
- Documentatie. Er moet voldoende documentatie beschikbaar zijn.
- Kwaliteit. De kwaliteit van het hulpmiddel moet objectief onderzocht en gemeten zijn, en een minimaal niveau overschrijden. Dat niveau operationaliseren we als de 'state of the art' van hetzelfde hulpmiddel voor andere talen.

De rest van dit deel van het rapport is onderverdeeld in vier secties: taaltechnologiemodules (sectie 7), taaltechnologiedata (sectie 8), spraaktechnologiemodules (sectie 9), en spraaktechnologiedata (sectie 10). De omvang van de verschillende onderdelen is niet overal dezelfde omdat niet overal evenveel te inventariseren valt. Om diezelfde reden zijn sommige aparte onderdelen uit Deel I samengenomen of lichtjes anders ingedeeld.

7 Taaltechnologie modules

7.1 State of the art internationaal

Een goede (maar inmiddels wat verouderde) beschrijving van de state of the art van taal- en spraaktechnologie is opgeschreven in 'Survey of the State of the Art in Human Language Technology'[33]. Een on line versie is te vinden op `http://cslu.cse.ogi.edu/HLTsurvey`.

Een interessante plaats op het internet is de Natural Language Software Registry (NLSR). Een beschrijving hiervan wordt hieronder gegeven. Vanuit deze website kun je ook terechtkomen op de belangrijkste repositories en groepen en organisaties die zich met taaltechnologie bezighouden.

- De Natural Language Software Registry (NLSR)
  `http://registry.dfki.de/`

  De Natural Language Software Registry (NLSR) levert een beknopte samenvatting en de sources van een grote hoeveelheid natuurlijke taalverwerkingssoftware die beschikbaar is voor de taaltechnologiegemeenschap. Het omvat zowel academische als commerciële software, met specificaties en condities over hoe de software te verkrijgen is. De NLSR concentreert zich vooral op taaltechnologie lingware, maar het laat de andere natuurlijke taal resources niet volledig buiten beschouwing.

  De NLSR is een database waar gebruikers kunnen browsen of zoeken in de database van producten. Producten zijn verdeeld in verschillende secties, zoals onder andere geschreven taal, gesproken taal, evaluatiehulpmiddelen etc. Binnen deze secties zijn de softwareproducten weer onderverdeeld per module, zoals POS tagging, morfologische analyse etc. De classificatie is grotendeels gebaseerd op de taxonomie zoals gehanteerd is in de eerder genoemde 'Survey of the State of the Art in Human Language Technology' [33].

- Informatie over lexica en corpora kan ook gevonden worden bij ELRA/ELDA
  `http://www.icp.inpg.fr/ELRA/`

  of de Linguistic Data Consortium (LDC)
  `http://www.ldc.upenn.edu/`

- EAGLES
  `http://www.ilc.pi.cnr.it/EAGLES/home.html`

  De Expert Advisory Group on Language Engineering Standards (EAGLES) is een initiatief van de Europese Commissie. Het beoogt het versnellen van de definitie van standaarden voor:

  - Grootschalige taal resources (zoals tekstcorpora, computationele lexica en spraakcorpora).
  - Manieren om kennis te manipuleren via computationeel linguïstische formalismen, mark-up talen en software tools.
  - Manieren om resources, hulpmiddelen en producten te evalueren.

- ELSNET
  `http://www.elsnet.org`

ELSNET onderhoudt een ander belangrijk portaal voor informatie over taal- en spraaktechnologie.

7.2 Specifieke criteria voor modules

Behalve aan de algemene criteria in sectie 6 moeten aan alle softwaremodules bijkomende criteria gesteld worden.

- Modulariteit en open architectuur. De taalkundige kennis en de controlestructuren, en eventuele verschillende niveaus van taalkundige kennis moeten modulair opgesteld zijn en gemakkelijk aanpasbaar en uitbreidbaar.
- Accuraatheid. Kwaliteit van modules kan vooral worden gemeten aan de hand van de accuraatheid ervan. Die kan gemeten worden in termen van aantal correcte antwoorden, *recall* (hoeveel van de gewenste outputs heeft het systeem correct geproduceerd), en *precision* (hoeveel van de outputs van het systeem waren correct).
- Efficiëntie. De tijd en geheugenruimte nodig voor normaal gebruik van de module moet acceptabel zijn.

7.3 Grafeem-naar-foneemomzetting

Bij het omzetten van de spelling van een woord naar een uitspraakrepresentatie, gedetailleerd genoeg om spraaksynthesesoftware aan te sturen, spelen een aantal subproblemen een rol: de morfologische structuur van een woord kan de uitspraak beïnvloeden ('bedel+en vs. be+'del+en), tegelijk met de uitspraak moet ook de woordklemtoon worden bepaald (die is in het Nederlands voor een gedeelte onvoorspelbaar), en de lettergreepstructuur van een woord speelt ook een rol, aangezien processen als assimilatie zich afspelen op syllabeniveau. Verder is er nogal wat (regionale) variabiliteit in uitspraak (bijvoorbeeld reductie van vocalen en stemloosheid van sommige consonanten) die voor toepassingen in spraaksynthese en -herkenning gemodelleerd zou moeten worden. Een grafeem-foneemomzetter, aangevuld met tekstanalysecomponenten, is in de eerste plaats een belangrijke taaltechnologische component binnen spraaksynthesesystemen, maar kan ook een rol spelen in andere toepassingen. Bijvoorbeeld, in spelfoutencorrectie kan het nuttig zijn om de uitspraak van woorden te berekenen omdat spelfouten vaak te wijten zijn aan uitspraakspelling.

Specifieke criteria voor grafeem-naar-foneem modules zijn de volgende:

- Algemene eigenschappen
  - Programmeertaal
  - Besturingssysteem
  - Ontwikkelmethode (inductief/deductief)
  - Gebruikte foneemset
  - Output (uitspraakvarianten?)
  - Componenten (morfologie, syllabestructuur, klemtoon)
- Beschikbaarheid

- Bruikbaarheid
  - Integreerbaarheid/modulariteit
  - Aanpasbaarheid (nieuwe regionale variant, foneemset)
- Documentatie
- Kwaliteit
  - Accuraatheid
  - Precision en recall per foneem
  - Snelheid
  - Geheugen

7.3.1 State of the art internationaal

Er zijn zowel kennisgebaseerde als statistische benaderingen van het probleem. Voor het Engels en vele andere talen lijkt een regelgebaseerde benadering standaard (bijv. [5]). Ook *finite-state transducers* [114] worden vaak gebruikt. Hierbij wordt de relatie tussen lexicale representatie en uitspraak gelegd met behulp van parallelle eindige automaten die de uitspraakregels implementeren. Statistische benaderingen maken vaak gebruik van analogiemodellen (zie [43] voor een overzicht en vergelijking voor het Engels). Ook voor deze toepassing in het Nederlands werden zowel kennisgebaseerde als statistische methodes gebruikt.

Op het niveau van academisch onderzoek werd voor het Nederlands onder meer gewerkt in het kennisgebaseerde paradigma aan een expertsysteem (frames + rules) benadering [40], en regelgebaseerd [106], in de statistische benadering met variabele lengte probabilistische regels [50], neurale netwerken [143], memory-based learning [41], regelinductiemethodes [71], en hybride vormen [20] (finite-state + regelinductie).

7.3.2 Inventaris beschikbare grafeem-foneemomzetters Nederlands

Voor zover de grafeem-foneemcomponenten geïntegreerd zijn in de context van een volledig spraaksynthesesysteem, zie sectie 9.

- TREETALK is een methode voor grafeem-foneemomzetting op woordniveau gebaseerd op IGTREEs ([41]). Het is een lerende methode die een uitspraakwoordenboek (spelling van woordvormen met corresponderende uitspraakrepresentaties) compact opslaat in een beslissingsboom, en voor onbekende woorden extrapoleert vanuit die kennisstructuur. De methode is ook toepasbaar voor klemtoontoekenning en lettergreepsplitsing in de context van grafeem-foneemomzetting. Deze taalonafhankelijke aanpak werd toegepast voor Vlaams (op basis van Fonilex[4]) en Nederlands (op basis van Celex[5]).

  Een on line demoversie is te vinden bij
  `http://ilk.kub.nl/demos.html`

[4] `http://bach.arts.kuleuven.ac.be/fonilex/`
[5] `http://www.kun.nl/celex/`

- Bij het MORPA cum MORPHON systeem gebeurt grafeem-foneemconversie in twee stappen. Allereerst splitst MORPA een woord op in een lijst van morfemen. Deze morfemen worden opgezocht in een lexicon. Elk morfeem is geassocieerd met zijn categorie en een fonemische transcriptie. De fonemische transcripties van de opeenvolgende morfemen worden aan elkaar geplakt om de onderliggende fonemische representatie van het woord te verkrijgen. Vervolgens past MORPHON een aantal fonologische regels toe op deze onderliggende representatie, om zo tot de uiteindelijke uitspraak van het woord te komen.
- Van Dale G2P: grafeem naar foneemomzetting op woordniveau die gebruik maakt van een leermethode en het uitspraakwoordenboek compact opslaat in een beslissingsboom. Evenals bij TREETALK kan ook klemtoontoekenning en lettergreepsplitsing worden meegenomen in het leerproces. De Van Dale G2P is getraind op de Van Dale uitspraakwoordenboeken. Beschikbaar voor gebruik onder UNIX en Windows (info: hlt-group@cs.utwente.nl).

7.3.3 Evaluatie

Zie ook de evaluatiesectie bij volledige spraaksynthesesystemen (sectie 9) voor evaluatie van grafeem-foneem-omzettingssystemen.

TreeTalk

- *Algemene eigenschappen*
  - *Programmeertaal*: C++
  - *Besturingssysteem*: Unix
  - *Ontwikkelmethode: (inductief/deductief)* Inductief, voor het Nederlands gebaseerd op CELEX en FONILEX.
  - *Gebruikte foneemset*: SAMPA
  - *Output (uitspraakvarianten?)*: Afhankelijk van trainingsmateriaal.
  - *Componenten (morfologie, syllabestructuur, klemtoon)*: Afhankelijk van trainingsmateriaal.
- *Beschikbaarheid*: Licentie op getrainde systemen. Daarvoor moet soms ook een licentie op het trainingsmateriaal worden genomen (in het geval van CELEX).
- *Bruikbaarheid*
  - *Integreerbaarheid/modulariteit*
  - *Aanpasbaarheid (nieuwe regionale variant, foneemset)*: Hertrainbaar.
- *Documentatie*: Systeembeschrijving.
- *Kwaliteit*
  - *Accuraatheid*: afhankelijk van kwaliteit training materiaal
  - *Precision en Recall per foneem*
  - *Snelheid*: real-time
  - *Geheugen*

7.3.4 Conclusie

De meeste software voor grafeem-naar-foneemomzetting is geïntegreerd met andere componenten in volledige spraaksynthesesystemen. TREETALK is een efficiënt, accuraat, en gemakkelijk hertrainbaar lerend systeem; de kwaliteit ervan en de mogelijke output hangt in grote mate af van de beschikbare trainingdata. Er bestaan getrainde systemen voor Nederlands en Vlaams. Het systeem is nog niet beschikbaar op andere platformen dan Unix. Hoewel de technologie goed ontwikkeld is (als onderdeel van complete spraaksynthese), valt de beschikbaarheid van losse grafeem-naar-foneemomzettingsmodules tegen.

7.4 Tekstvoorverwerking

Met tekstvoorverwerking bedoelen we hier een aantal processen die moeten of kunnen gebeuren vooraleer een taaltechnologiesysteem iets zinvols kan doen met inputtekst. Het gaat hier om:

1. Herkenning van documentstructuur (scheiding van grafische elementen en tekst, interpretatie of verwijdering van eventuele tags en codes).
2. Indelen van de tekst in zinnen.
3. Herkenning van lexicale items in de tekst
   (tokenisatie, herkenning van meer-woorduitdrukkingen).
4. Naamherkenning en -interpretatie.

Omwille van de beperkte beschikbaarheid van deze systemen voor het Nederlands geven we hier geen meer specifieke evaluatiecriteria. Deze zijn vergelijkbaar met die van andere modules, met dit verschil dat accuraatheid (in termen van recall en precisie) moet worden berekend op entiteiten als zinnen, woordvormen, en namen.

7.4.1 Segmentatie in zinnen en tokenisatie

Segmentatie van tekst in zinnen levert een oplossing voor bepaalde dubbelzinnige gevallen (bijv. waar een punt wordt gebruikt als deel van een afkorting en niet als aanduiding van het einde van een zin). Tokenisatie is een proces dat een gedetecteerde zin opdeelt in een reeks van tokens. Tokens zijn de woorden, spaties, punctuatietekens etc.

State of the art internationaal

Het segmenteren van tekst in zinnen is goed leerbaar met inductieve technieken:

- *Beslissingsbomen.* Riley [113] rapporteert 99.8% correct op het Brown Corpus.
- *Maximum entropy modellen.* MXTERMINATOR, een JAVA (JDK 1.1) implementatie van de zinsgrenzendetector zoals beschreven in Reynar & Rathnaparkhi [112] haalt 98.8% correct op het Wall Street Journal corpus, met een specifieke Engelse versie en 98.0% met een meer taalonafhankelijke versie. Mikheev rapporteert 99.2% correct op het Wall Street Journal corpus.

- *Neurale netwerken.* SATZ (`http://elib.cs.berkeley.edu/src/satz/`) is een serie van modules (geschreven in C) die dient om zinsgrenspunctuatie te disambigueren. Het is ontwikkeld om aanpasbaar te zijn aan nieuwe tekstgenres en talen, en gaat zo voorbij aan de beperkingen van regelgebaseerde technieken. Het programma gebruikt POS frequentie-informatie als input voor een neuraal netwerk, dat getraind is om grenzen te labelen. SATZ neemt een tekstfile als input en produceert als output de individuele tokens van de file op aparte regels. Voor een uitgebreide beschrijving van de gebruikte methodiek, zie Palmer & Hearst [108], waar 98.5% correct op het Wall Street Journal corpus wordt gerapporteerd.

  De source-code kan worden gedownload via:
  ftp://elib.cs.berkeley.edu/pub/archive/source_code/satz.tar.Z.

Er is evenwel nooit onderzocht hoe goed deze getrainde systemen presteren op andere soorten tekst dan die waarop werd getraind. Er zijn ook regelgebaseerde methodes die soortgelijke accuraatheid halen. Bijv. Grefenstette [66] rapporteert 99.1% correct op het Brown corpus.

Tokenisatie is een uiterst praktisch probleem dat meestal op een ad hoc manier wordt opgelost voor specifieke systemen en specifieke input. De gebruikte 'technologie' is vaak gebaseerd op reguliere expressies in scripting-talen. Verschillende van die systemen zijn vrij verkrijgbaar:

- Sed-script voor de Penn-treebankstijl van tokeniseren. Deze zijn te downloaden via `http://www.cis.upenn.edu/~treebank/tokenization.html`.
- Edinburgh LT TTT: Dit is een teksttokenisatiesysteem en toolset die werkt onder Solaris 2.5. De toolset is te verkrijgen via `http://www.ltg.ed.ac.uk/software/ttt/`.

Er zijn ook meer uitgebreide, taalonafhankelijke, oplossingen beschikbaar (Xerox, LinguistiX, Multexttools).

Inventaris beschikbare systemen Nederlands

- Xerox

  Bij Xerox is een tokenisator ontwikkeld voor het Engels, Frans, Spaans, Italiaans, Portugees, Nederlands en Duits. Tokenisatie gebeurt via finite-state transducers, gecompileerd uit tokenisatieregels. Dezelfde techniek gebruiken zij voor morfologische analyse (zie 7.6.2), met het enige verschil dat voor tokenisatie gebruik wordt gemaakt van *directed replacement* operatoren (zie [78] voor een beschrijving).

  Er is een on line demo versie van de tokeniseerder:
  `http://www.xrce.xerox.com/research/mltt/demos/dutch.html`

- LinguistX Tokenizer

  Deze tokeniseerder is onderdeel van het Inxight LinguistX Platform. Inxight is een dochteronderneming van Xerox (`http://www.inxight.com`). Het Platform is een collectie van softwarecomponenten die automatische taalidentificatie, tokenisatie, stemming en POS tagging uitvoeren. En dit alles voor verschillende talen waaronder het Nederlands.

De tokeniseerder breekt een document op in woorden, punctuatietokens, en categoriseert ieder token. Omdat iedere taal verschillende regels heeft voor het opbreken van tekst, is er een tokenisatie module voor iedere taal. Behalve tokens (met een label) levert de module ook de locatie en lengte van ieder token in de originele input.

Eigenschappen:

- Deelt de tekst op in syntactische eenheden (tokens)
- Accepteert 'plain-text' of HTML input
- Werkt met Unix-, Mac-, en Windows-documenten
- Werkt voor het Nederlands, Engels, Frans, Duits, Italiaans, Portugees, Spaans, Fins, Noors (Bokmal en Nynorsk), Zweeds, Deens, Chinees, Koreaans en Japans.
- Detecteert zinsgrenzen in plain-text and HTML
- Detecteert paragraafgrenzen in plain-text and HTML

Besturingssysteem en compilers:

- Windows 2000 NT met Microsoft Visual C++ 6.0
- Solaris 2.7/7.8 met GNU C++ 2.95.2
- Red Hat Linux 6.2 op een Intel met GNU C++ 2.95.2

LinguistX Platform bevat:

- Een taal module voor woord- en frase-analyse
- Windows DLL en Solaris 'static library'
- C++ API
- C++ voorbeeldapplicaties
- Documentatie

- MULTEXT

  De 'Multext multilingual segmenter tool' (MTSeg) is een configureerbare tokeniseerder, ontwikkeld om multilinguale tekst te analyseren. Het doel van de segmenter is om tekst te splitsen in woorden en speciale tokens, zoals afkortingen en getallen, alsook bepaalde multiwoordeenheden, en het opsporen en markeren van zinsgrenzen.

  De segmenteerder is ontwikkeld volgens de 'Software Lego' principes, die Multext aangenomen heeft. Met andere woorden, het was ontworpen als een 'multi-purpose tool' opgebouwd uit een reeks van subtools, die ieder een uniek, specifiek probleem oplossen. De segmenteerder is een script dat de serie van subtools aan elkaar ketent via de unix 'pipe'. De subtools kunnen onafhankelijk van elkaar gebruikt worden, en de gebruiker kan tools in de reeks toevoegen of verwijderen voor bepaalde applicaties.

  De subtools voeren processen uit zoals het opdelen van tekst bij spaties, het isoleren van punctuatie, het identificeren van afkortingen, het hercombineren van samenstellingen, etc. De regels die bepalen hoe punctuatie behandeld moet worden, hoe afkortingen en samenstellingen te identificeren, etc., zijn beschikbaar als data voor de juiste subtools via een set van taalspecifieke, gebruikersgedefinieerde resource files, en zijn zo geheel aanpasbaar.

De segmenter kan overweg met de output van de interpreter van de "SGML query language" (SgmlQL), met 'plain' tekst (die geen markup bevat), of met een voorgedefinieerde tabular format tekst.

Kort samengevat, de taken van de subtools zijn:

- splits de tekst in een serie tokens
- geef ieder van de tokens een label, wat aangeeft wat voor type het is (afkorting, punctuatie, etc.)
- detecteer zinsgrenzen

MtSeg is vrij beschikbaar voor niet-commerciële, niet-militaire doeleinden. Het is ontwikkeld op Sun Workstations onder Solaris 2.4 en is getest onder Solaris 2.x. Het zou makkelijk overzetbaar naar ander Unix-varianten moeten zijn, maar is nog niet getest.

Mtseg is te downloaden via
`http://www.lpl.univ-aix.fr/projects/multext/MtSeg/MSG1.html` Algemene informatie over Multext is te vinden bij
`http://www.lpl.univ-aix.fr/projects/multext/`

- L&H International Proofreader[6] (`http://www.lhsl.com/tech/icm/proofing/pr.asp`)
  en L&H IntelliScope Document Summarizer
  (`http://www.lhsl.com/tech/icm/retrieval/toolkit/ds.asp`)
  Dit zijn twee systemen die als onderdeel van spellingcontrole of samenvatten tekstvoorverwerkingssoftware bevatten, zoals onder andere tokenisatie.

7.4.2 Naamherkenning

Naamherkenning omvat het identificeren en classificeren van persoonlijke, geografische, institutionele en andere namen, tijden en data, geld en percentages, en is een belangrijke taak voor talrijke toepassingen.

Een voorbeeld van named entities in een tekst volgt hieronder:

```
<persoon>Luc Steels</persoon> is een <functie>professor</functie> in
Artifici\"ele Intelligentie aan de <organisatie>VUB</organisatie> in
<plaats>Brussel</plaats> en <functie>directeur</functie> van
<organisatie>Sony Research labs</organisatie> in <plaats>Parijs</plaats>.
```

State of the art internationaal

In de context van de Message Understanding Conference (`http://www.ldc.upenn.edu/Catalog/MUC.html`) werd veel onderzoek gestimuleerd en systemen geëvalueerd op het gebied van naamherkenning.

Er zijn drie benaderingen om systemen te ontwikkelen: Gebruikmaking van lijsten (gazetteers), regels, of een combinatie van beide. Een succesvol voorbeeld van die laatste aanpak is die van Mikheev et al. [99](`http://www.ltg.ed.ac.uk/software/ttt/index.html`).

[6] Een deel van de technologie van het bedrijf Lernout & Hauspie, dat tijdens de voorbereiding van dit rapport failliet ging, werd overgenomen door ScanSoft.

Cucerzan en Yarowsky [39] ontwikkelden een taalonafhankelijk bootstrapping-algoritme, gebaseerd op iteratief leren en herwaardering van contextuele en morfologische patronen. Een voorbeeld van een systeem voor naamextractie uit spraak is Kubala et al. [87]. Dit systeem maakt gebruik van Hidden Markov modellen.

Inventaris beschikbare systemen Nederlands

Op basis van het ILK corpus (Nederlandse kranten) werden aan de Universiteit Tilburg gazetteers en een systeem ontwikkeld voor naamherkenning voor het Nederlands [29]. Een versie op basis van Vlaamse kranten is in ontwikkeling bij het CNTS (UIA).

7.4.3 Evaluatie en conclusie

Tekstvoorverwerking is een heterogene verzameling deelproblemen. Sommige daarvan (zinsdetectie en tokenisatie) lijken goed oplosbaar zowel met inductieve als deductieve technieken, maar de algemene beschikbaarheid ervan is gering en er zijn geen grondige evaluaties. Modules voor robuuste naamherkenning zijn zo goed als onbestaand voor het Nederlands.

Het belang ervan is nochtans niet te onderschatten. Kwalitatief goede, generieke en gemakkelijk aanpasbare tools voor tekstvoorverwerking in de meest voorkomende teksttypes van het Nederlands (vooral herkenning van types van namen) zou een belangrijke versnelling betekenen van de ontwikkeling van toepassingen. Nu moet dit telkens opnieuw gebeuren, en meestal op een ad hoc manier.

7.5 Spellingcontrole en -normalisatie

Spellingscorrectie is al sinds enige jaren in bijna alle commerciële en niet-commerciële tekstverwerkingsprogramma's standaard ingebouwd. Het is een functionaliteit die gerekend kan worden tot de categorie van zogenaamde auteurstools, waartoe ook de zogenaamde grammarcheckers gerekend kunnen worden. Van recentere datum is de losse Nederlandse spellingscorrector (plug-in) die door Van Dale Lexicografie in samenwerking met TNO is ontwikkeld voor gebruik in combinatie met MS Word. Voor allerlei taal- en spraaktechnologisch onderzoek en ontwikkelwerk wordt gebruik gemaakt van tekstcorpora, onder meer bij de ontwikkeling van statistische modellen. Dergelijke corpora zijn vaak erg 'vervuild' door spellingsvarianten, spellingsfouten,etc., en ook is het aantal unieke woorden vaak groter dan de woordenlijsten die bij de bewerking van die corpora gebruikt kunnen worden. Daarom is het van belang om de corpora aan spellingsnormalisatie te onderwerpen. In deze sectie zullen een aantal normalisatietools voor het Nederlands besproken worden die daarbij een rol kunnen spelen.

- **Vervangingstool DRUID**

  Voor het ontwikkelen van taalmodellen wordt vaak gebruik gemaakt van grote corpora. Binnen het DRUID-project (`http://dis.tpd.tno.nl/druid/`) wordt een spraakherkenner ontwikkeld voor de ontsluiting van nieuwsprogramma's op radio en televisie. De taalmodellen voor deze herkenner worden getraind op een tekstcorpus van 300 miljoen woorden. De tool die binnen DRUID is ontwikkeld voor de vervanging van spellingsvarianten voldoet aan de eisen die dergelijke grote bestanden opleggen aan de programma-efficiëntie.

Algemene eigenschappen: Beschikbaar als Perl-script voor gebruik onder Unix. Input is tekst en variantenlijst, output is opgeschoonde tekst.

Beschikbaarheid: Voor inlichtingen: Parlevink Human Language Technology Group, Universiteit Twente (email: hltgroup@cs.utwente.nl). Links naar papers op `http://wwwhome.cs.utwente.nl/~ordelman`.

- **Lijst van spellingsvarianten Van Dale**

  Sinds de invoering van de nieuwe spelling (1995) zijn er extra veel spellingsvarianten in omloop, onder meer die van het type pannekoek/pannenkoek. Voor het ontwikkelen van taalmodellen voor o.m. spraakherkenning is het van belang de trainingscorpora te ontdoen van deze dubbelspellingen. Van Dale Lexicografie beschikt over lijsten van paren met varianten.

  Beschikbaarheid: Voor inlichtingen: Parlevink Human Language Technology Group, Universiteit Twente (email: hltgroup@cs.utwente.nl).

- **Uitspeller afkortingen**

  In tekstuele trainingscorpora kunnen allerlei uitdrukkingen in de vorm van een afkorting worden weergegeven. De gesproken variant zal echter vaak voluit worden uitgesproken. Voor de ontwikkeling van een akoestisch model en/of taalmodel is het van belang die afkortingen om te zetten naar een uitgespelde vorm die overeenkomt met de gesproken variant.

  Algemene eigenschappen:

  Beschikbaar als Perl-script voor gebruik onder Unix. Input is tekst en lijst afkortingen, output is tekst waarin afkortingen zijn uitgespeld. Voor inlichtingen: Voor inlichtingen: Parlevink Human Language Technology Group, Universiteit Twente (email: hltgroup@cs.utwente.nl).

- **Numbersolver**

  In tekstuele trainingscorpora kunnen allerlei numerieke uitdrukkingen voorkomen. In welke vorm die uitdrukkingen uitgesproken worden varieert per type uitdrukking. Voor de ontwikkeling van een taalmodel is het van belang om de tekstuele variant om te zetten naar een uitgespelde vorm die overeenkomt met de gesproken variant. Deze bewerking komt voornamelijk neer op het invoegen van spaties. Ook voor TTS-systemen is die functionaliteit van belang.

  Algemene eigenschappen: Beschikbaar als ddl-routine, te gebruiken onder Windows. Input is tekst, output is tekst waarin numerieke uitdrukkingen omgezet zijn naar een gesproken variant.

  Beschikbaarheid: Voor inlichtingen: Voor inlichtingen: Parlevink Human Language Technology Group, Universiteit Twente (email: hltgroup@cs.utwente.nl).

7.6 Lemmatisering en morfologische analyse

Morfologische analyse is een basistechnologie die veel soorten van tekstanalyse mogelijk maakt. Het herkennen van de structuur en morfosyntactische categorie van woorden is de eerste stap voor POS-tagging, parseren, vertaling en andere high-level applicaties. De twee centrale problemen bij morfologische analyse zijn:

- **herkenning van woordstructuur** Woorden zijn opgebouwd uit kleinere betekenisvolle eenheden; de morfemen. Woordvormen die niet in het lexicon staan (bijv. de meeste samenstellingen) moeten herleid worden tot hun morfeemstructuur en er moet een morfosyntactische klasse aan toegekend worden.
- **oplossen van morfologische en orthografische variatie** De vorm van een morfeem hangt vaak af van de context: *huis* wordt *huiz* in de context van het meervoudsuffix *en*.

Lemmatisering is een vereenvoudigde vorm van morfologische analyse waarbij van een morfologisch complex woord de mogelijke morfosyntactische klassen en citatievormen (bijv. infinitief bij werkwoorden, enkelvoud bij substantieven) worden geproduceerd.

Stemming is een nog eenvoudigere variant van lemmatisering waarbij voor iedere vorm van een morfologisch paradigma (bijv. lopen, liep, liepen, loopt, lopend, gelopen) de stam wordt opgeleverd (loop).

7.6.1 Specifieke evaluatiecriteria

- Algemene eigenschappen
  - Programmeertaal
  - Besturingssysteem
  - Ontwikkelmethode (inductief/deductief)
  - Gebruikte morfosyntactische klassen
  - Output (citatievorm, klasse, interne (hiërarchische) structuur, stam)
  - Voor welk teksttype
  - Componenten (segmentatieroutine, lexicon, spellingvariatieregels)
- Beschikbaarheid
- Bruikbaarheid
  - Integreerbaarheid/modulariteit
  - Aanpasbaarheid (nieuw teksttype, klassesysteem)
- Documentatie
- Kwaliteit
  - Accuraatheid
  - Precision en recall per woordsoort
  - Snelheid
  - Geheugen

7.6.2 State of the art internationaal

De meest gebruikte methode voor morfologische analyse is de zogenaamde *finite-state morphology* [85, 6, 114, 119, 77]. In deze aanpak wordt met behulp van een gefuseerde verzameling van finite-state transducers (waarvan elk een verschillende fonologische regel representeert) en een finite-state lexiconsysteem vertaald van oppervlaktevormen naar lexicale representaties en omgekeerd. De methode is gecommercialiseerd door Xerox, die uitgebreide morfologische analysers hebben gemaakt voor veel talen, waaronder het Nederlands. Het voordeel van lexicale transducers is dat ze *bidirectioneel* (hetzelfde netwerk voor zowel analyse als synthese), *snel* (duizenden woorden per seconde), en *compact* zijn.

Hoewel ook alternatieve methodes worden onderzocht (gebaseerd op inductief leren [133], of methodes als analyse door synthese), zijn finite state transducers de meest gebruikte (en bruikbare) methode.

7.6.3 Inventaris beschikbare systemen Nederlands

- **Samenstellingsplitser Druid**.

  In het kader van het project DRUID wordt gewerkt aan de ontwikkeling van een robuuste spraakherkenner voor het Nederlands ten behoeve van zogenaamde 'spoken document retrieval'. Het gaat om sprekeronafhankelijke, large vocabulary herkenning. De aanpak is gericht op woordherkenning op basis van een taalmodel. Voor het Nederlands is de herkenning van samenstellingen een probleem. Er is een grens aan het aantal woorden dat herkend kan worden (65k) en het aantal unieke woorden dat in principe herkend zou moeten kunnen worden ligt beduidend hoger, ook als de problematiek van de herkenning van eigennamen buiten beschouwing wordt gelaten. In DRUID wordt daarom gebruik gemaakt van een compoundsplitser. De splitser is gebaseerd op verschillende bronnen, waaronder 'harde' taalkundige kennis over waarschijnlijke samenstellingsgrenzen (bijvoorbeeld uitgangen als -schap, -heid,- ing), als op woordenlijsten.

  Algemene eigenschappen: De splitser is beschikbaar in twee varianten: als Perlscript voor gebruik onder Unix en onder Windows. Input is tekst en een woordenlijst van mogelijke samenstellende delen; output is tekst met gesplitste samenstellingen.

  Beschikbaarheid: Voor inlichtingen: Parlevink Human Language Technology Group, Universiteit Twente (email: hltgroup@cs.utwente.nl).

  De splitser is goed integreerbaar en kan getraind worden met verschillende woordenlijsten. Cf. links naar papers
  `http://wwwhome.cs.utwente.nl/~ordelman`.

- **Atranos**.

  Binnen het Atranos-project (`http://atranos.esat.kuleuven.ac.be/`) zijn door het Centrum voor Computerlinguïstiek in Leuven een samenstellingsplitser en een samenstellingbouwer ontwikkeld. De modules zijn geprogrammeerd in Perl en draaien onder Linux. De samensteller werkt op basis van een lexicon van 36.000 woorddelen dat ook is ontwikkeld binnen het Atranos-project.

  Beschikbaarheid: De software kan aangevraagd worden door een email te zenden aan Vincent Vandeghinste (vincent.vandeghinste@ccl.kuleuven.ac.be).

- Nederlandse Porter Stemmer ([86])

  Deze stemmer is ontwikkeld als onderdeel van het UPLIFT project (`http://let.ruu.nl/~uplift/`). Dit programma bepaalt de stam van verbogen woorden en geeft geen informatie over de categorie of morfologische kenmerken van een woord. De (taalkundige) accuratesse van het systeem is beperkt doordat geen woordenboek wordt gebruikt. Het is een aanpassing van de regelgebaseerde Porter stemmer voor het Engels (een multi-stap stemmer zonder uitzonderingenlijst). Uitbreidingen betreffen pre- en infixen, en het systeem kan verbindingsstreepjes en diakritische tekens aan.

  Beschikbaarheid: De source code van de Porter Stemmer is verkrijgbaar via `http://let.uu.nl/~uplift/dstem.tar.gz` onder de GNULIB-licentie.

- Xerox

  Xerox heeft programma's ontwikkeld voor morfologische analyse van diverse Europese talen, waaronder het Nederlands. Het probleem van morfologische analyse wordt opgelost door middel van eindige automaten: (i) de toegestane combinatie van morfemen wordt geëncodeerd in een finite-state netwerk; (ii) de regels die de vorm van ieder morfeem bepalen worden geïmplementeerd als finite-state transducers; (iii) het lexiconnetwerk en de regeltransducers worden gevormd tot een enkele automaat, een lexicale transducer, die alle morfologische informatie van een taal bevat, inclusief derivatie, inflectie en samenstellingen. Een lijst met morfologische categorieën is ook op de webpagina te vinden. Deze analyseerder is onderdeel van de Xerox POS-tagger.

  Beschikbaarheid: Er is een on line demoversie::
  `http://www.xrce.xerox.com/research/mltt/demos/dutch.html`

  Voor commerciële licenties en toepassingen is de Xerox-dochter Inxight verantwoordelijk `http://www.inxight.com/`

- MORANE

  Dit systeem is ontwikkeld door Peter-Arno Coppen van de KU Nijmegen, afdeling computerlinguïstiek [34]. Het is een morfologisch analysesysteem waarbij de eerste slag gebruik maakt van spellingsregels.

- Morfo-Analyzer.

  Morfo-analyzer is ontwikkeld door Rob Heemels aan de KUN, afdeling computerlinguïstiek [67]. Dit is een systeem dat Nederlandse woordvormen analyseert op basis van basisgrootheden, de zgn. MACRO's. Ook de analyse van samenstellingen is in principe opgelost. De beperkende factor is de analysetijd. Elke analyse wordt gerelateerd aan basiswoordvormen in de vier aangehechte dynamische lexica.

- MBLEM and MBMA: Memory-Based Lemmatization and Morphological Analysis

  Deze twee systemen zijn ontwikkeld bij de ILK Research Group van de KUB. MBLEM is een lemmatizer voor het Nederlands, Engels, en Duits. Het systeem converteert vervoegde woordvormen naar hun lemma's (voor naamwoorden het enkelvoud, voor werkwoorden de infinitiefvorm). MBMA analyseert de morfologie van Nederlandse woorden (zie [133]). MBMA voert classificatiegebaseerde segmentatie en spellingsveranderingen

uit, vindt de inflectionele kenmerken en kent de morfologische klasse toe. De systemen zijn getraind op CELEX-data.

Een on line demoversie is te vinden bij
`http://ilk.kub.nl/demos.html`.

Beschikbaarheid: Licentie.

- SPIRIT (ASCII - Fi Systems Belgium - Gerrit Potoms)

  Één van de componenten van SPIRIT is een morfologische analyser: woorden worden geïdentificeerd en teruggebracht tot hun basisvorm. Ze gebruiken hiervoor een morfologisch woordenboek (CELEX).

  Beschikbaarheid: SPIRIT is commercieel verkrijgbaar voor het Nederlands in een 'limited edition'. De full edition wordt verwacht in 2001.

- Lingbench (`http://www.natlantech.com`)

  Een onderdeel van de Natlantech-technologie is een morfologische analyseerder, alsmede een parser. Dit zijn taalonafhankelijke software modules die reeds in hoge mate functioneren.

  Op het ogenblik van schrijven is Natlantech NV failliet. Er worden echter druk pogingen gedaan om de technologie verder te zetten. Het lag in de bedoeling om ook voor het Nederlands te ontwikkelen.

- Euroglot

  Euroglot is een meertalig vertaalsysteem dat beschikbaar is in de talen Nederlands, Engels, Frans, Duits, Spaans en Italiaans. Onderdeel hiervan is een morfologiemodule. In de morfologiemodule worden niet alleen alle vervoegingen van werkwoorden en verbuigingen van alle zelfstandige en bijvoeglijke naamwoorden gegenereerd maar ook herkend.

  Euroglot is te vinden via `http://www.euroglot.nl/nl/default.htm`, waar ook prijslijsten raadpleegbaar zijn.

Hieronder volgen nog twee applicaties die toestaan een analyseerder te maken voor het Nederlands.

- Mmorph

  Mmorph is een applicatie die onderdeel uitmaakt van MULTEXT (zie ook 7.4.1). Met Mmorph is het mogelijk om lexica te maken en te modificeren, en om teksten te annoteren met lexicale informatie. Het programma koppelt woordvormen zoals gevonden in een tekst, aan een 'entry' in een lexicale database, die arbitraire informatie bevat uitgedrukt in attributen en waarden. De lexicale database wordt gecreëerd vanuit een set van initiële lexicale 'entries' en een set van structurele regels.

  Beschikbaarheid: Het programma Mmorph en de documentatie zijn vrij te downloaden vanaf `http://issco-www.unige.ch/projects/MULTEXT.html`.

- LEXA: Corpus Processing Software

  De Lexa suite is ontwikkeld door Raymond Hickey aan de Universiteit van Essen, Duitsland. Lexa bestaat uit zes verschillende pakketten. Het eerste pakket is ook het

belangrijkste. Het zorgt voor de lexicale analyse en lemmatisering. LEXA laat toe om elke ASCII-tekst automatisch te annoteren en te lemmatiseren, om frequentielijsten aan te leggen van de types en tokens die voorkomen in de ingeladen tekst, om lexicale dichtheidstabellen op te stellen of om tekstuele data om te zetten in een database-omgeving. De resultaten van al deze bewerkingen kunnen als elektronische bestanden worden opgeslagen en later worden geraadpleegd.

Beschikbaarheid: Lexa kan vrij gedownload worden van de volgende site: `http://www.hit.uib.no/lexainf.html`. De suite wordt geïnstalleerd onder Windows.

7.6.4 Evaluatie en conclusie

Hoewel er voor het Nederlands verschillende systemen voor morfologische analyse bestaan, lijken er geen vrij beschikbaar te zijn (met uitzondering van de Nederlandse Porter Stemmer). Het lijkt evenwel niet uitgesloten dat betaalbare licenties op systemen voor morfologische analyse beschikbaar zijn. Morfologische analyse is alleszins een uitvoerig onderzocht probleem waarvoor goede oplossingen beschikbaar zijn.

7.7 Morfosyntactische disambiguering (POS tagging)

Morfosyntactische disambiguering (POS tagging) is de fruitvlieg van de computertaalkunde: het is een prototypisch probleem voor taaltechnologische toepassingen (selectie in context). Gegeven een woord en de context waarin het voorkomt, moet de contextueel correcte woordsoort van het woord worden bepaald. Dit is geen eenvoudig probleem: hoewel voor een typische inventaris van woordsoorten slechts een 10% van de woordvormen (types) ambigu is, kan dat oplopen tot 50% voor de woordvorm *tokens*.

Een POS tagger veronderstelt een lexicon waarin voor elk woord de mogelijke woordsoorten te vinden zijn (eventueel met hun waarschijnlijkheid), een beslissingsmodule die de disambiguering doet en een component die een woordsoort gokt voor woorden die niet in het lexicon staan. Een uitvoerige inleiding tot dit probleem en de verschillende benaderingen voor het oplossen ervan, is te vinden in [138]. Output van een tagger is voor elk woord de contextueel correcte woordsoort, of in sommige systemen, een klein aantal overblijvende mogelijke woordsoorten.

POS tagging is nuttig als eerste stap in een meer volledige syntactische analyse van tekst, of als een alternatief voor volledige syntactische analyse in verschillende toepassingen: information retrieval en informatie-extractie, verwerving van terminologie, spraaksynthese, spelling-correctie enz.

Zoals bij de meeste modules zijn zowel deductieve als inductieve benaderingen geprobeerd bij de ontwikkeling van POS taggers. Verschillende statistische of machine learning methodes zijn toegepast op dit probleem en hebben aanleiding gegeven tot het beschikbaar maken van tagger-generatoren: software die getraind kan worden op een met woordsoorten geannoteerd corpus, en dan een tagger oplevert die nieuwe tekst volgens dezelfde systematiek kan analyseren. Deze tagger-generatoren zijn uiteraard taalonafhankelijk en kunnen dus gebruikt worden voor de constructie van taggers voor het Nederlands. In de deductieve benadering valt vooral de *constraint grammar* benadering van [76] op, een methode gebaseerd op eliminatie van contextueel onmogelijke woordsoorten met behulp van met de hand gemaakte regels.

7.7.1 Specifieke evaluatiecriteria

We geven de volgende interpretatie aan de eerder vermelde algemene criteria en voegen er de specifieke criteria en eigenschappen aan toe in onze evaluatie en vergelijking van POS taggers voor het Nederlands. Voor elk van de geïnventariseerde taggers worden zoveel mogelijk van deze criteria onderzocht.

- Algemene eigenschappen
  - Programmeertaal
  - Besturingssysteem
  - Ontwikkelmethode (inductief/deductief)
  - Gebruikte tag set
  - Output (1 tag per woord, beperkte ambiguïteit, probabiliteiten)
  - Voor welk teksttype
  - Componenten (tokenisering, lexicon, disambiguator, module voor onbekende woorden, idiomenmodule)
- Beschikbaarheid
- Bruikbaarheid
  - Integreerbaarheid/modulariteit
  - Aanpasbaarheid (nieuw teksttype, tag set)
- Documentatie
- Kwaliteit
  - Accuraatheid
  - Precision en recall per woordsoort
  - Snelheid
  - Geheugen

7.7.2 State of the art internationaal

De accuraatheid van POS taggers hangt niet alleen af van de kwaliteit van de gebruikte methode, maar ook van het type van de taal (bijv. Slavische talen hebben een meer complexe morfosyntaxis dan Germaanse talen), en van de gekozen inventaris van woordsoorten; zowel omvang als contextuele disambigueerbaarheid ervan. Een grote tagset is niet noodzakelijk moeilijk te disambigueren of omgekeerd. Hiermee moet rekening worden gehouden bij het vergelijken van de resultaten. Dit fenomeen maakt het eveneens moeilijk om iets algemeens te zeggen over de state of the art van POS tagging. Voor Engelse nieuwstekst (Wall Street Journal) en met een middelmatig gedetailleerde tag set (Penn Treebank tag set) is 97% accuraatheid haalbaar (60% op zinsniveau), met hoge efficiëntie (tienduizenden woorden per seconde). Hoewel beweerd wordt door de voorstanders ervan dat handgemaakte regelgebaseerde *constraint grammars* superieure accuraatheid hebben ten opzichte van inductieve

methodes (statistisch of gebaseerd op leertechnieken), is dit nooit duidelijk aangetoond, en wordt in de meeste projecten gebruik gemaakt van deze inductieve technieken, getraind op (semi-automatisch) geannoteerde corpora, met redelijk hoge accuraatheid.

Toch is POS tagging een verre van opgelost probleem. Bij de overblijvende 3+% fouten gaat het vaak om cruciale ambiguïteiten, syntactische constructies waarvoor meer dan lokale context nodig is om ze op te lossen, onbekende woorden en semantische ambiguïteiten. De accuraatheid van taggers is soms bedroevend klein wanneer ze worden getest op andere tekstsoorten dan diegene waarop ze werden getraind of ontworpen. Wanneer wordt gewerkt met lerende taggers bevat het leermateriaal vaak veel fouten of is de annotatie inconsistent over verschillende annotatoren. Toevoeging van meer informatie in de input van lerende taggers heeft een exponentieel effect op de *sparseness* van de data, een belangrijk probleem voor lerende systemen, en vele talen hebben een erg uitgebreide tag set nodig en hebben geen of weinig geannoteerd materiaal beschikbaar.

7.7.3 Inventaris beschikbare taggers Nederlands

Dit overzicht is gebaseerd op [146], uitgebreid met informatie die we hebben gekregen tijdens het huidige project. We maken een onderverdeling in lerende en met de hand geconstrueerde taggers, en starten met een globale beschrijving. In de volgende sectie worden deze taggers geëvalueerd aan de hand van de hoger genoemde criteria.

- *Handgemaakte taggers.* Soms gecombineerd met statistische componenten.
  - D-Tale tagger.
    De *Dutch TAgger-LEmmatizer* voor het Nederlands is een programma dat werd ontwikkeld door de afdeling lexicologie aan de VU, voor Van Dale. Het is een regelgebaseerde tagger/lemmatizer met een groot woordvormenlexicon (140.000 woordvormen) en een vaste eigen tagset.

- *Getrainde taggers.* Deze tagger-generatoren hebben een geannoteerd trainingscorpus nodig. Voor het Nederlands zijn een aantal van deze corpora beschikbaar (zie Sectie 8.3.3). Het meeste onderzoek naar lerende taggers voor het Nederlands is gebeurd op het Eindhoven corpus met de WOTAN tag set.
  - Gebaseerd op n-gram / Hidden Markov Model technieken. In deze methodes wordt op basis van de frequentie van n-grammen van woordsoorten (meestal bigrammen of trigrammen) in een geannoteerd corpus en op basis van de lexicale waarschijnlijkheid van woord-woordsoortparen, per zin berekend welke reeks woordsoorten het meest waarschijnlijk is. De implementatie hiervan gebeurt in een probabilistische finite state automaat (hmm) en met algemene zoekmethodes (varianten van dynamic programming).
    * Xerox tagger
      Xerox Research Center Europe heeft trigram-gebaseerde taggers ontwikkeld voor een groot aantal Europese talen, waaronder ook het Nederlands. Tagging gebeurt in 3 stappen: tokenisatie, lexicale opzoeking (waarbij morfologische analyse een onderdeel is) en disambiguatie via HMM-techniek.
      `http://www.xrce.xerox.com/research/mltt/fsnlp/tagger.html`

    * KEPER tagger
      De KEPER tagger is een bigram-tagger gecombineerd met morfologische analyse, ontwikkeld door Polderland BV. KEPER werkt met een eigen tagset, die ontwikkeld is met het oog op Information Retrieval toepassingen.
      `http://www.polderland.nl`
    * CORRie tagger
      Deze tagger generator, ontwikkeld door Theo Vosse in Leiden, is een standaard trigram HMM tagger, plus "een poging om structuur te herkennen aan de hand van functiewoorden". Onbekende woorden worden gegokt op basis van het einde van het woord. De tagger is gebaseerd op een kleine tagset van hoofdcategorieën en is getraind op één miljoen woorden aan handmatig door het INL geannoteerde krantenteksten.
    * WOTAN tagger
      Deze tagger (generator), gebaseerd op Hidden Markov Modellen (HMM) en een memory-based module voor onbekende woorden, is ontwikkeld bij de afdeling Taal en Spraak van de KUN [8]. Op het moment is deze tagger getraind op het Eindhoven corpus en geannoteerd met de WOTAN-I tagset. Een nieuwe versie, gebaseerd op de uitgebreidere WOTAN-II tagset is in de maak.
    * PAROLE tagger
      Het INL beschikt over een Nederlandse tagger Dutchtale, gebaseerd op een lexicon, morfologische analyse, en een hybride van een regelgebaseerde en statistische disambigueringscomponent. Op het moment wordt er een modernere opvolger op basis van de PAROLE tagset en HMM-technnieken ontwikkeld. Deze tagger is in een vergevorderd stadium.
      `http://www.inl.nl/`
    * TnT (Trigrams 'n Tags)
      TnT is een trigram HMM tagger, ontwikkeld door Thorsten Brants van de Universiteit van Saarbrücken [26]. De software is een onderdeel van het Annotate-platform voor corpusannotatie. De tagger maakt gebruik van lineaire interpolatie voor smoothing van kansen en van suffixen van woorden om onbekende woorden te gokken. Deze tagger is in Tilburg getraind op met WOTAN geannoteerd materiaal.
      `http://www.coli.uni-sb.de/~thorsten/tnt/`
    * Van Dale tagger
      Van Dale Data heeft een standaard HMM tagger ontwikkeld in Perl met bijkomende regels om veel voorkomende fouten te elimineren. De tagger maakt gebruik van een groot lexicon (database van Van Dale Data) en is getraind op het Persdatacorpus (meer dan 80 miljoen woorden).

– Gebaseerd op andere statistische leertechnieken.

    * Brill tagger
      Deze tagger-generator, gemaakt door Eric Brill (Johns Hopkins, nu Microsoft) leert transformatieregels voor de omzetting van de output van een initiële tagger naar een correctere output [27]. De regels worden geleerd op basis van discrepanties tussen een door de initiële tagger getagd corpus en een 'gouden standaard' en de ruimte van mogelijke regels wordt beperkt door vooraf gedefinieerde regeltemplaten. Zowel in Groningen als in Tilburg is deze

tagger reeds getraind op Nederlands materiaal.
`http://www.cs.jhu.edu/~brill/`

* Memory-Based tagger
  De MBT tagger-generator, ontwikkeld door de ILK groep aan de KUB en UIA [42], werkt op basis van Memory-Based Learning, analogisch redeneren met in het geheugen opgeslagen voorbeelden. Deze tagger is eenvoudig hertrainbaar op een nieuw corpus, en is momenteel beschikbaar met de WOTAN-I tagset (getraind op 600 duizend woorden Eindhoven corpus) en WOTAN-II (getraind op 150 duizend woorden).
* MXPOST (Maximum Entropy tagger)
  Deze tagger (generator), gemaakt door Adwait Ratnaparkhi (UPenn, nu IBM), is in Tilburg getraind op WOTAN-I en WOTAN-II materiaal. De tagger is eenvoudig trainbaar op nieuwe corpora, en werkt zonder expliciet lexicon. Onbekende woorden worden gegokt op basis van een aantal vorm-features.
  `http://www.cis.upenn.edu/~adwait/statnlp.html`
* CGN Tagger Fabriek
  Ten behoeve van de POS tagging van het Corpus Gesproken Nederlands (CGN) is een 'tagger-fabriek' opgezet in Tilburg [56], op basis van de bevindingen in [146], waarin verschillende lerende methodes (MXPOST, MBT, TnT, Brill) en verschillende informatiebronnen (lexicons en taggers getraind op ander materiaal dan CGN) worden gecombineerd met tweede-niveau leermethodes (stacked learning) om de automatische tagging van het corpus met weinig met de hand gecorrigeerde data op te starten met redelijke accuraatheid. Deze complexe leermethode wordt herhaald met telkens meer data tot op een bepaald moment de beste individuele lerende methode (de verwachting is dat dat TnT zal zijn) het niet veel slechter meer doet dan de combinatiemethode.

7.7.4 Evaluatie

Relatief weinig methodologisch correct opgestelde vergelijkende en evaluatieve studies zijn beschikbaar voor POS tagging voor het Nederlands. In [146] wordt een evaluatie van taggers ondernomen met als doel de selectie van een tagger voor het CGN corpus. Een andere vergelijking van tagger generatoren is te vinden in [139] waar een aantal tagger-generatoren werd getest, en TnT als beste naar voor kwam met 92-95% voor twee versies van de WOTAN-I tagset op het Eindhoven Corpus. Voor de alsnog beperkte CGN corpusdata en de CGN tagset [55] levert TnT ook de beste resultaten op (91% met 20,000 woorden training data). In deze laatste experimenten doen combinatiemethodes het echter nog steeds beter dan TnT (foutenreducties tot 30% ten opzichte van TnT).

D-Tale tagger .

- *Algemene eigenschappen*
  - *Programmeertaal*: C
  - *Besturingssysteem*: UNIX
  - *Ontwikkelmethode*: deductief

  - *Gebruikte tag set*: vaste eigen
  - *Output*: meer dan 1 tag in geval van ambiguïteit
  - *Teksttype*: kranten, stukken wetenschappelijk proza, en scripts van soapseries
  - *Componenten*: lexicon, disambigueerder

- *Beschikbaarheid*: ?

- *Bruikbaarheid*

  - *Integreerbaarheid/modulariteit*: ja
  - *Aanpasbaarheid*: Niet opnieuw trainbaar, aangezien de disambiguatieregels met de hand zijn opgesteld (In een Constraint Grammar formalisme à la [76]), maar is eventueel wel handmatig aan te passen. Het lexicon is uitbreidbaar, en kan ook multi-word units bevatten.

- *Documentatie*

- *Kwaliteit*

  - *Accuraatheid*: 93% (82.4% op CGN-tagset)
  - *Precision en recall per woordsoort*
  - *Snelheid*: 'niet al te snel'
  - *Geheugen*

PAROLE tagger

- *Algemene eigenschappen*

  - *Programmeertaal*: Perl en C++
  - *Besturingssysteem*: Solaris
  - *Ontwikkelmethode*: inductief
  - *Gebruikte tag set*: PAROLE tagset: POS met features (zie `http://www.inl.nl`)
  - *Output*: 1 tag met features per token
  - *Teksttype*: PAROLE-corpus (20 mln woorden): gevarieerde samenstelling, geschreven Nederlandse tekst, vnl. vroege jaren 1990 (tot 1996)
  - *Componenten*: tokenisering, lexicon, disambigueerder, onbekende woorden module, geen idiomenmodule

- *Beschikbaarheid*: Omdat het een combinatietagger is met gebruik van software van anderen (conform hun voorwaarden), bijv. TnT en Timbl, kan hierover niet beslist worden zonder die anderen.

- *Bruikbaarheid*

  - *Integreerbaarheid/modulariteit*: laag

  - *Aanpasbaarheid*: afzonderlijke taggers in principe aanpasbaar, het geheel d.w.z. inclusief de lemmatiseerder is dat niet zonder meer, omdat gebruik gemaakt wordt van een lexicon met de PAROLE tagset.

- *Documentatie*: nog niet want nog in ontwikkeling
- *Kwaliteit*
  - *Accuraatheid*: nog niet bekend
  - *Precision en recall per woordsoort*: nog niet bekend
  - *Snelheid*: trage initialisatie, daarna ruim 500 tokens/seconde, inclusief het lemmatiseren
  - *Geheugen*: 40 Mb, inclusief lemmatiseren

Xerox tagger

- *Algemene eigenschappen*
  - *Programmeertaal*: XeLDA client/server platform
  - *Besturingssysteem*: UNIX en Windows
  - *Ontwikkelmethode*: trigram gebaseerd
  - *Gebruikte tag set*: Xerox tagset (49 elementen) en te bekijken bij `http://www.xrce.xerox.com/research/mltt/demos/doc/pos-dut-1.txt`
  - *Output*: 1 tag
  - *Teksttype*
  - *Componenten*: tokenisering, lemmatisering, lexicon
- *Beschikbaarheid*: De Nederlandse tagger is beschikbaar onder een onderzoekslicentie voor een eenmalig bedrag van US $ 2000 voor één jaar, gratis verlengbaar. Updates en nieuwe versies zullen opnieuw betaald moeten worden. De tagger is in principe hertrainbaar. De tool die hiervoor en voor incrementele aanpassing aan het lexicon benodigd is, valt echter buiten de genoemde onderzoekslicentie, tenzij er een samenwerkingsovereenkomst met Xerox wordt afgesloten. Op `http://www.rxrc.xerox.com/research/mltt/demos/dutch.html` is een on line demoversie te vinden.
- *Bruikbaarheid*
  - *Integreerbaarheid/modulariteit*
  - *Aanpasbaarheid*: ja, maar in samenwerking met XEROX.
- *Documentatie*
- *Kwaliteit*
  - *Accuraatheid*: 78.8% op CGN-tagset
  - *Precision en recall per woordsoort*
  - *Snelheid*: duizenden woorden per seconde voor tagging en enkele seconden voor hertrainen
  - *Geheugen*

KEPER tagger

- *Algemene eigenschappen*
  - *Programmeertaal*
  - *Besturingssysteem*: Windows en diverse UNIX-varianten en is op aanvraag ook voor andere platforms geschikt te maken.
  - *Ontwikkelmethode*: Bigram-tagger gecombineerd met morfologische analyse.
  - *Gebruikte tag set*: Eigen tagset.
    Het systeem is hertrainbaar op een nieuw corpus met een andere tagset, maar dit zou door Polderland uitgevoerd moeten worden.
  - *Output*: 1 tag per woord, tenzij disambiguering wordt uitgeschakeld, dan geeft hij alle mogelijke tags terug.
  - *Teksttype*
  - *Componenten*: Tokenisatie, lemmatisatie, lexicon
- *Beschikbaarheid*: Deze tagger is beschikbaar voor CGN voor een vaste prijs van Hfl. 18000,- en aanpassingen tegen uurtarief.
- *Bruikbaarheid*
  - *Integreerbaarheid/modulariteit*
  - *Aanpasbaarheid*: De training van de statistische parameters kan incrementeel gebeuren, maar het tagging-lexicon is vast geïntegreerd in het systeem. Wel kan de gebruiker met een aparte lexiconfaciliteit een eigen woordenlijst toevoegen, die het systeemlexicon kan 'overrulen'. Het trainen van de tagger op een nieuw corpus gebeurt onder supervisie van een menselijke operator.
- *Documentatie*
- *Kwaliteit*
  - *Accuraatheid*: 73.7% op CGN-tagset
  - *Precision en recall per woordsoort*
  - *Snelheid*: 45 minuten voor 1 Mbyte tekst op een SUN SPARCstation uit 1992
  - *Geheugen*

Over de trainingssnelheid meldt Polderland: "Bij begin van de training ligt het tempo rond de 5 woorden per seconde. Het loopt snel op. Later neemt de snelheid geleidelijk minder toe en convergeert tot het eindtempo dat een veelvoud van het begintempo is. Trainingsmateriaal kan hergebruikt worden, waarbij het tempo bijna zo hoog is als voor de gewone productiesnelheid. Aangezien KEPER nog niet is getraind op andere soorten tekst heeft Polderland nog weinig ervaring met het trainingstempo. De trainingsfaciliteit is eenvoudig te bedienen."

Het outputformaat is aan te passen met behulp van een style file. Keper maakt optioneel gebruik van een gecustomiseerde preprocessor. De standaard input is gewone ASCII.

CORRie tagger

- *Algemene eigenschappen*
  - *Programmeertaal*: C++
  - *Besturingssysteem*: Onafhankelijk
  - *Ontwikkelmethode*: Trigram HMM tagger
  - *Gebruikte tagset*: De gebruikte tagset bevat hoofdcategorieën plus simpele morfologische informatie van werkwoordsmorfologie plus een klasse per functiewoord/categorie (dus bv. 'het' zit in de klassen 'het/pvnw' en "het/lidw").
  - *Output*: De tagger kan de N-beste tags per woord teruggeven, voorzien van hun waarschijnlijkheid.
  - *Teksttype*: Krantenteksten
  - *Componenten*: (corpusafhankelijke) zinssplitser en tokenizer, woordenboekmodule
- *Beschikbaarheid*: De software is in principe beschikbaar voor CGN, al zou er nog overlegd moeten worden over de precieze condities, aangezien de trainingscorpora eigendom zijn van het INL.
- *Bruikbaarheid*
  - *Integreerbaarheid/modulariteit*: De software is portable naar allerlei platformen en het is in principe mogelijk deze aan te passen aan specifieke vereisten van de annoteeromgeving.
  - *Aanpasbaarheid*: De tagger kan eenvoudig getraind worden op een nieuw corpus met een andere tagset.
- *Documentatie*
- *Kwaliteit*
  - *Accuraatheid*: 86.7% op CGN-tagset
  - *Precision en recall per woordsoort*
  - *Snelheid*: Onbekend.
  - *Geheugen*: Een paar MB (afhankelijk van de grootte van de tagset en trainingsdata); als het lexicon in het geheugen geladen moet worden (voor grotere snelheid), kost dat natuurlijk veel meer geheugen.

Memory-Based tagger

De tagger tokeniseert niet, en lemmatiseert niet, maar er is een tokenisatie-preprocessor en een aparte Memory-Based morfologische analyse (MBMA) module beschikbaar (zie `http://ilk.kub.nl/` voor een demo) die extrapoleert vanuit CELEX. Onbekende woorden worden gegokt op basis van een aantal vorm-features (prefix, hoofdletter?,suffix, getallen?, hyphen?).

Er wordt gewerkt aan een versie die SGML-invoer en -uitvoer aankan.

- *Algemene eigenschappen*

  - *Programmeertaal*: C++ code en Perlscripts
  - *Besturingssysteem*: Verschillende UNIX-varianten
  - *Ontwikkelmethode*: Inductief
  - *Gebruikte tag set*: WOTAN-I en WOTAN-II
  - *Output*: Meerdere tags per woord teruggeven, vergezeld van een zekerheidsmaat.
  - *Teksttype*: Hertrainbaar
  - *Componenten*: Lexicon
- *Beschikbaarheid*
- *Bruikbaarheid*
  - *Integreerbaarheid/modulariteit*: Gemakkelijk te porten naar andere platforms (C++ code en Perl-scripts)
  - *Aanpasbaarheid*: Het lexicon is incrementeel uitbreidbaar.
- *Documentatie*
- *Kwaliteit*
  - *Accuraatheid*: 96% (MBT-WOTAN-I: 87.7% op CGN tagset; MBT-WOTAN-II: 82.9% op CGN-tagset: bij een automatische vertaalslag van WOTAN-tags naar CGN-tags.)
  - *Precision en Recall per woordsoort*
  - *Snelheid*: Het trainen duurt enkele minuten, en de snelheid bij het taggen is meer dan twintigduizend woorden per seconde op een Pentium II PC.
  - *Geheugen*

MXPOST (Maximum Entropy tagger)

In een vergelijk van vier bekende taggertechnieken (van Halteren et al., 1998) bleek deze tagger steeds significant accurater dan de andere (Brill, HMM, MBT). Recentere experimenten laten echter zien dat TnT (zie hieronder) vaak net iets accurater is.

- *Algemene eigenschappen*
  - *Programmeertaal*: Voorgecompileerde Java bytecode
  - *Besturingssysteem*: Platformonafhankelijk
  - *Ontwikkelmethode*: Inductief
  - *Gebruikte tag set*: WOTAN-I en WOTAN-II
  - *Output*
  - *Teksttype*: Hertrainbaar
  - *Componenten*

- *Beschikbaarheid*: De software is voor onderzoeksdoeleinden gratis te downloaden van het internet
- *Bruikbaarheid*
  - *Integreerbaarheid/modulariteit*
  - *Aanpasbaarheid*
- *Documentatie*
- *Kwaliteit*
  - *Accuraatheid*: MX-WOTAN-I: 86.9% op CGN-tagset; MX-WOTAN-II: 81.8% op CGN-tagset (automatische vertaalslag)
  - *Precision en recall per woordsoort*
  - *Snelheid*: Het trainen van de tagger duurt voor een groot corpus enkele dagen. Taggen duurt ongeveer een seconde per zin van 30 woorden op een Solaris Pentium II PC.
  - *Geheugen*

TnT (Trigrams 'n Tags)

- *Algemene eigenschappen*
  - *Programmeertaal*: Solaris ANSI C
  - *Besturingssysteem*: Unix/Linux
  - *Ontwikkelmethode*: inductief
  - *Gebruikte tag set*: WOTAN
  - *Output*: 1 tag per woord of meerdere tags samen met een kansverdeling
  - *Teksttype*: Hertrainbaar
  - *Componenten*
- *Beschikbaarheid*: licentie is gratis voor niet-commercieel gebruik.
- *Bruikbaarheid*
  - *Integreerbaarheid/modulariteit*
  - *Aanpasbaarheid*: ja
- *Documentatie*: on line (`http://www.coli.uni-sb.de/~tnt/`)
- *Kwaliteit*
  - *Accuraatheid*: TnT-WOTAN-I: 89.9% op CGN tagset; TnT-WOTAN-II: 83.9% op CGN-tagset (automatische vertaalslag)
  - *Precision en recall per woordsoort*
  - *Snelheid*: Het trainen van de tagger is een kwestie van enkele seconden en taggen gaat met enkele duizenden woorden per seconde.
  - *Geheugen*

Van Dale Data HMM tagger

- *Algemene eigenschappen*
  - *Programmeertaal*: Perl
  - *Besturingssysteem*: UNIX en Windows
  - *Ontwikkelmethode*: inductief/deductief
  - *Gebruikte tag set*: vaste eigen
  - *Output*
  - *Teksttype*: persdata
  - *Componenten*: standaard HMM aanpak en foutcorrectieregels
- *Beschikbaarheid*: ?
- *Bruikbaarheid*
  - *Integreerbaarheid/modulariteit*: ja
  - *Aanpasbaarheid*: Hertrainbaar
- *Documentatie*
- *Kwaliteit*
  - *Accuraatheid*: 96%
  - *Precision en recall per woordsoort*
  - *Snelheid*: redelijk snel
  - *Geheugen*

7.7.5 Conclusie

POS tagging is een breed toepasbare lingware-module die met redelijke accuraatheid beschikbaar is. Dat impliceert niet dat het een opgelost probleem is, er is nog een lange onderzoeksweg te gaan voor POS taggers beschikbaar komen met de noodzakelijke accuraatheid en flexibiliteit (bijv. aanpasbaarheid aan verschillende teksttypes en POS tag sets). Het is opvallend dat er zelfs voor onderzoek weinig of geen publiek beschikbare taggers zijn voor het Nederlands, er is helemaal niets in het publieke domein voor commercieel gebruik. Het is evenwel mogelijk taggers te kopen voor redelijke bedragen, of er zelf een te maken met behulp van inductieve technieken (pakketten die deze technieken implementeren zijn evenwel ook niet gratis).

We adviseren niet om van de constructie van een tagger voor het Nederlands een topprioriteit te maken, hoewel deze lingware een centrale plaats in de BATAVO verdient. Wanneer werk wordt gemaakt van een goed uitgewerkte TREEBANK voor het geschreven Nederlands, zal in combinatie met het Corpus Gesproken Nederlands voldoende trainingsmateriaal aanwezig zijn om degelijke open source taggers te maken met behulp van inductieve technieken. Vooral belangrijk lijkt de ontwikkeling van afbeeldingen tussen verschillende gangbare tagsets om zoveel mogelijk gebruik te kunnen maken van bestaand geannoteerd trainingsmateriaal.

7.8 Syntactische analyse

In deze sectie behandelen we modules die gericht zijn op syntactische analyse van tekst, d.w.z. het herkennen en benoemen van zinsdelen en de relaties tussen zinsdelen.

Syntactische analyse is van belang voor toepassingen waarbij de structuur van zinnen een rol speelt. Te denken valt aan de automatische identificatie en correctie van grammaticale fouten in documenten (zoals het foutief gebruik van d of t op het eind van een woord), dialoogsystemen (waarbij bijvoorbeeld vragen van een gebruiker moeten worden geanalyseerd en begrepen), en automatische vertaling (waarbij syntactische analyse van de originele tekst de basis vormt voor een grammaticaal correcte vertaling).

Andere toepassingen waar syntactische analyse tot op zekere hoogte een rol speelt zijn document retrieval (het vinden van (web-) documenten die relevant zijn voor een bepaalde zoekvraag), information extraction (het extraheren van delen van documenten die speciaal relevant zijn voor een zoekvraag), en question answering (het extraheren van delen van teksten die een zoekvraag daadwerkelijk beantwoorden), en tekstclassificatie (het automatisch classificeren van documenten (zoals e-mails) op onderwerp). Veel van het onderzoek naar deze onderwerpen beschouwt documenten slechts als verzamelingen losse woorden. Dit gezichtspunt betekent dat er soms belangrijke informatie wordt genegeerd. Wanneer men wil weten wie de toekomstige koningin van Nederland is, is het belangrijk te kunnen zoeken naar documenten waarin deze woorden niet alleen los maar ook in een bepaald verband ('toekomstige koningin van Nederland') voorkomen. Het automatisch herkennen van dergelijke phrases in een tekst kan de effectiviteit van document retrieval en verwante taken bevorderen, maar vereist wel dat dergelijke reeksen woorden als syntactische eenheden herkend kunnen worden.

Een toepassing waar syntactische analyse een rol begint te spelen is spraakherkenning. De taalmodellen die worden gebruikt om te voorspellen wat het volgende woord in een uiting is zijn meestal gebaseerd op n-grammen. Een taalkundig meer aansprekend, en in theorie accurater, alternatief maakt ook gebruik van grammaticale kennis.

Syntactische analyse kan gericht zijn op het analyseren van volledige zinnen of uitingen, met als doel alle zinsdelen te benoemen en met elkaar in verband te brengen. Omdat dit moeilijk is, zijn er ook systemen ontwikkeld die slechts bepaalde soorten zinsdelen (typisch naamwoordelijke constituenten en voorzetsel constituenten) proberen te herkennen en ontleden. De laatste vorm van ontleden noemen we wel *partial parsing* of *chunking* [4]. Het herkennen van relatief eenvoudige losse zinsdelen is verwant aan taken als *named entity recognition* (sectie 6.4.2).

Automatische syntactische analyse maakt over het algemeen gebruik van een grammatica en een parser. De grammatica bevat een definitie van de syntactische regels van de taal en van het woordenboek. De parser heeft als doel zinnen te ontleden volgens de regels van de grammatica. Wanneer er voor een zin meerdere analyses mogelijk zijn, is het van belang een uitspraak te doen over wat de beste of meest waarschijnlijke analyse is. Dit noemen we (syntactische) disambiguatie.

Het opstellen en implementeren van grammatica's die een redelijk groot deel van de syntaxis van een taal afdekken is moeilijk en arbeidsintensief. Met name voor het Engels zijn ook systemen ontwikkeld die gebruik maken van een grammatica die automatisch is afgeleid uit een zogenaamde *treebank*. Een treebank is een corpus geannoteerd met syntactische informatie (m.n. constituentstructuur) (zie sectie 8). Een treebank bevat impliciet een definitie van een grammatica. Bovendien valt uit een treebank af te leiden welke regels vaak en welke regels minder vaak worden gebruikt. Dit helpt om het disambiguatieprobleem op te lossen,

doordat b.v. aan analyses met frequente regels de voorkeur kan worden gegeven. Tenslotte is een treebank nuttig voor evaluatie, omdat de resultaten van automatische analyse kunnen worden vergeleken met de annotatie in het corpus.

Syntactische analyse kan worden geholpen door de woorden in de te analyseren zin reeds te voorzien van woordsoorten. Sommige systemen zijn zo ontworpen dat ze vereisen dat de input voorzien is van woordsoorten (en dus eerst door een POS tagger is geleid, zie sectie 6.7).

7.8.1 Specifieke evaluatiecriteria

- Specifieke Eigenschappen
  - Taalkundig kader
  - Eisen aan de input (met of zonder POS tags, alleen woorden uit een bepaalde woordenlijst of niet)
  - Volledige of partiële analyse?
  - Omvang van het woordenboek
  - disambiguatie (aan- of afwezig, op basis van data-georiënteerde, statistische methoden?)
- Algemene eigenschappen
  - Programmeertaal en besturingssysteem
  - Beschikbaarheid
  - Bruikbaarheid (Integreerbaar en Modulair, Aanpasbaar (teksttype, outputformaat))
- Documentatie
- Kwaliteit
  - Zijn er gegevens bekend over accuraatheid met betrekking tot tests zoals *(labelled, non-crossing) bracketing* (d.w.z. het bepalen van de 'haakjes'-structuur aan een zin, gezien als een grove maat voor constituentstructuur), of *dependency relations* (d.w.z. het bepalen van grammaticale relaties tussen woorden en constituenten)?
  - Snelheid en geheugen

7.8.2 State of the art internationaal

Er zijn voor het Engels een aantal systemen beschikbaar die gebruik maken van een handmatig opgestelde grammatica. Bekende voorbeelden zijn de SRI *Core Language Engine* (CLE) en de (publiek beschikbare) XTAG grammatica op basis van *Tree Adjoining Grammar*. CLE is onder andere gebruikt in (gesproken) dialoogsystemen en automatische vertaalsystemen.

Het Wall Street Journal corpus is de basis van een aantal systemen die gebruikmaken van een statistische grammatica, automatisch afgeleid uit de treebank. Omdat al deze systemen gebruik maken van dezelfde syntactische analyses, is vergelijking en evaluatie mogelijk. Er kan ongeveer 88% recall en precision behaald worden voor het herkennen van zinsdelen inclusief categorie *labelled bracketing*.

Een bekende taak waarbij slechts een partiële analyse wordt uitgevoerd, is het herkennen van nominale constituenten (NPs) in tekst. Op basis van het WSJ-corpus is hiernaar ook uitvoerig onderzoek gedaan, met als beste resultaten een precision en recall van rond de 93%.

Voor het Engels zijn ook grammatica's beschikbaar die zich richten op het identificeren van grammaticale (of dependentie-) relaties. Zulke relaties zijn bijvoorbeeld expliciet beschikbaar in het Suzanne-corpus. De beste systemen geven een precision en recall van rond de 88%.

7.8.3 Inventaris beschikbare computationele grammatica's voor het Nederlands

- **Alpino**
  Omschrijving: "Alpino is a wide-coverage computational analyser of Dutch which aims at accurate, full, parsing of unrestricted text. The grammar produces dependency structures, thus providing a reasonably abstract and theory-neutral level of linguistic representation."

  Meer informatie: `http://odur.let.rug.nl/~vannoord/papers/alpino.pdf`

- **AMAZON-CASUS**
  Omschrijving: AMAZON is een grammatica die het Nederlands redelijk breed afdekt: ongeveer 95% van de ingevoerde tekst kan door het systeem ontleed worden. De AMAZON-component is een shallow parser. Met behulp van de component CASUS worden de door AMAZON opgeleverde analyses verder verfijnd om zo tot een totale parsering te komen.

  Meer informatie: `http://lands.let.kun.nl/amazon/`

- **Carp Technologies parseermodules**
  Omschrijving: Carp Technologies ontwikkelt technologieën voor het verwerken van menselijke taal door computers. Carp Technologies biedt enkele van deze technologieën ook aan in de vorm van softwarecomponenten, waarmee ontwikkelaars zelf applicaties kunnen ontwikkelen die menselijke taal begrijpen en verwerken. Carp Technologies heeft onder meer een module voor het parseren van natuurlijke taal ontwikkeld.

  Meer informatie: `http://www.carp-technologies.nl/low/nl/taalsoftware.html`

- **Corrie**
  Omschrijving: Theo Vosse ontwikkelde de spelling- en grammaticacorrector CORRIE voor het Nederlands (Vosse, 1994). Onderdeel van het programma is een (Tomita)-parser voor het Nederlands.

  Meer informatie: Vosse, T., (1994) *The Word Connection*. Proefschrift, Rijksuniversiteit Leiden.

- **Delilah**
  Omschrijving: Delilah is een ontleder voor het Nederlands, ontwikkeld door Crit Cremers en Maarten Hijzelendoorn van de vakgroep Algemene Taalwetenschap, Rijksuniversiteit Leiden.

  Meer informatie: `http://fonetiek-6.leidenuniv.nl/hijzlndr/delilah.html`

- **LS-GRAM+**
  Omschrijving: Binnen het door de EU-commissie gefinancierde project LS-GRAM+ (Large Scale Grammars for EU Languages) zijn grammaticale resources ontwikkeld voor negen Europese talen waaronder het Nederlands. Het project liep van januari 1994 tot juli 1996. De grammatica's zijn ontwikkeld op basis van corpusgegevens. De Nederlandse bijdrage aan het project bestond uit een samenwerking van het Centre for Computational Linguistics van de KU Leuven en SST (Foundation of Speech Technology). De Nederlandse lingware-module heeft een lexiconcomponent en een component voor *phrase structure*.

  Meer informatie:
  `http://www.iai.uni-sb.de/LS-GRAM/home.html` en `http://www.ccl.kuleuven.ac.be/about/LS-GRAM.html`

- **OVIS-parser**
  Omschrijving: Deze "bi-directional, head-driven parser for constraint-based grammars" werd ontwikkeld voor het OVIS-systeem: een Nederlands dialoogsysteem waarin informatie over het openbaar vervoer via de telefoon kan worden verkregen.

  Meer informatie:
  `http://odur.let.rug.nl/~vannoord/papers/cl97/`, `http://odur.let.rug.nl:4321/tstplan/c6.html` en `http://grid.let.rug.nl:4321/`

- **Performance Grammar**
  Omschrijving: De Performance Grammar (PG) is een psycholinguïstisch gemotiveerd grammaticaformalisme voor natuurlijke talen. Tot nu toe zijn Nederlandse, Duitse en Engelse fragmenten uitgewerkt.

  Meer informatie: `http://www.liacs.nl/~cvbreuge/ToKeN2000/`

- **SynTag**
  Omschrijving: Binnen het project TwentyOne van TNO is een snelle PSG-parser ontwikkeld die gebruikt wordt voor NP-extractie. Deze module maakt gebruik van een simpele grammatica voor het Nederlands, door het Twentse Parlevink geschreven, en een taalonafhankelijke parser-generator, ontwikkeld bij TNO. Inmiddels zijn de parser en de grammatica overgedaan aan het bedrijf Irion. Hier wordt de parser verkocht onder de naam SynTag.

  Meer informatie: `http://twentyone.tpd.tno.nl/twentyone` en `http://www.irion.nl`

7.8.4 Evaluatie

Alpino

- Specifieke Eigenschappen
  - Taalkundig kader: Head-driven phrase structure grammar.
  - Eisen aan de input: Willekeurige tekst.
  - Volledige of partiële analyse?: Volledige analyse.

  - Omvang van het woordenboek: Tenminste 150.000 woorden op basis van Celex en Parole, met valentiepatronen.
  - disambiguatie: disambiguatie op basis van statistische methoden.

- Algemene eigenschappen
  - Programmeertaal en besturingssysteem: Prolog op Unix/Linux.
  - Beschikbaarheid: Beschikbaar op termijn.
  - Bruikbaarheid: Het systeem wordt voorlopig alleen gebruikt als hulpmiddel bij syntactische annotatie van een corpus. Voor toepassingen waarbij snelheid van belang is, zullen experimenten worden uitgevoerd met contextvrije of reguliere approximaties van de grammatica.

- Documentatie: [22, 21, 110]

- Kwaliteit
  - Accuraatheid: 75 tot 85 % accuratesse voor het identificeren van dependency relations (afhankelijk van tekstsoort).
  - Snelheid en geheugen: Gemiddeld ongeveer 10 sec per zin (bij een gemiddelde zinslengte van 20 woorden).

Amazon-Casus

- Specifieke Eigenschappen
  - Taalkundig kader: Structuralistisch.
  - Eisen aan de input:
  - Volledige of partiële analyse?: Volledige analyse.
  - Omvang van het woordenboek: Ongeveer 300.000 woorden op basis van Celex.
  - disambiguatie: Nog steeds geldt in AMAZON dat de ambiguïteit zoveel mogelijk wordt beperkt, door bij structurele ambiguïteit te kiezen voor onderspecificatie - een conceptuele beschrijving van de structuur waaruit alle mogelijkheden gegenereerd kunnen worden - of voor de meest waarschijnlijke oplossing. Voorbeelden van de omzeiling van structurele ambiguïteit zijn de beschrijving van coördinatie, die alleen rechtsrecursief wordt geanalyseerd, en de inperking van de aanhechting van voorzetselgroepen (PP's) als complementen bij zelfstandig naamwoordgroepen.

- Algemene eigenschappen
  - Programmeertaal en besturingssysteem: De gebruikte parsergenerator is in C. Deze draait onder DOS en UNIX.
  - Beschikbaarheid: Alle rechten met betrekking tot AMAZON en CASUS berusten bij de sectie Computerlinguïstiek van de afdeling Taal en Spraak aan de Katholieke Universiteit Nijmegen.

  - Bruikbaarheid: Het huidige AMAZON/CASUS-systeem kan snel zinnen analyseren. Door het scheiden van theorie en algoritme, is het eenvoudig te begrijpen, te onderhouden en uit te breiden.

- Documentatie: Enkele regels in de broncode en enkele los daarvan verschenen documenten. Op `http://lands.let.kun.nl/TSpublic/dreumel/amazon\_document.nl.html` is een eerste aanzet te vinden om broncode en commentaar te combineren. Uiteindelijk moet dit ertoe leiden dat de gehele grammatica grondig gedocumenteerd is in de broncode.

- Kwaliteit

  - Accuraatheid: Het is een grammatica die het Nederlands redelijk breed afdekt: ongeveer 95% van de ingevoerde tekst kan door het systeem ontleed worden.
  - Snelheid en geheugen: Het systeem is snel.

Carp Technologies parseermodules

- Specifieke Eigenschappen

  - Taalkundig kader: Er worden verschillende grammatica's voor verschillende toepassingen gebruikt. De grammatica's voor dialoogsystemen zijn gebaseerd op semantiek (d.w.z. de non-terminals zijn dan semantische labels i.p.v. taalkundige). Voor analyse van teksten zijn taalkundige labels gebruikt. Die lijken op de grammatica's uit Nijmegen (Peter-Arno Coppen), maar produceren minder diepe parse trees (omdat ze minder labels bevatten). De activiteiten m.b.t. generatieve grammatica's zijn bij Carp Technologies nog slechts in het experimenteerstadium. De parser kan type 0 grammatica's aan en is daarmee even krachtig als een Turing Machine. De parser gebruikt een zo efficiënt mogelijke parseringsmethode afhankelijk van de complexiteit van de grammatica.
  - Eisen aan de input: De invoer bestaat uit platte tekst, opgemaakte html of opgemaakte rtf-bestanden. Hierin hoeven geen POS-tags aanwezig te zijn aangezien de parser zelf POS-tags aanbrengt. Verder wordt naast een uitgebreid lexicon gebruikgemaakt van morfologische analyse zodat woorden die niet in het lexicon voorkomen meestal toch goed worden verwerkt. De invoer wordt dus niet beperkt door het lexicon.
  - Volledige of partiële analyse?: Er zijn zowel grammatica's voor een partiële als voor een volledige analyse. De parser is robuust, dus als een grammatica voor volledige parsering wordt gebruikt en een zeer complexe invoerzin hier niet mee kan worden geparseerd, zal de parser toch proberen in ieder geval een oppervlakkige analyse te maken.
  - Omvang van het woordenboek: Afhankelijk van de toepassing 40.000 tot 500.000 woordvormen.
  - disambiguatie: disambiguatie gebeurt op dit moment regelgebaseerd, maar er wordt geëxperimenteerd met statistische methoden.

- Algemene eigenschappen

  - Programmeertaal en besturingssysteem: De taalverwerkende modules zijn volledig in Java geïmplementeerd en draaien op elk platform. In de toepassingen waarin ze worden gebruikt, wordt doorgaans een kleine hoeveelheid platformafhankelijke C++ code gebruikt waardoor *porten* erg gemakkelijk is.
  - Beschikbaarheid: De modulen voor parsing maken deel uit van producten die zijn ontwikkeld bij Carp Technologies, maar zouden eventueel ook als losse componenten geleverd kunnen worden.
  - Bruikbaarheid:
- Documentatie:
- Kwaliteit
  - Accuraatheid:
  - Snelheid en geheugen:

Corrie

(Aanvullende informatie uit 'Testing CORRIE for SCARRIE': `http://www.ling.uu.se/wp/wp3a.pdf`. Deze tekst gaat over de versie van CORRIE van 29 oktober 1997)

- Specifieke Eigenschappen
  - Taalkundig kader: De grammatica is een *augmented context-free grammar* (een CFG waaraan attributen zijn toegevoegd om zaken als persoon en getal te kunnen beschrijven) met plusminus 500 regels en 14 regels die speciaal voor het doen van correctie zijn toegevoegd.
  - Eisen aan de input:
  - Volledige of partiële analyse?:
  - Omvang van het woordenboek:
  - disambiguatie:
- Algemene eigenschappen
  - Programmeertaal en besturingssysteem: C, Unix.
  - Bruikbaarheid: Het toevoegen van een nieuwe herkenningsmodule bleek moeilijk te zijn volgens 'Testing CORRIE for SCARRIE'. Bestaande onderdelen die in principe niets met de nieuwe module te maken hebben moeten toch worden aangepast. De correctiemodule van CORRIE kan wel door een andere worden vervangen.
- Documentatie: Volgens 'Testing CORRIE for SCARRIE' was de documentatie niet compleet: "The manual pages of the software distribution lack features we expect from good documentation: good structure, completeness and correctness. We have been unable to find any documentation on new make pron program. We also noticed the absence of comment lines in code at places where they would have been very helpful. However nearly all tests were conducted without having to consult the CORRIE programmer so by investing time in reading both the code and the manual pages it is possible to obtain a basic idea of how the program works."

- Kwaliteit
  - Accuraatheid: Het systeem is getest op verschillende documenten (o.a. juridische teksten,wetenschappelijke boeken en scripties, en (6 megabyte) nieuwsberichten).
  - Snelheid en geheugen:

Delilah

- Specifieke Eigenschappen
  - Taalkundig kader: Het systeem past een categoriale grammatica - een categoriale lijst grammatica of minimale categoriale grammatica [36] - van enkele belangrijke zinsbouwverschijnselen in het Nederlands toe.
  - Eisen aan de input:
  - Volledige of partiële analyse?: Delilah ontleedt op basis van het lexicon onder meer allerlei vormen van linkse verplaatsing, werkwoordelijke verstrengeling in al z'n facetten, en vrije nevenschikking.
  - Omvang van het woordenboek: De ontleder wordt gevoed door een zeer bescheiden maar uit categoriaal oogpunt veelzijdig lexicon.
  - disambiguatie:
- Algemene eigenschappen
  - Programmeertaal en besturingssysteem:
  - Bruikbaarheid:
- Documentatie:
- Kwaliteit
  - Accuraatheid:
  - Snelheid en geheugen:

LS-GRAM+

- Specifieke Eigenschappen
  - Taalkundig kader: Head-Driven Phrase Structure Grammar.
  - Eisen aan de input:
  - Volledige of partiële analyse?:
  - Omvang van het woordenboek:
  - disambiguatie:
- Algemene eigenschappen
  - Programmeertaal en besturingssysteem:
  - Bruikbaarheid:

- Documentatie: Gedetailleerde documentatie is te downloaden op de website.
- Kwaliteit
  - Accuraatheid:
  - Snelheid en geheugen:

OVIS-parser

- Specifieke Eigenschappen
  - Taalkundig kader: De OVIS-grammatica is geïmplementeerd in een eenvoudig *constraint-based* formalisme: DCG.
  - Eisen aan de input: De input is een woordgraaf zoals die wordt afgeleverd door een spraakherkenner. De woordgraaf kan alleen woorden bevatten die ook in het vocabulaire van de spraakherkenner zitten (ongeveer 2000 woorden).
  - Volledige of partiële analyse?: Volledige analyse waar mogelijk, anders partiële analyse.
  - Omvang van het woordenboek: Het OVIS-lexicon bevat de namen van 500 treinstations en 500 woorden. Ambiguïteit meegerekend komt de omvang van het lexicon op ongeveer 1500 tot 2500 ingangen.
  - disambiguatie: op basis van heuristieken
- Algemene eigenschappen
  - Programmeertaal en besturingssysteem: Prolog
  - Beschikbaarheid: Het systeem wordt niet meer ondersteund.
  - Bruikbaarheid:
- Documentatie: [104, 105, 140]
- Kwaliteit
  - Accuraatheid: Ongeveer 84% word accuracy en 83% semantic accuracy.
  - Snelheid en geheugen: Gemiddeld 5 seconden per *word graph* in de langzaamste configuratie, 0,2 seconden per *word graph* in de snelste (iets minder accurate) configuratie.

Performance Grammar

- Specifieke Eigenschappen
  - Taalkundig kader: Psycholinguïstisch.
  - Eisen aan de input:
  - Volledige of partiële analyse?: Partiële analyse.
  - Omvang van het woordenboek: Op dit moment beperkt. Circa 2000 woorden, uitbreiding op termijn via uitwisselingsverbanden.

  - disambiguatie: Aanwezig.

- Algemene eigenschappen:
  - Programmeertaal en besturingssysteem:
  - Beschikbaarheid: Via `http://www.liacs.nl/~cvbreuge/pgw`
  - Bruikbaarheid:

- Documentatie: Hier wordt aan gewerkt, maar de documentatie heeft op dit moment nog geen prioriteit.

- Kwaliteit
  - Accuraatheid:
  - Snelheid en geheugen:

SynTag

- Specifieke Eigenschappen:
  - Taalkundig kader:
  - Eisen aan de input:
  - Volledige of partiële analyse?:
  - Omvang van het woordenboek:
  - disambiguatie:

- Algemene eigenschappen:
  - Programmeertaal en besturingssysteem:
  - Beschikbaarheid: De parser wordt per taal en per grammatica automatisch gegenereerd door de parser generator. Op basis van grammatica's kunnen de parsers hierdoor gemakkelijk worden aangepast.
  - Bruikbaarheid:

- Documentatie:

- Kwaliteit
  - Accuraatheid:
  - Snelheid en geheugen:

Tijdens de LOT Winterschool 2001 is er een poging ondernomen de systemen Alpino, Amazon, en Delilah (en de CGN annotatie software) te vergelijken. Bij gebrek aan een *treebank* voor het Nederlands was het vooralsnog onmogelijk een serieuze kwantitatieve evaluatie uit te voeren. Uit deze bijeenkomst kwam wel naar voren dat de annotatievoorschriften van CGN een goede richtlijn kunnen zijn voor het creëren van zo'n treebank (die dan, i.t.t. CGN, vooral geschreven tekst zou moeten bevatten).

7.8.5 Conclusies

Het grootste probleem voor computationele syntactische analyse voor het Nederlands is dat geen van de genoemde systemen het onderwerp is geweest van een objectieve en serieuze evaluatie. (De enige uitzondering is wellicht het OVIS-systeem, maar dat is een systeem met een zeer beperkt domein.) Daarnaast kan worden geconstateerd dat de meeste beschreven systemen ofwel niet meer actief ondersteund worden ofwel in ontwikkeling zijn. Concreet betekent dit dat er momenteel geen modules voor syntactische analyse zijn die toegankelijk of bruikbaar zijn voor derden.

Naast deze tekortkomingen schieten veel (met name academische) systemen ook tekort waar het gaat om *coverage* (vaak is sprake van grammatica's voor een beperkt domein met een beperkte woordenschat). De systemen met een grotere *coverage* voeren vaak slechts een oppervlakkige syntactische analyse uit. In tegenstelling tot het buitenland, waar het gebruik van statistische technieken de dominante trend is, staat het gebruik van statistiek hier nog in de kinderschoenen.

Om verandering aan te brengen in deze stand van zaken lijkt het vooral noodzakelijk te investeren in geannoteerd corpusmateriaal dat kan dienen als basis voor de training van statistische (modules van) parsers en als basis voor evaluatiecriteria.

7.9 Semantische en pragmatische analyse

In deze sectie behandelen we modules die gericht zijn op de semantische en pragmatische analyse van tekst en spraak. Semantische en pragmatische analyse is van belang voor toepassingen waarbij de betekenis van uitingen, de context waarin uitingen worden gedaan, en de bedoeling van de gebruiker, een rol spelen.

Semantische analyse heeft als doel tekst en spraak om te zetten in semantische representaties, zoals bijvoorbeeld logische formules. Semantische representaties kunnen op hun beurt worden omgezet in een applicatiespecifiek formaat, zoals bijvoorbeeld SQL. Semantische analyse vervult dus een belangrijke rol in dialoogsystemen die als front-end fungeren voor een bepaalde applicatie.

De context van een uiting is belangrijk voor systemen die proberen een natuurlijke dialoog met de gebruiker te voeren. Dit betekent bijvoorbeeld dat voor een vraag als "gaat er nog een latere trein?" de voorafgaande dialoog kan worden gebruikt voor het interpreteren van "latere trein".

Pragmatische analyse is vooral gericht op het vaststellen van de bedoeling van de gebruiker, en op het produceren van antwoorden door een dialoogsysteem die aansluiten bij de kennis en het doel van de gebruiker. In een dialoogsysteem is het bijvoorbeeld van belang bevestigingen, ontkenningen, groeten, correcties, vragen, etc. juist te kunnen herkennen.

Bepaalde aspecten van semantische analyse spelen een rol in automatisch vertalen. Het Nederlandse woord *instelling* vertaalt bijvoorbeeld onder meer naar het Engelse *setting, attitude* en *organization*. Het vaststellen van de juiste betekenis van een woord (*word sense disambiguation*) is daarom essentieel voor het kiezen van de juiste vertaling. Het bepalen van de interpretatie van persoonlijke voornaamwoorden (*ik, zij, hem, zich, ons,..*) (*pronoun resolution*) speelt een rol in de vertaling van zinnen als *Jan zag dat Marie zichzelf in moeilijkheden bracht*, waarbij *zichzelf* in het Engels als *himself* of *herself* vertaald kan worden.

Systemen voor semantische en pragmatische analyse die een vrij gedetailleerde en precieze analyse van de invoer uitvoeren, maken over het algemeen gebruik van de resultaten

van syntactische analyse. Syntactische analyse helpt bijvoorbeeld om de juiste structuur en grammaticale relaties aan de woorden in een zin toe te kennen. Semantische analyse bouwt vervolgens voort op de beslissingen die in de syntactische analyse zijn genomen.

Componenten die een oppervlakkiger analyse uitvoeren, maken soms gebruik van eenvoudiger technieken, gebaseerd op *keyword* of *concept spotting*. In dat geval wordt de betekenis van een zin bepaald op basis van een klein aantal trefwoorden in de zin, met weinig of geen aandacht voor de relaties tussen die woorden. Dit kan voldoende zijn voor eenvoudige dialoogsystemen.

Word sense disambiguation maakt veelal gebruik van statistische technieken. De juiste betekenis van een ambigu woord wordt dan voorspeld door te kijken naar de omringende woorden, en de statistieken die zijn verzameld (op basis van een corpus) over de kans dat een gegeven betekenis met deze contextwoorden voorkomt.

Resolutie van pronomina maakt gebruik van syntactische informatie in combinatie met statistische gegevens. Statistiek voorspelt bijvoorbeeld dat onderwerpen vaker als antecedent voor een pronomen fungeren dan nominale constituenten met andere grammaticale rollen, dat de antecedent relatief vaak de voorafgaande nominale constituent is, etc.

7.9.1 Specifieke evaluatiecriteria

- Afhankelijk van syntactische analyse?
- Is er sprake van een taalspecifieke techniek, of worden algemene (statistische) technieken gebruikt?
- Op basis van corpora?
- Evaluatiegegevens?

7.9.2 State of the art internationaal

Een probleem bij de evaluatie van semantische modules is dat deze enerzijds vaak nauw verweven zijn met een syntactische module, en anderzijds vaak een output opleveren die specifiek is voor een bepaald formalisme of voor een bepaalde applicatie.

SRI Cambridge werkt sinds ongeveer 10 jaar aan de *Core Language Engine*, een *wide-coverage* grammatica met semantische component voor het Engels. De output van het systeem is een zogenaamde *quasi logical form*. Met behulp van deze module zijn verschillende applicaties gemaakt op het gebied van natuurlijke taal-interfaces en automatisch vertalen, zowel op basis van tekst als op basis van spraak.

Binnen het Nederlandse NWO-TST project is een evaluatie uitgevoerd waarbij de semantische representaties die door twee natuurlijke taal-modules worden opgeleverd zijn vergeleken met een geannoteerd corpus. De toepassing was een spraak-interface voor informatie over het openbaar vervoer.

Voor deelproblemen op het gebied van semantische interpretatie zijn ook evaluaties uitgevoerd.

Het Senseval-experiment evalueert verschillende systemen voor *word sense disambiguation*. Voor de evaluatie zijn een aantal ambigue woorden in een corpus geannoteerd met de juiste betekenis. Betekenisonderscheidingen zijn ontleend aan WordNet. Een van de conclusies die uit dit experiment kunnen worden getrokken is dat het trekken van grenzen tussen verschillende betekenissen van een woord vaak lastig is.

Daarnaast zijn er pogingen ondernomen om materiaal te annoteren met informatie over anaforische en pronominale relaties. Dergelijk materiaal kan worden gebruikt om modules voor de resolutie van pronomina te evalueren.

7.9.3 Inventarisatie voor het Nederlands

- **Irion WSD-module**
  Omschrijving: Irion Technologies BV ontwikkelt een module voor woordbetekenisdisambiguering welke zal worden toegepast binnen een cross-lingual retrieval systeem.
  Meer informatie: `http://www.irion.nl`

- **Carp Technologies semantische module**
  Omschrijving: Carp Technologies ontwikkelt technologieën voor het verwerken van menselijke taal door computers. Carp Technologies biedt enkele van deze technologieën ook aan in de vorm van softwarecomponenten, waarmee ontwikkelaars zelf applicaties kunnen ontwikkelen die menselijke taal begrijpen en verwerken. Carp Technologies heeft onder meer een semantische module ontwikkeld.
  Meer informatie:
  `http://www.carp-technologies.nl/low/nl/taalsoftware.html`

7.9.4 Evaluatie

IRION WSD module

- Afhankelijk van syntactische analyse? Informatie over de precieze wijze waarop de disambiguering plaatsvindt wordt niet prijsgegeven.
- Is er sprake van een taalspecifieke techniek, of worden algemene (statistische) technieken gebruikt? Idem.
- Op basis van corpora? Idem.
- Evaluatiegegevens? Nee.
- Beschikbaarheid: Irion hoopt eind 2001 een commerciële versie van de module klaar te hebben. Deze zal verkocht worden als een black box die is geïntegreerd in een multilinguaal semantisch netwerk en eventueel een automatische *classifier*. Het is mogelijk om dat netwerk met de disambiguering te kopen of te leasen. Dit kan ook als onderdeel van een ander product, bijvoorbeeld een cross-lingual retrieval systeem. Precieze methodes hangen af van de combinatie: alleen semantisch netwerk, netwerk met classificatie, netwerk en classificatie in cross-lingual retrieval engine.

Carp Technologies semantische module

- Afhankelijk van syntactische analyse? Ja.
- Is er sprake van een taalspecifieke techniek, of worden algemene (statistische) technieken gebruikt? Zowel taalspecifieke technieken als statische technieken.
- Op basis van corpora? Ja, indien gebruik wordt gemaakt van statistische technieken.

- Evaluatiegegevens? Geen recente evaluatiegegevens beschikbaar.
- Beschikbaarheid: De modulen voor semantische analyse maken deel uit van producten die zijn ontwikkeld bij Carp Technologies, maar zouden eventueel ook als losse componenten geleverd kunnen worden.

7.9.5 Conclusies

Semantische analyse van het Nederlands is een vrijwel onontgonnen terrein. Gezien het feit dat hierboven twee bedrijven genoemd worden die actief zijn op dit terrein kan wel worden geconcludeerd dat er een duidelijke behoefte is aan software die teksten ook op semantisch en pragmatisch niveau kunnen analyseren. Veel systemen voor semantische en pragmatische analyse bouwen op de resultaten van (oppervlakkige) syntactische analyse. Syntactische analysemodules zijn voor het Nederlands echter vrijwel niet beschikbaar.

Om te voorzien in de behoefte aan componenten voor semantische en pragmatische analyse lijkt een syntactische component onmisbaar. Het Euronet-project heeft een lexicale database opgeleverd die gebruikt kan worden als basis voor WSD. Om WSD voor het Nederlands verder te ontwikkelen lijkt het opzetten van een geannoteerd corpus een eerste vereiste.

7.10 Generatie

Tekstgeneratie is het produceren van een tekst op basis van niet-talige informatie. Te denken valt bijvoorbeeld aan het produceren van een weerbericht in tekstvorm zoals dat wordt afgedrukt in de krant of op Teletekst. Een dergelijke tekst is het resultaat van de interpretatie van numerieke meteorologische gegevens. Verwante toepassingen (die wel bekend staan onder de noemer *rapportgeneratie*) zijn het maken van beursberichten en het produceren van brieven waarvan de inhoud gebaseerd is op klantgegevens. Tekstgeneratie kan ook worden gebruikt voor het produceren van een beschrijving van de inhoud van een database, of van uitleg over de redenering van een kennissysteem.

Een andere belangrijke applicatie is automatisch vertalen. Bij het omzetten van een Engelse tekst in het Nederlands kan men niet volstaan met het woord voor woord omzetten van de brontekst. De Engelse tekst zal eerst syntactisch geanalyseerd moeten worden, en op basis van die analyse kan een correcte Nederlandse zin gegenereerd worden.

Tenslotte speelt tekstgeneratie een beperkte rol in spraaksystemen die informatie over banksaldo's, rekening bij- of afschrijvingen, vluchtinformatie, etc. geven.

Tekstgeneratie wordt vaak opgedeeld in twee componenten. De eerste component is verantwoordelijk voor *wat* er gezegd moet gaan worden (die bijvoorbeeld beslist welke informatie deel moet uitmaken van de tekst, en in welke volgorde deze informatie gepresenteerd moet worden). Een tweede component is verantwoordelijk voor *hoe* iets gezegd gaat worden. De laatste component maakt in het eenvoudigste geval gebruik van vaste zinspatronen waarbinnen bijvoorbeeld alleen een datum, bedrag, of naam wordt ingevuld, en in meer geavanceerde systemen van een taalkundig gebaseerd generatie-algoritme. Goede tekstplanning kan helpen om een tekst compact en informatief te maken. Goede formulering van een tekst (bijvoorbeeld in de context van automatisch vertalen) helpt storende taalfouten voorkomen en komt in het algemeen de leesbaarheid en begrijpelijkheid van een tekst ten goede.

Een toepassing die nauw verwant is aan tekstgeneratie, is automatisch samenvatten van teksten. Net als tekstgeneratie is het van belang beslissingen te nemen over *wat* er gezegd moet worden. Voor de formulering van de samenvatting wordt vaak gebruikgemaakt van relevante

zinnen uit het originele document, maar het is voorstelbaar dat hiervoor in de toekomst ook generatie wordt gebruikt.

7.10.1 State of the art internationaal

De pagina (`http://www.dynamicmultimedia.com.au/siggen/`) van de ACL *special interest group in generation* bevat links naar een aantal demonstraties van tekstgeneratie, waaronder systemen die rondleiding verzorgen in een virtueel museum (en daarbij in hun uitleg rekening houden met de kennis van de gebruiker en de afgelegde weg), systemen die beursberichten, weerberichten, en (Nederlandstalige) voetbalverslagen produceren, en een systeem dat persoonlijke brieven produceert die de geadresseerde oproepen te stoppen met roken.

7.10.2 Evaluatiecriteria

- Inclusief of exclusief tekstplanning?
- Tekstformulering op basis van *template filling* of taalkundige kennis?
- Gekoppeld aan spraakgeneratie?

7.10.3 Inventarisatie en evaluatie

- **LGM Omschrijving:** LGM (Language Generation Module) is een module van Polderland waarmee gestructureerde data wordt omgezet in natuurlijke taal. De module is oorspronkelijk op het IPO ontwikkeld. De LGM produceert teksten die aanwijzingen bevatten voor prosodie (plaatsing van accenten en frasegrenzen) en houdt rekening met de manier waarop het eerdere informatie heeft 'vertaald'. De laatste versie van deze module is geschikt voor gebruik met verschillende informatiesystemen. Zo kunnen bijvoorbeeld ook weerberichten worden gegenereerd.

  Meer informatie:
  `http://www.polderland.nl`
  `http://wwwhome.cs.utwente.nl/~theune/` en
  `http://wwwhome.cs.utwente.nl/theune/GG/GG_index.html`.

  Inclusief of exclusief tekstplanning? Inclusief tekstplanning. De LGM bevat (in tegenstelling tot de meeste andere NLG-systemen) geen aparte tekstplanningsmodule die van tevoren een globaal tekstplan maakt, dat later verder wordt ingevuld. Dit betekent echter niet dat er geen tekstplanning plaatsvindt. Dit gebeurt echter niet in één keer, maar 'incrementeel'; stapje voor stapje. Gegeven de reeds gegenereerde tekst, wordt er steeds per zin bepaald hoe de tekst kan worden voortgezet. Dit hangt af van (1) de condities op de syntactische templaten, die (tot op zekere hoogte) de volgorde van de zinnen in de tekst bepalen, en (2) de 'topic' informatie die met de templaten is geassocieerd, en die bepaalt hoe de zinnen in een tekst gegroepeerd kunnen worden (zinnen met hetzelfde onderwerp worden bij elkaar gehouden).

  Tekstformulering op basis van template filling of taalkundige kennis? Een combinatie van beide. Er worden templaten gebruikt, maar deze bevatten taalkundige kennis (in de vorm van een syntactische structuur) en de manier waarop ze worden ingevuld wordt mede bepaald op taalkundige gronden. Voorbeelden zijn het gebruik

van verschillende soorten verwijzende uitdrukkingen, controle op antecedenten en op naleving van de Binding Theorie. Zaken zoals lexical choice, keuze van de zinsstructuur en morfologie gebeuren echter niet op taalkundig verantwoorde wijze, maar zijn 'hard ingebakken' in de templaten.

Gekoppeld aan spraakgeneratie? De LGM kan als zelfstandig systeem gebruikt worden. Doordat er ook automatisch prosodische annotatie wordt geproduceerd, kan de LGM overigens goed worden gecombineerd met spraakuitvoer.

Beschikbaarheid: Voor commercieel gebruik is de LGM niet vrij beschikbaar; als een bedrijf de module wil gebruiken voor een applicatie, dient er met Polderland contact op te worden genomen. De module is platformonafhankelijk.

- **ToKeN2000-zinnengenerator**

  Omschrijving: In het kader van ToKeN2000 wordt een Nederlandstalige zinnengenerator gebouwd, die gebaseerd is op het psycholinguïstisch gemotiveerde Performance Grammar formalisme. Het betreft de voortgezette ontwikkeling van het interactieve zinsbouwsysteem dat 'Performance Grammar Workbench' (PGW) heet. De PGW en de zinnengenerator zijn bedoeld om te fungeren als het syntactische 'hart' van toepassingssystemen waarin gesproken of geschreven zinnen on line vervaardigd moeten kunnen worden en aan de gebruiker gepresenteerd. Dit geldt voor informatiesystemen die een gevarieerde taaluitvoer vereisen zodat niet volstaan kan worden met gepreformatteerde teksten. Daarnaast zijn toepassingen mogelijk in computerondersteund taalonderwijs en ter ondersteuning van taalcommunicatie van gehandicapten.

  De zinnengenerator is voorlopig alleen bedoeld voor het Nederlands, maar is wel zo opgezet dat voor een andere taal de software niet aangepast hoeft te worden, alleen de lexicale data.

  Meer informatie:
  `http://www.liacs.nl/~cvbreuge/ToKeN2000/` en `http://www.liacs.nl/~cvbreuge/pgw/`

  Inclusief of exclusief tekstplanning? Exclusief tekstplanning. De generator is vooralsnog vooral een zinsbouwmodule.

  Tekstformulering op basis van*template filling* of taalkundige kennis? Taalkundige kennis. De input is een logische vorm die al wel zeer dicht bij het zinsniveau ligt, retoriek is hierin al uitgekristalliseerd.

  Gekoppeld aan spraakgeneratie? Ja, er wordt gebruikgemaakt van de spraakmachine van het IPO in Eindhoven. De prosodieplaatsing moet nog geïmplementeerd worden

  Beschikbaarheid: Niet te koop, alleen voor onderzoeksdoeleinden beschikbaar op aanvraag.

- **CONPAS**

  Omschrijving: Het systeem waar CONPAS deel van uitmaakt bestaat uit vier modules: (1) een domeinspecifieke tekstplanner (voor treinreisbeschrijvingen, en voor botanische beschrijvingen van wilde bloemen); (2) een syntactische realisator die abstracte beschrijvingen van de syntactische structuur omzet naar grammaticale zinnen;

(3) een prosodische realisator die o.a. accenten en grenzen plaatst; (4) een systeem voor spraaksynthese (KUNTTS).

De eerste module is domeinafhankelijk, terwijl alle andere modules redelijk domeinonafhankelijk zijn, en in principe dus geschikt zijn voor hergebruik. De domeinonafhankelijke modules vormen samen CONPAS. Het systeem is vrij beschikbaar maar de ontwerper wijst erop dat het een onderzoekssysteem betreft dat zeker niet foutloos is, weinig of niet gedocumenteerd is en dat het moeilijk draaiende te krijgen kan zijn op een andere machine. Ook heeft hij weinig tijd om het systeem te ondersteunen.

Inclusief of exclusief tekstplanning? Inclusief.

Tekstformulering op basis van *template filling* of taalkundige kennis? Op basis van taalkundige kennis.

Gekoppeld aan spraakgeneratie? Ja.

Documentatie: [96, 95]

De grammatica's zijn niet gedocumenteerd.

- **Sinope**

  Omschrijving: Sinope is een systeem dat volledig automatisch samenvattingen kan maken van willekeurige Engelse- en Nederlandstalige teksten. Het gebruikt hiervoor NLP-technieken (parsing, semantische analyse en tekstgenerering). De samenvatting van een tekst van enkele pagina's wordt binnen enkele seconden gegenereerd.

  Eerst neemt de summarizer de semantische structuur die de semantische analyse van Sinope heeft opgeleverd en 'snoeit' die zodat de belangrijkste delen overblijven. De tekstgenerator gebruikt dan de output van de summarizer – een samengevatte semantische structuur - om een nieuwe tekst te genereren. Daarnaast is de generator verantwoordelijk voor het verenigen van zinnen.

  Meer informatie:
  `http://www.carp-technologies.nl/low/nl/sinope.html` en
  `http://www.carp-technologies.nl/SumatraTWLT14paper/SumatraTWLT14.html`

  Inclusief of exclusief tekstplanning? Sinope gebruikt de volgende twee stappen bij het genereren van een tekst:

  1. Content determination & tekstplanning
     Content determinatie gebeurt door de samenvatmodule die bepaalt welke informatie-elementen belangrijk zijn en welke niet. De literatuur over tekstgeneratie beschrijft meestal hoe een nieuwe tekst kan worden gegenereerd uitgaande van een of andere datastructuur. De tekstgenerator moet deze structuur omzetten in een gestructureerde tekst. Maar in het geval van Sinope is een gestructureerde tekst reeds aanwezig in de vorm van de originele tekst, hetgeen de tekstplanning vereenvoudigt.
  2. Sentence planning.
     (a) Toevoegen van anaforische expressies (ter voorkoming van repeterende noun phrases).
     (b) Sentence aggregation (samenvoegen van twee of meer zinnen).

(c) Genereren paragraafstructuur.

Tekstformulering op basis van *template filling* of taalkundige kennis? Tekstformulering is beperkt tot het genereren van een coherente en leesbare tekst. In de praktijk betekent dit dat kop en staart worden aangepast. Hiervoor wordt zowel gebruik gemaakt van templates als taalkundige kennis. Voor de twee hiervoor genoemde stappen bij het genereren van een tekst wordt uitsluitend gebruikgemaakt van taalkundige kennis.

Gekoppeld aan spraakgeneratie? Er zijn plannen om Sinope te koppelen aan een e-mailvoorlezer.

Beschikbaarheid: Te koop.

7.10.4 Evaluatie

De toegenomen belangstelling voor intelligente informatie-extractie zal er toe leiden dat ook het belang van generatie en combinaties van generatie en samenvatten, zal toenemen.

Het Nederlands beschikt over een klein aantal onderzoekers die *state-of-the-art* onderzoek verrichten naar generatie, zowel vanuit een taalkundig en cognitief perspectief, als vanuit een meer toegepast perspectief. Daarnaast is er belangstelling vanuit het bedrijfsleven voor generatie.

Het is niet geheel duidelijk wat de status van tekstgeneratie binnen een BATAVO is. Eén van de problemen is dat de scheidslijn tussen applicatiespecifieke onderdelen en generieke componenten moeilijk te trekken is. Generators die gebruikmaken van een onafhankelijke grammatica zijn duidelijk gebaat bij een goede computationele grammatica en een elektronisch lexicon voor het Nederlands. Veel van de toegepaste en praktisch bruikbare systemen maken echter gebruik van een beperkte grammatica die is toegesneden op het domein en de generatie-taak.

Uit het overzicht blijkt verder dat combinaties van generatie en samenvatten in het onderzoek ondervertegenwoordigd zijn. Omdat met name deze combinatie van belang lijkt voor intelligente informatie-extractie lijkt het verstandig onderzoek in deze richting te stimuleren.

7.11 Vertaalcomponenten

Automatisch vertalen is al jaren een onderzoeksterrein waarvoor zowel bij taaltechnologen als bij potentiële gebruikers grote belangstelling bestaat. De meeste software die is ontwikkeld is echter niet geschikt voor vertaling van tekst vanuit of naar het Nederlands. Bovendien is de bestaande vertaaltechnologie meestal alleen geschikt voor beperkte domeinen, is de kwaliteit van de vertalingen belabberd, en het gebruik beperkt tot een bepaalde werkomgeving. Wel zijn er allerlei ondersteunende hulpmiddelen beschikbaar voor vertalers die gebaseerd zijn op de opslag van bestaande vertalingen van teksten, zogenaamde 'translation memory systemen'. Bij de ontwikkeling daarvan gaat het niet zozeer om de inzet van taalspecifieke componenten, maar om ondersteuning van het vertaalwerk en het beheren van bestanden. Voor een overzicht

van het veld zie: John Hutchins, 'The State of Machine Translation in Europe and Future Prospects', In: LeJournal (online publication of HLTCentral), January 2002: http://www.hltcentral.org/page-917.shtml

Eind 1998 sloten de Taalunie en de Europese Commissie een overeenkomst over de gezamenlijke uitvoering van het project NL-Translex. NL-Translex is gericht op de ontwikkeling van een vertaalsysteem tussen het Nederlands en het Engels, Frans en eventueel het Duits, waarbij het Nederlands bron- en doeltaal is. De stand van zaken binnen NL-Tanslex is te volgen via http://www.taalunie.org/_/werkt/technologie.html#NL-Translex.

8 Taaltechnologie data

8.1 Lexica en thesauri

Elektronische toegang tot informatie die normaalgesproken in een woordenboek kan worden gevonden is belangrijk voor bijna alle toepassingen op het gebied van taal- en spraaktechnologie.

De aard van de benodigde lexicale informatie hangt erg van de toepassing af. Voor spellingcorrectie volstaat een eenvoudige woordenlijst, voor automatisch vertalen is een tweetalige woordenlijst nodig, voor tekstclassificatie of *information retrieval* kunnen lijsten met namen (ook wel *gazetteers* genoemd) van personen, organisaties, of geografische aanduidingen nuttig zijn, evenals woordenboeken met semantische informatie (over bijvoorbeeld synoniemen of ingedeeld op onderwerp (*sport, politiek, financieel, etc.*) nuttig zijn, voor woordsoortdisambiguatie en syntactische analyse is informatie over syntactische categorie en overige syntactische kenmerken van belang, en voor tekst-naar-spraak is informatie over de uitspraak van woorden nodig.

Traditionele woordenboeken zijn in de eerste plaats gericht op menselijke taalgebruikers, en leggen daarom de nadruk op kwesties als spelling, uitspraak, en betekenis, waar voor het laatste gebruik kan worden gemaakt van omschrijvingen. Voor toepassingen binnen TST is soms andere informatie nodig, of is het nodig informatie in een welomschreven formaat aan te bieden. Informatie over de frequentie van een woord in een bepaald corpus is bijvoorbeeld voor TST van groot belang, maar speelt in 'gewone' woordenboeken niet of nauwelijks een rol. Informatie over de uitspraak of betekenis van woorden is voor taaltechnologische toepassingen vooral nuttig wanneer er een vast omschreven formaat wordt gehanteerd.

Traditionele woordenboeken zijn ook om een andere reden niet direct geschikt voor taaltechnologische toepassingen. Een 'gewoon' woordenboek kan bijvoorbeeld alleen voor automatische spellingcontrole gebruikt worden wanneer de data op zo'n manier beschikbaar worden gemaakt dat ze geïntegreerd kunnen worden met een tekstverwerker. Beschikbaarheid op CD-ROM is hiervoor geen garantie. Wanneer men voor taaltechnologische doeleinden direct toegang wil tot de lexicale informatie die in een woordenboek is opgeslagen, zal men dus vaak apart toestemming en toegang moeten zien te verkrijgen van de uitgever van het woordenboek.

De databestanden die hieronder genoemd worden onderscheiden zich van gewone woordenboeken (al dan niet op CD-ROM) doordat het materiaal toegankelijk is voor gebruik in een taaltechnologische toepassing.

8.1.1 State of the art internationaal

Voor het Engels zijn er lexicale databases die voorzien in verschillende soorten van informatie. Bestaande woordenboeken zijn deels ook geschikt voor gebruik in TST. De elektronische versies van LDOCE (Longman Dictionary of Contemporary English) en Cobuild worden bijvoorbeeld gebruikt voor woordsoortdisambiguering en uitspraakinformatie.

WordNet is een zeer veel gebruikte lexicale database waarin semantische informatie is opgenomen. De database is georganiseerd als een verzameling concepten, met voor ieder concept een of meer woorden die dit concept uitdrukken, en een overzicht van deel-geheel relaties (bijvoorbeeld: een vinger is een deel van hand). WordNet wordt onder andere gebruikt als de basis voor woordbetekenisdisambiguering. ComLex bevat systematische informatie over syntactische valentie voor zo'n 38.000 lemma's. Lexicale databases die speciaal gericht zijn op gebruik voor spraaktechnologie worden behandeld in sectie 10.

8.1.2 Evaluatiecriteria

Volgende evaluatiecriteria zijn speciaal voor lexica van belang:

- Aard van de aanwezige informatie (uitspraak, tweetalig, valentie, betekenis, etc.)
- Omvang van de database
- Aansluiting bij een internationale standaard
- Formaat (tekst, database, XML, etc.)
- Beschikbaarheid
- Documentatie

8.1.3 Inventarisatie

Woordenlijsten voor correctie en afbreken

- **Sdu/Elektronisch Groene Boekje**

  Omschrijving: Dit is de elektronische versie van de Woordenlijst Nederlandse Taal (Woordenlijst, 1996). De lijst is vooral bedoeld om de juiste spelling van een woord op te zoeken.
  Beschikbaarheid: Distributie door de SDU.

  Meer informatie: `http://www.sdu.nl/`

  Referentie: Woordenlijst, (1996). *Woordenlijst Nederlandse Taal*. Sdu, Den Haag.

- **Sdu/Standaard Spellingschijf**

  Omschrijving: Dit programma is bedoeld voor spellingcorrectie. Het programma is bedoeld voor de tekstverwerker WordPerfect, en zorgt voor een *update* van de woordenlijst die door WordPerfect wordt geleverd. De regels van de nieuwe spelling worden toegepast en nieuwe woorden uit de Woordenlijst Nederlandse taal worden toegevoegd.
  Beschikbaarheid: Distributie door de SDU.

  Meer informatie: `http://www.sdu.nl/`

- **TALO**

  Omschrijving: De TALO woordenboeken die de spellercorrectors ondersteunen beslaan een zo uitgebreid mogelijke reikwijdte van een gebruikt idioom, inclusief de

recentst gebruikte vorm van een taal en basiswoordenboeken die elke toegelaten woordvorm bevatten. Deze woordenboeken zijn complementair en zijn ook uitgebreid met woorden en uitdrukkingen die in specifieke domeinen, zoals bijvoorbeeld de juridische of commerciële sector, worden gebruikt.

De woordenboeken zijn niet standaard als ASCII-bestand beschikbaar. Wel kan het woordenboek via de software van Talo geraadpleegd worden. Een deel van de inhoud is bedoeld voor omspeldoeleinden zoals het omzetten van oude naar nieuwe spelling of van Van Dale naar Groene Boekje. Een ander deel betreft collocaties. Voor zowel het omspellen als het corrigeren van collocaties is de software van Talo nodig. Slechts in uitzonderingen kunnen de originele bestanden worden geleverd.

Meer informatie: `http://www.talo.nl/`

- **Van Dale Groot woordenboek der Nederlandse taal op cd-rom**

  Omschrijving: Deze Plusversie van Het Groot woordenboek der Nederlandse taal van Van Dale bevat ook een spellingcorrector voor MS-Word.

  Beschikbaarheid: Distributie door Van Dale.

  Meer informatie: `http://www.vandale.nl`

- **Words-L.**

  Omschrijving: Op de webpagina van WORDS-L worden een aantal woordenlijsten beschikbaar gesteld die kunnen worden gebruikt voor spellingcorrectie en het afbreken van woorden in combinatie met een aantal gangbare tekstverwerkers (Word, WordPerfect, Latex) en correctieprogramma's (ispell). Het initiatief is ontstaan uit onvrede over pakketten die door commerciële leveranciers worden aangeboden. Een collectief heeft zich vervolgens tot taak gesteld bestaande woordenlijsten (al dan niet beschikbaar in het publieke domein) te combineren, uit te breiden en te corrigeren. De site bevat ook een nuttig overzicht van bestaande pakketten en, met name, de tekortkomingen van verschillende producten.

  De drie woordenlijsten, zowel in ISO Latin1 code (latin1) als in de code gebruikt op IBM PC's (ibmpc437), bevatten respectievelijk 130623, 173762 en 222872 woorden. De eerste lijst (woorden.min) bevat uitsluitend woorden die voorkomen in het Elektronische Groene boekje. De tweede lijst bevat daarnaast ook de woordvormen, zoals *werkt* en *werkte*, die in de eerste lijst ontbreken. De derde en grootste lijst bevat ongeveer 50000 extra woorden. In deze lijst zijn de afbreekplaatsen aangegeven met een verhoogde punt. Voor spellingscontrole lijken de eerste en tweede lijst het meest geschikt. Eventueel kan ook de grootste lijst worden gebruikt maar dan moeten de daarin voorkomende afbreekstreepjes worden verwijderd.
  Beschikbaarheid: Distributie via de webpagina van Words-L.

  Meer informatie: `http://www.goddijn.com/words.htm`

Lijsten met terminologie

- **Buitenlandse aardrijkskundige namen in het Nederlands**

  Omschrijving: Deze lijst bevat de namen van landen met opgave van de daarbij behorende bijvoeglijke naamwoorden en inwoneraanduidingen, namen van hoofdsteden,

alsmede de Nederlandse vormen van namen van belangrijke steden, regio's en andere geografische entiteiten in het buitenland.

Beschikbaarheid: De lijst is nu nog een tekstbestand in Word. Het is de bedoeling dat het geheel binnenkort elektronisch beschikbaar wordt gesteld, óf via de website van de Taalunie, óf als onderdeel van het Taalunieversum.

Meer informatie: `http://www.taalunie.org` en www.taalunieversum.org

- **CoTerm**

  Omschrijving: In 1998 werd de Commissie Terminologie (CoTerm) ingesteld die de Nederlandse Taalunie moet bijstaan bij de voorbereiding en uitvoering van haar terminologiebeleid en die het Nederlandse taalgebied in internationale werkverbanden zal vertegenwoordigen (zie ook `http://www.taalunie.nl/_/werkt/terminologie.html#coterm`). De Commissie wordt belast met de inhoudelijke begeleiding van de drie terminologieprojecten van de Taalunie: NL-Translex, VIPTerm en het DOT-project. NL-Translex is een vertaalsysteem (zie hiervoor de paragraaf over vertaalcomponenten) en VIPTerm ('Virtueel InformatiePunt Terminologie') is een project dat als doel heeft alle beschikbare informatie en documentatie over terminologie binnen het Nederlandstalige gebied te inventariseren en te ontsluiten.

  - **DOT-project**

    Omschrijving: DOT staat voor 'Databank met OverheidsTerminologie'. Binnen het DOT-project wil men de verschillen tussen de in België en Nederland gebruikte overheidsterminologieën zo goed mogelijk in kaart brengen en de onderliggende conceptuele stelsels vergelijken. Dit onderzoek naar en de beschrijving van het gehele terrein van de Nederlandstalige overheidsterminologie zou uiteindelijk moeten uitmonden in een centrale databank voor overheidsterminologie. Het project omvat drie componenten:

    1. Het opstellen van het 'datamodel', het model voor de representatie van de concepten die door de termen worden gedekt. Omdat gestreefd wordt naar een 'multifunctionele' databank, bruikbaar in diverse toepassingen, voor verschillende gebruikersgroepen en in uiteenlopende software-omgevingen, moet bij het werk aan het datamodel met een groot aantal gegevens en parameters rekening gehouden worden.
    2. Het bouwen van een 'prototype' waarmee het 'datamodel' getoetst wordt. Het 'prototype' krijgt de vorm van een op zichzelf staande, in de praktijk bruikbare, termenbank. De gebruikersorganisaties die bij het project betrokken worden, leveren daarvoor teksten en termen aan en benutten het prototype in de dagelijkse praktijk bij hun activiteiten die omgang met terminologie vergen.
    3. Het verder ontwikkelen van software voor het automatisch verzamelen en rubriceren van termen (termextractie).

    Beschikbaarheid: De componenten zijn nog niet beschikbaar.

    Meer informatie: `http://www.taalunie.nl/_/werkt/terminologie.html#overheidsterminologie`

- **E-ANS lijst van aardrijkskundige namen en afleidingen daarvan**

Omschrijving: Deze lijst bevat de namen van aardrijkskundige eenheden, het daarvan afgeleide onverbogen adjectief en de mannelijke inwonersnaam in het enkelvoud. Bijvoorbeeld: *Costa Rica - Costaricaans - Costaricaan.*

Meer informatie: `http://www.kun.nl/e-ans`

- **Polderland woordenlijsten met vaktermen**

  Omschrijving: Polderland biedt woordenlijsten aan voor de volgende vakgebieden: juridisch; medisch; zakelijk-bestuurlijk; sociaal-maatschappelijk en technisch-wetenschappelijk.

  Meer informatie: `http://www.polderland.nl/systemen/index.html`

Woordenlijsten met taalkundige informatie

- **Atranos**

  Omschrijving: Binnen het Atranos-project is door het Centrum voor Computerlingustiek in Leuven (CCL) een lexicon samengesteld dat een 36.000-tal woorddelen bevat.

  Beschikbaarheid: Het lexicon kan aangevraagd worden door een email te zenden aan Vincent Vandeghinste (vincent.vandeghinste@ccl.kuleuven.ac.be).

  Meer informatie: `http://atranos.esat.kuleuven.ac.be/`

- **Carp Technologies**

  Omschrijving: Voor hun taalverwerkende software maakt Carp Technologies onder andere gebruik van zelf ontwikkelde digitale lexica die zijn samengesteld uit een groot aantal bronnen. Hierbij wordt gebruikgemaakt van intelligente morfologische algoritmen. De lexica zijn ook apart te koop. Ze bevatten doorgaans naast lemma's ook alle woordvormen met bijbehorende categorieën en kenmerken. Carp Technologies heeft lexica voor het Nederlands, Engels en Duits (40.000 tot 500.000 woordvormen). Daarnaast heeft Carp Technologies de Engelse versie van WordNET met behulp van intelligente statistische technieken automatisch vertaald naar een Nederlandse versie. Tenslotte kan men nog een groot aantal bilinguale lexica leveren.

  Meer informatie: `http://www.carp-technologies.nl/low/nl/digitalelexica.html`

- **CELEX**

  Omschrijving: CELEX (Centre for Lexical Information) heeft elektronische databases ontwikkeld die verschillende types van lexicale informatie over het hedendaagse Nederlands, Engels en Duits bevatten. Het Nederlandse deel (400.000 woordvormen) is voornamelijk afgeleid uit het INL 50 miljoen woorden corpus. Er is gedetailleerde informatie beschikbaar over de orthografie, fonologie, morfologie (flexie en derivatie), syntaxis en woordfrequentie. Ook bevat CELEX informatie over syntactische en semantische subcategorisatie en over valentie. Voor de homografen zijn de woordfrequenties gedisambigueerd op basis van een 42.4 miljoen woorden tellend corpus van het INL. Een belangrijke eigenschap aan de databases is dat alle informatie gerepresenteerd is om tegemoet te komen aan de formele en strikte voorwaarden voor computationele toepassingen.

Beschikbaarheid: De CD-ROM wordt verkocht via de Linguistic Data Consortium. Het materiaal is ook on line beschikbaar via `http://www.mpi.nl/world/celex`. Voor CELEX kan momenteel geen enkele ondersteuning worden geleverd.

Meer informatie: `http://www.icp.inpg.fr/ELRA/cata/text_det.html#celex`, `http://www.kun.nl/celex/` en `http://www.ldc.upenn.edu/`

- **CGN-lexicon**

  CELEX en het RBN-lexicon zijn inmiddels samengevoegd en vormen samen het CGN-lexicon. Uiteindelijk zal het CGN-lexicon in de 'breedte' (aantal kenmerken) en in de 'lengte' (aantal woorden) helemaal toegesneden zijn op het CGN-corpus (zie het gedeelte over tekstcorpora). Zo zullen alle woorden uit CELEX en RBN die niet in het corpus voorkomen en alle kenmerken die niet relevant worden geacht voor het corpus worden verwijderd. Het CGN-lexicon zal zo dus de spelling hanteren die is afgesproken binnen het CGN (die afwijkt van het Groene Boekje) en alleen morfologie, syntax e.d. weergeven voor zover dat te koppelen is aan het CGN.

  Beschikbaarheid: Waarschijnlijk vanaf april 2002 op de cd-roms van het CGN-corpus. Projectmedewerkers kunnen het lexicon ook nu al krijgen.

- **Compound dictionary Parlevink**

  Omschrijving: Dit lexicon bestaat uit een lijst van zo'n 350.000 samenstellingen met een 'vertaling' in constituenten erachter. De lijst is gemaakt op basis van een lijst met samenstellingen van TNO aangevuld met samenstellingen uit 150 miljoen woorden krantenmateriaal.

  Meer informatie: `http://parlevink.cs.utwente.nl`

- **Contrastive verb valency dictionary**

  Omschrijving: De interdepartementale onderzoeksgroep CONTRAGRAM van de Universiteit Gent ontwikkelt momenteel een contrastief woordenboek m.b.t. de werkwoordenvalenties van het Nederlands, het Frans en het Engels. De *contrastive verb valency dictionary* kan voor taaltechnologische doeleinden gebruikt worden. Wel moet er dan eerst een omzetting gebeuren naar een databankformaat. Deze omzetting kan volgens de makers echter eenvoudig worden uitgevoerd.

  Meer informatie: `http://bank.rug.ac.be/contragram/cvvd.htm`

- **Eurodicautom**

  Omschrijving: Eurodicautom is de multilinguale terminologische database van de *European Commission's Translation Service*. De database bevat informatie voor twaalf talen en wordt constant bijgewerkt. Hoewel de kern betrekking heeft op onderwerpen uit de Europese Unie, beslaat de database verder een breed spectrum aan menselijke kennis. De database bevat technische termen, afkortingen, acroniemen en idioom. Het aantal Nederlandse ingangen is 504440.

  Beschikbaarheid: Via de website: `http://eurodic.ip.lu/cgi-bin/edicbin/EuroDicWWW.pl`

  Meer informatie: `http://eurodic.ip.lu/cgi-bin/edicbin/EuroDicWWW.pl`

- **EuroWordNet**

  EuroWordNet is een multilinguale lexicale database met *wordnets* voor verschillende Europese talen (Nederlands, Italiaans, Spaans, Duits, Frans, Tsjechisch en Estisch). De *wordnets* zijn gestructureerd op dezelfde manier als de Amerikaanse *wordnets* voor het Engels (Miller et al 1990): ze bevatten *synsets* (sets van synoniemen) met semantische relaties ertussen. De database kan onder andere worden gebruikt voor monolinguale en cross-linguale *information retrieval*.

  Omvang: Ruim 40.000 woorden.
  Beschikbaarheid: Een licentie voor de *wordnets* kan bij ELDA/ELRA verkregen worden.

  Formaat: Het EuroWordNet-formaat wordt gedefinieerd door de EuroWordNet database editor Polaris. De specificaties staan in de user-manual van de database.

  Meer informatie: `http://www.hum.uva.nl/~ewn/` en `http://www.icp.grenet.fr/ELRA/home.html`

  Referentie: Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, en K. J. Miller, (1990).

  Introduction to wordnet: an on line lexical database. *International Journal of Lexicography* 3(2): 235 - 244.

- **ILK Lexicon**

  Omschrijving: Binnen het project ILK Corpus van de onderzoeksgroep *Induction of Linguistic Knowledge (ILK)* van de KUB (zie het gedeelte over tekstcorpora) wordt ook een lexicon ontwikkeld. Het is de bedoeling dat het lexicon het Nederlands goed afdekt, en dat het bovendien inductief is. De omvang is 1.273.128 unieke woordvormen en andere *tokens.*

  Het lexicon bevat de spelling van woordvormen, morfologie en syntactische klasse voor een kernlexicon dat is afgeleid uit CELEX. Tevens bevat het informatie - afgeleid uit het tekstcorpus - over frequentie en n-grammen. In het lexicon is rekening gehouden met de mogelijkheid de inhoud van de lege cellen in de spreadsheet af te leiden uit de gevulde cellen met machine-learningmethodes.

  Meer informatie: `http://ilk.kub.nl/`

- **L&H**

  Omschrijving: L& H biedt meer dan 100 elektronische referentiewerken aan waaronder woordenboeken, thesauri en encyclopedieën.

  Meer informatie: `http://www.lhsl.com/`

- **LanTmark (L0004)**

  Omschrijving: Het lexicon LanTmark heeft de volgende samenstelling: nomina (50.000), verba (7.000), adjectieven (6.000) en adverbia (1.000). Elk lemma bevat morfologische, syntactische en semantische informatie.

  Omvang: 64000 ingangen.
  Formaat: ASCII met een ISO 8859-1 karakterset.

  Meer informatie: `http://www.icp.grenet.fr/ELRA/cata/text_det.html#lantdutch`

- **8.1.4 PAROLE-lexicon**

  Omschrijving: Naast corpora (zie het gedeelte over tekstcorpora) heeft het PAROLE-project ook een lexicon voortgebracht van 20.200 ingangen die voorzien zijn van morfosyntactische en syntactische informatie: 12.000 zelfstandige naamwoorden, 3000 werkwoorden, 3000 adjectieven, 500 bijwoorden en 1500 ingangen met een andere woordsoortcategorie. Het Nederlandse lexicon, in SGML-formaat, is medio 1999 beschikbaar gekomen op cd-rom. Voor onderzoekers en onderzoeksgroepen in Nederland en België, die het lexicon uitsluitend voor niet-commerciële onderzoeksdoeleinden willen gebruiken, is het verkrijgbaar bij het INL tegen een gereduceerde prijs van 200 ECU (Hfl. 440,-). Het PAROLE-lexicon wordt ook gedistribueerd door ELRA/ELDA onder productnummer ELRA-L0031.

  Het project SIMPLE dat in augustus 1998 is gestart, beoogt de toevoeging van semantische informatie aan de set PAROLE-lexica. De typen semantische informatie zijn geselecteerd vanuit het perspectief van de relevantie voor taaltechnologische toepassingen.

  Meer informatie: `http://www.inl.nl/corp/parole.htm`
  `http://www.inl.nl/europa/projecten.htm`
  `http://www.ub.es/gilcub/SIMPLE/simple.html` en
  `http://www.icp.inpg.fr/ELRA/cata/text_det.html#dutparollex`

  Zie voor een omschrijving van de inhoud van het lexicon de documentatie op `http://www.inl.nl/PAROLE/doc_A1.html` en `http://www.inl.nl/PAROLE/doc_A2.html`.

- **PROTON**

  Omschrijving: Binnen het project PROTON van het Centrum voor Computerlinguïstiek (CCL) van de KU Leuven zijn twee databases opgebouwd die data bevatten over werkwoordsvalentie, een voor het Frans met 8500 ingangen voor 3700 werkwoorden en een voor het Nederlands met 6299 ingangen voor 4200 werkwoorden. Beide databases bevatten morfologische en syntactische informatie.

  Meer informatie: `http://www.ccl.kuleuven.ac.be/about/PROTON.html`

- **RBN**

  Omschrijving: Het Referentiebestand Nederlands is een multifunctionele lexicale databank met informatie met betrekking tot morfologie, syntaxis, combinatoriek, semantiek en pragmatiek, welke op een expliciete, formele wijze is weergegeven in de vorm van feature-value paren (45.000 lemmata). De valentiedata die het RBN bevat worden momenteel omgezet naar de CGN-standaard.

  Beschikbaarheid: Alleen toegankelijk na het nodige overleg. Het huidige bestand kan draaiend onder de editor OMBI worden aangeboden.

  Omvang: 45.000 trefwoorden

  Meer informatie: `http://www.taalunie.org/_/werkt/rbn.html`

  Referentie: - Vliet, Hennie van der (te verschijnen). Het Referentiebestand Nederlands. *Trefwoord*, 1999.

- **TRIPTIC**

  Omschrijving: Triptic is een parallelle lexicale database met preposities. Deze werd ontwikkeld voor de contrastieve analyse van preposities in het Engels, Frans en Nederlands. Voor elke geanalyseerde prepositie werd een ruime context toegevoegd: 1) de zin waarin de prepositie voorkomt en 2) de paragraaf waarin die zin voorkomt. Hiervoor werden uit het Namur-corpus (zie het gedeelte over tekstcorpora) een aantal voorbeeldzinnen met vertaling geselecteerd en in de databank geplaatst. Alle onderzochte preposities zijn zowel syntactisch als semantisch geannoteerd. De 'context' van de preposities is gelemmatiseerd om opzoekwerk te vergemakkelijken.

  Beschikbaarheid: De beschikbaarheid is momenteel onduidelijk.

  Meer informatie: `http://bank.rug.ac.be/contragram/newslet3.html#FORUM`

8.1.5 Evaluatie en conclusies

Het Nederlands beschikt over lexicale databanken die vergelijkbaar of zelfs beter zijn dan wat er beschikbaar is voor andere talen, met name op het gebied van woordenlijsten voor spelling en afbreken, morfologische informatie, en syntaxis (valentie) en semantiek.

Lexicale informatie die speciaal gericht is op de behoeften van *information retrieval* en verwante gebieden lijkt minder algemeen voorhanden. Verschillende groepen hebben lexicale informatie verzameld voor analyse van samenstellingen, en voor het accuraat herkennen van namen van personen en organisaties en voor het herkennen van geografische aanduidingen. Er lijkt een redelijk grote behoefte te bestaan aan met name dit soort van informatie bij een aantal groepen. Het is daarom wellicht te overwegen om hiervoor een goede en uitgebreide database te ontwikkelen en beschikbaar te maken.

8.2 Multilinguale lexica

Multilinguale lexica vormen de basis van applicaties waarin iets vertaald moet worden, zoals natuurlijk automatische vertaalsystemen ('machine translation'), maar ook bijvoorbeeld systemen voor het zoeken in anderstalige informatie ('cross-language information retrieval'). Dit overzicht beperkt zich tot woordenboeken die informatie geven over vertaalrelaties van en/of naar het Nederlands en die in elektronische vorm beschikbaar zijn. Multilinguale lexica komen doorgaans tot stand door middel van zorgvuldige handmatige arbeid, maar ze kunnen ook automatisch worden afgeleid uit parallelle corpora (een of meer teksten die in minstens twee talen beschikbaar zijn). Dit hoofdstuk bevat daarom naast multilinguale lexica ook verwijzingen naar parallelle corpora en naar software voor de automatische extractie van multilinguale lexica.

- **Ergane**

  Ergane is een gratis meertalig woordenboekprogramma voor Windows dat is ontwikkeld door Gerard van Wilgen. Het kan niet alleen gebruikt worden als vertaalwoordenboek, maar ook als hulpmiddel om studenten te helpen met het leren van de woorden van vreemde talen. Verder is het met dit programma mogelijk om woordenlijsten te genereren. Door het gebruik van de kunstmatige taal Esperanto als een interlingua, wordt voorkomen dat er aparte woordenlijsten voor elk taalpaar nodig zijn. De woordenboeken bevatten lexicale informatie zoals woordsoort en homonymierelaties.

Algemene eigenschappen: Besturingssysteem Windows. Ergane ondersteunt vertaling van en naar maar liefst 57 talen, hoewel van sommige talen maar enkele vertalingen aanwezig zijn. Het Nederlandse lexicon heeft 56.000 ingangen, Portugees, Engels en Duits ruim 15.000 en Frans ruim 10.000. Opvallend zijn de lexica voor Fries, Afrikaans en Papiamento, elk rond de 5.000 ingangen groot. Domeinwoorden uit het domein reizen hebben voor sommige talenparen bijzondere aandacht gekregen.

Beschikbaarheid: Ergane kost niets en is beschikbaar via `http://www.travlang.com/Ergane/frames-nl`. De ermee te genereren woordenlijsten zijn vrij van auteursrecht en kunnen worden gekopieerd, gedistribueerd en veranderd zonder wettelijke beperkingen. Men kan ze op iedere wijze gebruiken die men wil, zelfs voor commerciële doeleinden. Gebruikers hebben het recht het programma te modificeren en het te gebruiken als basis voor de ontwikkeling van eigen software. Ook hier zijn geen wettelijke beperkingen van toepassing.

- **FreeDict**

  Freedict levert woordenboeken voor een groot aantal taalparen, met name van en naar Engels. De freedict programma's zijn ontwikkeld voor Windows, maar woordenlijsten worden apart bijgeleverd.

  Algemene eigenschappen: Besturingssysteem Windows 95/98/NT4 Ja V2.5 Windows 3.1. Van en naar Nederlands zijn er woordenboeken beschikbaar voor Engels, Frans, Duits, Spaans, Italiaans en Portugees. De lexica hebben tussen de 2.000 en 14.000 ingangen.

  Beschikbaarheid: Freedict is gratis en is beschikbaar als freeware via `http://www.freedict.com/woordenboek`. De software en woordenlijsten zijn waarschijnlijk vrij van auteursrecht.

- **VLIS: Van Dale Lexicaal Informatiesysteem**

  De VLIS database van Van Dale Lexicografie is een relationele database die alle lexicale kennis bevat die gebruikt wordt voor het uitgeven van de Van Dale vertaalwoordenboeken. De database is gebaseerd op Nederlandse ingangen met vertalingen naar equivalente lemma's in de vreemde talen. Het aantal lemma's in de vreemde talen ligt tussen de 180.000 en 300.000. Het aantal Nederlandse ingangen is (door synonymen) zelfs nog groter.

  Algemene eigenschappen: Oracle database. Nederlands naar vreemde taal (Duits, Engels, Frans, Spaans en Italiaans).

  Beschikbaarheid: De Van Dale VLIS database is niet zonder meer beschikbaar voor derden. De database vormt wel de basis van commerciële vertaal- en retrievalsoftware die geleverd wordt door Irion Technologies in Delft (`http://www.irion.nl`).

8.2.1 Automatische extractie van multilinguale lexica

Automatische extractie van multilinguale lexica uit parallelle corpora doorloopt doorgaans drie stappen. In de eerste stap worden de tokens (meestal de woorden, maar soms ook bijv. nominale constituenten) en de zinnen geïdentificeerd. In de tweede stap worden corresponderende zinnen met elkaar verbonden, wat niet triviaal is omdat soms bijvoorbeeld twee zinnen in de brontaal naar één zin in de doeltaal vertaald kunnen zijn. In de derde en laatste stap

worden de woorden uit de parallelle zinnen met elkaar verbonden. Algoritmes voor zins- en woordverbinding (ook wel aangeduid als alignment) worden beschreven door [60, 28, 69].

Beschikbaarheid:

De broncode van het zinsverbindingsprogramma van Gale en Church (1993) is beschikbaar als bijlage bij het artikel. Het woordverbindingsprogramma van Hiemstra (1997) is vrij beschikbaar via het web (`http://www.cs.utwente.nl/~irgroup/align/`)

8.2.2 Parallelle corpora

Parallelle corpora kunnen als bron dienen voor de extractie van multilinguale lexica, zoals hierboven beschreven. Deze sectie introduceert kort enkele bronnen waarmee op een goedkope manier een parallel corpus zou kunnen worden samengesteld. Vermeldenswaardig is een methode voor het automatisch zoeken van parallelle pagina's op het web beschreven door [103].

Beschikbaarheid: Parallelle corpora waarbij het Nederlands een van de talen is zijn met enige moeite bij elkaar te sprokkelen via bijvoorbeeld de website van de Europese Unie (`http://europa.eu.int`) of die van de Belgische overheid (`http://www.gov.be`). Op beide sites zijn de documenten in meer dan een taal beschikbaar. Het is niet geheel duidelijk of deze documenten vrij zijn van auteursrecht, maar multilinguale lexica die ermee geconstrueerd worden waarschijnlijk wel. Een Engels-Nederlands webcorpus bestaande uit bijna 3000 parallelle webpagina's werd door Wessel Kraaij (zie [70] gevonden met behulp van de methode van [103]. Dit corpus is beschikbaar voor onderzoeksdoeleinden. Voor inlichtingen: Wessel Kraaij (TNO TPD, Delft; email: kraaij@tpd.tno.nl).

8.3 Tekstcorpora

Taaltechnologie maakt intensief gebruik van informatie die kan worden afgeleid uit tekstcorpora. Informatie over de woordfrequenties, frequentie van letter- of woordcombinaties, etc., is nuttig voor allerlei toepassingen, en kan worden verkregen op basis van ruwe, niet geannoteerde, tekst. Een corpus dat is voorzien van extra annotatie, bijvoorbeeld voor woordsoort, zins- en constituentgrenzen, relaties tussen constituenten, woordbetekenis, etc., kan worden gebruikt om statistische modules voor woordsoortdisambiguatie, syntactische disambiguatie of woordbetekenisdisambiguering af te leiden.

8.3.1 State of the art internationaal

Corpora zijn de afgelopen jaren steeds omvangrijker geworden. Het British National Corpus omvat zo'n 100 miljoen woorden, en is zorgvuldig samengesteld uit verschillende tekstsoorten.

Het bekendste geannoteerde corpus is de Penn Treebank, waarvan onder andere het Wall Street Journal corpus deel uitmaakt. In dit corpus (omvang ongeveer 1 miljoen woorden) is het materiaal voorzien van syntactische annotatie (constituenten voorzien van syntactische categorie).

Een aantal corpora zijn ontwikkeld met het oog op evaluatie. Het Senseval-corpus bevat materiaal waarin woorden zijn geannoteerd op betekenis (op basis van WordNet), en is gebruikt om verschillende systemen voor *word sense disambiguation* te testen. Voor conferenties als MUC en TREC worden omvangrijke datasets beschikbaar gemaakt, om bijvoorbeeld systemen voor *cross language information retrieval* of *named entity recognition* te evalueren.

Voor toepassingen op het gebied van automatisch vertalen zijn parallelle corpora beschikbaar, die dezelfde tekst in verschillende vertalingen bevatten.

8.3.2 Evaluatiecriteria

- Omvang van het corpus,
- Inhoud van het corpus (een bepaalde bron, tekstsoort (nieuws), of een gebalanceerd corpus),
- Geannoteerd of niet, aard van de annotatie (woordsoort, syntaxis, etc.),
- Aansluiting bij een internationale standaard,
- Formaat (tekst, database, XML, etc.),
- Beschikbaarheid,
- Documentatie

8.3.3 Inventarisatie geannoteerde corpora

- **ANNO**

  Omschrijving: Het ANNO-corpus (*een publieke, geannoteerde gegevensbank voor het Nederlands*) werd ontwikkeld in het kader van het Vlaams korte termijnprogramma Spraak- en Taaltechnologie voor het Nederlands (STTN). Gezien de aard van dit programma, dat de nadruk legt op spraaktechnologie, is gekozen voor een corpus dat dicht aansluit bij de spreektaal. Het corpus bestaat uit de tekst van BRTN-radio nieuwsuitzendingen en uitzendingen van Actueel. De transcripties van interviews binnen die uitzendingen betreffen spontane uitingen. Er is een kleine demo beschikbaar die is gebaseerd op een enkel bestand.

  Meer informatie: [117] en `http://www.ccl.kuleuven.ac.be/about/ANNO.html`

- **CGN**

  Omschrijving: Doel van het project *Corpus Gesproken Nederlands* is het opbouwen van een geannoteerd corpus met ca. 10 miljoen woorden gesproken Nederlands, waarvan tweederde deel afkomstig is uit Nederland, en eenderde uit Vlaanderen.

  Meer informatie: `http://lands.let.kun.nl/cgn/home.htm` en `http://www.elda.fr/index.html`

- **CHILDES**

  Omschrijving: Het Child Language Data Exchange System (CHILDES) is een internationale faciliteit van de Carnegie Mellon University (Pittsburgh). Het systeem is ontwikkeld met als doel de studie naar het leren van taal door kinderen en volwassenen te vergemakkelijken. Het systeem bevat een grote database met conversationele interacties, een systeem om deze interacties te transcriberen en een aantal programma's om de data te analyseren. Het Nederlands deel van dit corpus bevat uitsluitend kindertaal.

  Meer informatie: [92] en `http://atila-www.uia.ac.be/childes/`

- **Eindhoven (Uit den Boogaart) corpus (1975)**

  Omschrijving: Dit corpus werd opgesteld om een nauwkeurig idee te krijgen van in Nederland (niet in Vlaanderen) veel voorkomend taalgebruik d.m.v. frequentietellingen van woorden. Het onderzoek werd uitgevoerd door de Werkgroep Frequentie-Onderzoek van het Nederlands, gesubsidieerd door Z.W.O. (het Nederlandse Fonds voor Zuiver Wetenschappelijk Onderzoek, nu N.W.O.) en de Technische Hogeschool Eindhoven (geschreven taal) en het Instituut voor Dialectologie, Volks- en Naamkunde van de Koninklijke Nederlandse Academie voor Wetenschappen te Amsterdam (nu: Meertensinstituut) (gesproken taal). Het geschreven deel bevat fragmenten van in totaal 600.000 woorden uit de periode 1964-1971. Het gesproken deel (ook wel aangeduid als "Corpus De Jong") is aanzienlijk kleiner. Oorspronkelijk in 1975 in boekvorm verschenen als *Woordfrequenties in Geschreven en Gesproken Nederlands* [129, 44].

- **ILK-corpus**

  Omschrijving: Het ILK-corpus bestaat uit een verzameling krantenartikelen. De artikelen zijn voor een gedeelte afkomstig uit het archief van de voormalige VNU Dagbladenunie (nu onderdeel van Wegener). Daarnaast bevat het corpus materiaal van de websites van twee Nederlandse kranten. Laatstgenoemde verzameling is verkregen met behulp van een webrobot. Het corpus is ontwikkeld binnen en door de onderzoeksgroep *Induction of Linguistic Knowledge(ILK)* van de KUB.
  Meer informatie: `http://ilk.kub.nl/`

- **INL-Corpora**

  - **5 miljoen woorden corpus 1994**

    Omschrijving: Algemeen Nederlands uit de periode 1989-1994. Ongeveer 250 tekstbronnen (periodieken en boeken), geclassificeerd naar publicatiemedium en onderwerp. (Dit corpus heeft een andere samenstelling dan het niet-geannoteerde ECI/MCI corpus.)
    Meer informatie: `http://www.inl.nl/corp/corp.htm`

  - **27 miljoen woorden corpus 1995**

    Omschrijving: 27 miljoen woorden corpus 1995 (NRC-corpus)

    Meer informatie: `http://www.inl.nl/corp/corp.htm`

  - **38 miljoen woorden corpus 1996**

    Omschrijving: Gevarieerde samenstelling met 3 hoofdcomponenten: krantenteksten (1992-1995), juridische component (1814-1989), gevarieerd samengestelde component (1970-1995). De teksten zijn geclassificeerd volgens onderwerp en publicatiemedium.

    Meer informatie: `http://www.inl.nl/corp/corp.htm`

  - **50 miljoen woorden corpus**

    Omschrijving: Algemeen Nederlands uit de periode 1970-1990. Ongeveer 1500 tekstbronnen ontleend aan ruim 400 boeken over gevarieerde onderwerpen.

    Meer informatie: Kruyt (1995) op de INL-website: `http://www.inl.nl`

  - **Geïntegreerde Taalbank 8ste - 21ste Eeuws Nederlands**

Omschrijving: Sinds juli 1999 wordt de *Geïntegreerde Taalbank 8 ste −21 ste Eeuws Nederlands* voorbereid. De *Geïntegreerde Taalbank* wordt een flexibel raadpleegbare databank van het vroegste tot het modernste Nederlands. Die databank zal diverse soorten taaldata bevatten: elektronische woordenboeken (o.a. VMNW,MNW en WNT), tekstbestanden en bestanden met taalkundige gegevens (lexica).

Meer informatie: `http://www.inl.nl/taalbank/index.htm`

– **PAROLE Distributable corpus (3 miljoen woorden)**

Omschrijving: Het INL heeft in het kader van het Europese project PAROLE (*Preparatory Action for Linguistic Resources Organization for Language Engineering*) een PAROLE-lexicon en een PAROLE-corpus ontwikkeld, volgens richtlijnen die ook zijn toegepast in de PAROLE-lexica en PAROLE-corpora van ruim tien andere West-Europese talen. Het PAROLE Distributable Corpus is een deelverzameling van het PAROLE Reference corpus.

Meer informatie: `http://www.inl.nl/corp/parole.htm`

– **PAROLE Reference corpus (20 miljoen woorden)**

Omschrijving: Het PAROLE Reference corpus is tot stand gekomen in het kader van het grootschalige Europese corpusproject PAROLE (MLAP/LE2-4017).

Meer informatie: `http://www.inl.nl/corp/parole.htm`

8.3.4 Evaluatie geannoteerde corpora

- **ANNO**

Omvang van het corpus: ruim 640.000 woorden.

Geannoteerd of niet, aard van de annotatie: Het gehele corpus is voorzien van morfosyntactische en fonetische annotaties. Deze annotatie is automatisch aangebracht (m.b.v. de WOTAN -tagger en Treetalk grafeem-foneemomzetter), en deels gecorrigeerd. Verder zijn delen van de tekst geannoteerd met discourse-informatie. Ook zijn er delen met morfologische informatie geannoteerd door gebruik te maken van het KEPER-systeem. Tenslotte zijn er delen geannoteerd met door het METAL-systeem aangebrachte syntactische informatie.

Beschikbaarheid: 1997: "Het corpus zal beschikbaar gesteld worden zodra de auteursrechtelijke kwesties zijn geregeld."

Inhoud van het corpus: Het corpus sluit dicht aan bij de spreektaal. Het bestaat uit de tekst van BRTN-radio nieuwsuitzendingen en uitzendingen van Actueel. De transcripties van interviews binnen die uitzendingen betreffen spontane uitingen.

Aansluiting bij een internationale standaard:

Formaat: Teksttypologie en tekststructuur zullen worden geformaliseerd in een SGML *Document Type Description* .

Documentatie:

- **CGN**

Omvang van het corpus: 10 miljoen woorden (zo'n duizend uren spraak).

Geannoteerd of niet, aard van de annotatie: Al het materiaal wordt orthografisch getranscribeerd, terwijl er tevens een oplijning plaatsvindt waarbij de orthografische transcriptie gekoppeld wordt 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 brede fonetische transcriptie wordt vervaardigd, er een geverifieerde oplijning 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. Bij de derde release is 50.000 woorden aan syntactisch geannoteerde data beschikbaar, uiteindelijk moet een kerncorpus van 1 miljoen woorden syntactisch worden geannoteerd. De CGN-annotatie stelt zich neutraal op wat betreft theoriegebonden oplossingen voor het op elkaar betrekken van constituentstructuur en dependentierelaties. De syntactische annotatie is semi-automatisch aangebracht.

Beschikbaarheid: Geheel beschikbaar in 2003. Maar tussentijds komen er ook al delen beschikbaar. Distributie van tussentijdse releases gebeurt door ELRA/ELDA in opdracht van de Nederlandse Taalunie, eigenaar van het CGN. Het corpus wordt beschikbaar gesteld voor wetenschappelijk onderzoek en voor de ontwikkeling van commerciële producten.

Inhoud van het corpus: Gesproken Nederlands. Tweederde deel afkomstig uit Nederland, en eenderde uit Vlaanderen.

Aansluiting bij een internationale standaard: Binnen het project is een eigen CGN-tagset gedefinieerd die ca. 300 tags omvat en die aansluit bij de praktijk van de ANS [54]. De tagset is conform de EAGLES-richtlijnen.

Formaat: Data zijn gecodeerd in XML-formaat.

Documentatie: Documentatie wordt elektronisch meegeleverd.

- **CHILDES**

  Omvang van het corpus: De database bevat momenteel data van een gering aantal Nederlandse kinderen. Deze data zijn in 7 verschillende projecten verzameld.

  Geannoteerd of niet, aard van de annotatie: Sommige van de databases uit CHILDES bevatten ook morfologische en/of syntactische annotatie. Deze annotatie wordt aangebracht op een aparte regel. Het formaat van de annotatie heet CHAT. CHAT is speciaal voor CHILDES ontwikkeld, met bijzondere aspecten van eerstetaalverwerving als uitgangspunt.

  Beschikbaarheid: Het corpus is verkrijgbaar op de CHILDES-site in Antwerpen en bij de initiatiefnemers in Pittsburgh, Carnegie Mellon. Voor onderzoeksdoeleinden is het corpus ook toegankelijk op het Nijmeegse Max Planck Instituut (MPI). De meest recente cd-versie van CHILDES is de editie van april 1998.

  Inhoud van het corpus: Het gaat om longitudinale data: de kinderen zijn met een regelmatige interval voor een periode van een paar jaar getapet. De spraak is voornamelijk spontaan.

  Aansluiting bij een internationale standaard: Nee, eigen CHAT-formaat ontwikkeld.

Formaat: ASCII met annotatie volgens CHAT.

Documentatie: Documentatie is beschikbaar op de CD.

- **Eindhoven (Uit den Boogaart) corpus (1975)**

  Omvang van het corpus: Ca. 600.000 woorden geschreven, ca. 120.000 woorden gesproken taal.

  Geannoteerd of niet, aard van de annotatie: Woordsoort en flexievorm.

  Beschikbaarheid: Op verschillende instituten, Taal & Spraak KUN, Max Planck Instituut (CELEX), Systeemhuis TU Eindhoven, Mathematisch Centrum Amsterdam, is een versie van het corpus aanwezig; het is onduidelijk of er copyright op het corpus rust. Waarschijnlijk is dit niet het geval voor wetenschappelijk gebruik.

  Inhoud van het corpus: Geschreven en getranscribeerd gesproken Nederlands uit de perioden 1964-1971 en 1960-1973.

  Aansluiting bij een internationale standaard:

  Formaat:

  Documentatie:

- **ILK-corpus**

  Omvang van het corpus: Het corpus bevat meer dan 230,000 artikelen en 120 miljoen woorden.

  Geannoteerd of niet, aard van de annotatie: Het geannoteerde deel bestaat uit 10.000 zinnen van januari 1998 van het Brabants Dagblad. Het merendeel van de annotatie is met de hand aangebracht. Er is annotatie voor de volgende domeinen: 'prosodie' (met name, de positie van *breaks* en accenten) in gesynthetiseerde spraak, "named-entity recognition and labelling" en 'base-NP chunking'.

  Beschikbaarheid: Het corpus is verkregen van Wegener Dagbladen onder de voorwaarde dat niets van het materiaal mag worden gereproduceerd. Dat blokkeert vooralsnog publiek gebruik. Meer duidelijkheid hierover moet nog volgen.

  Inhoud van het corpus: Materiaal van Wegener Dagbladen: ANDA, Brabants Dagblad, Eindhovens Dagblad, Gelderlander, De Groene Amsterdammer op het web, en Volkskrant op het web.

  Aansluiting bij een internationale standaard: Annotatie volgens Text Encoding Initiative (TEI).

  Morfo-syntax: voor wat betreft tagging, lemmatisering en syntactische annotatie is geprobeerd zoveel mogelijk de CGN-manuals en tagsets te gebruiken, met aanpassingen voor gesproken Nederlands. Daarnaast wordt parallel ook nog altijd de WOTAN-tagset uit Nijmegen gebruikt.

  Prosodie: hiervoor wordt een eigen weg gevolgd die redelijk dicht bij de CGN-annotatie voor prosodie ligt.

  Named entities: hiervoor wordt een eigen weg gevolgd.

  Formaat: In principe worden alle data opgeslagen als kolomsgeordende ASCII-bestanden (vergelijkbaar met CELEX). Het materiaal kan ook te allen tijde als matrix in een database worden ingevoerd.

Documentatie: De documentatie is in ontwikkeling en zal nog niet publiek worden gemaakt.

- **INL-Corpora**

  - **5 miljoen woorden corpus 1994**

    Omvang van het corpus: 5 miljoen woorden.

    Geannoteerd of niet, aard van de annotatie: Verrijkt met trefwoord (lemmatisering) en DutchTale woordsoort (enkelvoudige woordsoortcategorieën). Woordsoortcodes en lemmavormen zijn automatisch toegekend. Op het materiaal zijn nauwelijks extra correctieslagen uitgevoerd. Dit geldt zowel voor de teksten zelf als voor de linguïstische gegevens woordsoort en lemma.

    Beschikbaarheid: Via internet raadpleegbaar corpus, d.m.v. retrievalprogramma. De restricties in gebruik zijn opgelegd door copyright en contractbepalingen met de tekstleveranciers. De toegang tot het corpus is gratis voor academisch onderwijs en voor niet-commerciële onderzoeksdoeleinden.

    Inhoud van het corpus: Algemeen Nederlands, 1989-1994, 17 tekstbronnen, geclassificeerd naar publicatiemedium en onderwerp.

    Aansluiting bij een internationale standaard:

    Formaat:

    Documentatie:

  - **27 miljoen woorden krantencorpus 1995**

    Omvang van het corpus: 27 miljoen woorden.

    Geannoteerd of niet, aard van de annotatie: Verrijkt met trefwoord en DutchTale woordsoort (enkelvoudige woordsoortcategorieën).

    Beschikbaarheid: Via internet raadpleegbaar corpus, d.m.v. retrievalprogramma. De restricties in gebruik zijn opgelegd door copyright en contract bepalingen met de tekstleveranciers. De toegang tot het corpus is gratis voor academisch onderwijs en voor niet-commerciële onderzoeksdoeleinden.

    Inhoud van het corpus: Materiaal uit het NRC

    Aansluiting bij een internationale standaard:

    Formaat:

    Documentatie:

  - **38 miljoen woorden corpus 1996**

    Omvang van het corpus: 38 miljoen woorden.

    Geannoteerd of niet, aard van de annotatie: De teksten zijn automatisch taalkundig verrijkt met een lemma (trefwoordvorm) en twee woordsoorttoekenningen: een globale (13 woordsoortcategorieën) en een verfijnde (met subcategorisatie) conform de MECOLB standaard. Er zijn nauwelijks correctieslagen uitgevoerd.

    Beschikbaarheid: Via internet raadpleegbaar corpus, d.m.v. retrievalprogramma. De restricties in gebruik zijn opgelegd door copyright en contractbepalingen met de tekstleveranciers. De toegang tot het corpus is gratis voor academisch onderwijs en voor niet-commerciële onderzoeksdoeleinden.

    Inhoud van het corpus: Krantenteksten (1992-1995), juridische component (1814-1989), gevarieerd samengestelde component (1970-1995).

Aansluiting bij een internationale standaard:

Formaat:

Documentatie:

– **50 miljoen woorden corpus 1996**

Omvang van het corpus: 50 miljoen woorden.

Geannoteerd of niet, aard van de annotatie: Voor een deel (15 miljoen woorden) taalkundig verrijkt. Deze annotatie (woordsoort en lemma) is automatisch aangebracht.

Beschikbaarheid: Alleen op INL raadpleegbaar, via retrievalprogramma (dus niet vrije tekst). De restricties in gebruik zijn opgelegd door copyright en contractbepalingen met de tekstleveranciers. Dit corpus kent grote beperkingen in het externe gebruik ervan.

Inhoud van het corpus: Algemeen Nederlands, 1970-1990, 17 boeken over gevarieerde onderwerpen (30% fictie).

Aansluiting bij een internationale standaard:

Formaat:

Documentatie:

– **Geïntegreerde Taalbank 8ste - 21ste Eeuws Nederlands**

Omvang van het corpus:

Geannoteerd of niet, aard van de annotatie:

Beschikbaarheid: De restricties in gebruik zijn opgelegd door copyright en contractbepalingen met de tekstleveranciers.

Inhoud van het corpus:

Aansluiting bij een internationale standaard:

Formaat:

Documentatie:

– **PAROLE-corpus 3 miljoen woorden**

Omvang van het corpus: 3 miljoen woorden.

Geannoteerd of niet, aard van de annotatie: Het corpus bevat PoS-tagging en lemmatisering voor 250.000 woorden.

Beschikbaarheid: Integrale teksten toegankelijk op cd-rom. Het PAROLE Distributable Corpus wordt eveneens gedistribueerd door ELRA/ELDA, onder productnummer ELRA-W0019.

Inhoud van het corpus: Het corpus bevat teksten uit de categorieën 'boeken', 'kranten', 'tijdschriften' en 'gemengd'. De teksten zijn niet ouder dan 1970.

Aansluiting bij een internationale standaard: De mark-up is conform de PAROLE-DTD, welke weer compatibel is met de DTD van de Text Encoding Initiative (TEI) en de DTD van de Corpus Encoding Standard (CES).

Formaat: SGML.

Documentatie: Beschikbaar op `http://www.inl.nl/PAROLE/distrudoc.html`

– **PAROLE-corpus 20 miljoen woorden**

Omvang van het corpus: 20 miljoen woorden.

Geannoteerd of niet, aard van de annotatie: Verrijkt met TEI-codering, trefwoord en PAROLE-woordsoort (woordsoortcategorieën met subcategorisatie).
Beschikbaarheid: Begin 2002 operationeel. Zal dan raadpleegbaar zijn via internet.
Inhoud van het corpus: Het corpus bevat teksten uit de categorien 'boeken', 'kranten', 'tijdschriften' en 'gemengd'. De teksten zijn niet ouder dan 1970.
Aansluiting bij een internationale standaard: TEI-codering en POS-annotatie gebaseerd op EAGLES.
Formaat: SGML.
Documentatie: Het document getiteld 'Dutch Corpus Documentation' bevat een overzicht van de bronnen en een aantal bijzonderheden over de selectie van teksten, over TEI-tagging en over taalkundige verrijking.

8.3.5 Inventarisatie niet-geannoteerde corpora

- **CONDIV-corpus**

  Omschrijving: Het CONDIV-corpus is een elektronisch toegankelijke, regionaal, stilistisch en diachroon gecontroleerde materiaalverzameling van ongeveer 47.000.000 woorden geschreven Nederlands, die speciaal ten behoeve van het CONDIV-project ontwikkeld werd.

  Meer informatie: `http://allserv.rug.ac.be/~jtaeldem/BelgNedl.html` `http://www.niederlandistik.fu-berlin.de/digitaal/digitaal-11.html`

- **ECI/MCI**

  Omschrijving: De cd-rom *European Corpus Initiative Multilingual Corpus I* bevat drie Nederlandse corpora.

  Meer informatie: `http://www.elsnet.org/eci.html` en `http://www.ldc.upenn.edu/`

- **MLCC**

  Omschrijving: Het eerste deel (ELRA-W0006) van MLCC (Multilingual and Parallel Corpora) bevat artikels uit zes Europese kranten, waaronder 8,5 miljoen woorden uit *Het Financieel Dagblad* (1992-1993). Het tweede deel (ELRA-W0007) bestaat uit een parallel corpus van vertaalde tekst in negen Europese talen, verdeeld in twee subcorpora: geschreven vragen (ongeveer 1,1 miljoen Nederlandse woorden uit 1993) en parlementaire debatten (5 tot 8 miljoen Nederlandse woorden uit de periode 1992-1994). De parlementaire debatten zijn dus vertalingen van gesproken taal. Al het materiaal komt uit *the Official Journal of the European Communities* en is beschikbaar gesteld door de Europese Commissie.

  Meer informatie: `http://www.icp.inpg.fr/ELRA/cata/text_det.html#mlcc`

- **Namur-corpus**

  Omschrijving: Het Namur-corpus is een parallel corpus van Nederlandse, Engelse en Franse tekst, zowel fictie als non-fictie. Het is opgebouwd ten behoeve van een contrastieve studie naar preposities.

  Meer informatie: `http://bank.rug.ac.be/contragram/newslet3.html#FORUM`

8.3.6 Evaluatie niet-geannoteerde corpora

- **CONDIV-corpus**

  Omvang van het corpus: 47.000.000 woorden.

  Geannoteerd of niet, aard van de annotatie: Niet geannoteerd.

  Beschikbaarheid: Geïnteresseerden kunnen voor onderzoeksdoeleinden gebruikmaken van het internetmateriaal in het corpus. Voor inlichtingen en aanvragen, e-mail Stefan Grondelaers (stefan.grondelaers@arts.kuleuven.ac.be) of Vicky Van Den Heede (vicky.vandenheede@rug.ac.be).

  Inhoud van het corpus: Tekst uit Nederlandse en Vlaamse kranten en tekst afkomstig van internet (Usenet en Internet Relay Chat).

  Aansluiting bij een internationale standaard:

  Formaat: De teksten zijn beschikbaar als ASCII-bestanden.

  Documentatie:

- **ECI/MCI**

  Omvang van het corpus: dut01 bevat 600K tokens, dut02 bevat 5203K tokens en dut03 bevat 128K tokens.

  Geannoteerd of niet, aard van de annotatie: Geen annotatie.

  Beschikbaarheid: Distributie door Elsnet (`http://www.elsnet.org/resources/`) en het Linguistic Data Consortium (`http://www.ldc.upenn.edu/`).

  Inhoud van het corpus:

  - dut01: "Newspaper, Dutch, Articles from the student newspaper Universiteitskrant of the University of Groningen from the academic years 1990/1991 and 1991/1992."
  - dut02: "Mixed, Dutch, A large Dutch corpus from INL including transcripts of radio programs, newspaper and magazine issues and some technical texts."
  - dut03: "Mixed, Dutch, A continuation of dut02."

  Aansluiting bij een internationale standaard:

  Formaat:

  Documentatie:

- **MLCC**

  Omvang van het corpus: 8,5 miljoen, plus ongeveer 1,1 miljoen, plus 5 tot 8 miljoen woorden.

  Geannoteerd of niet, aard van de annotatie: Geen linguïstische annotatie.

  Beschikbaarheid: Verkrijgbaar voor niet-commerciële doeleinden voor 450 Euro voor leden van ELRA en 1200 Euro voor niet-leden.

  Inhoud van het corpus: Deel één bevat tekst uit *Het Financieel Dagblad* uit de periode 1992-1993. Deel twee bevat geschreven vragen en parlementaire debatten uit *the Official Journal of the European Communities*.

Aansluiting bij een internationale standaard: SGML-conform TEI P3. Van het deel parlementaire debatten zijn enkel de headers conform TEI.

Formaat: SGML

Documentatie:

- **Namur-corpus**

  Omvang van het corpus: 2,000,000 woorden.

  Geannoteerd of niet, aard van de annotatie: Het Namur-corpus is louter een opgelijnd corpus, zonder extra annotatie. De oplijning is tot op paragraafniveau. Het corpus is dus niet linguïstisch geannoteerd.

  Beschikbaarheid: Het corpus is niet vrij beschikbaar zolang een aantal praktische problemen niet is opgelost. Mogelijk in de toekomst dus wel vrij beschikbaar.

  Inhoud van het corpus: De helft van de tekst uit het corpus is fictie, de andere helft non-fictie.

  Aansluiting bij een internationale standaard:

  Formaat:

  Documentatie:

8.4 Conclusie

Corpora spelen een cruciale rol in het huidige onderzoek op het gebied van TST. Het is daarom verontrustend dat juist op dit punt de situatie voor het Nederlands niet erg goed lijkt, met name voor geschreven taal. Er zijn een aantal kleinere, en deels verouderde, corpora beschikbaar, en enkele grotere corpora zijn alleen onder tamelijk stringente condities beschikbaar. In alle gevallen is de annotatie minimaal. Vergeleken met andere talen is het Nederlands op dit punt duidelijk in een achterstandssituatie terecht gekomen.

Het CGN zal weliswaar een zekere hoeveelheid syntactisch geannoteerd materiaal opleveren, maar gezien de aard van de bron (gesproken taal) kan dit nooit een substituut zijn voor een geannoteerd corpus geschreven Nederlands.

De ontwikkeling van modules voor syntactische en semantische analyse voor het Nederlands wordt zeer gehinderd door het feit dat er geen geschikt corpusmateriaal voorhanden is. Er is daarom alle reden de ontwikkeling van een omvangrijk corpus geschreven Nederlands de komende jaren de hoogste prioriteit te geven.

8.5 Testcorpora

Met de term corpus wordt in deze inventarisatie bedoeld een natuurlijk, of min of meer spontaan verkregen verzameling tekst of spraak in natuurlijke taal voor het afleiden van taalkundige kennis of het testen van taalkundige hypotheses. Een testcorpus onderscheidt zich van een corpus door de handmatige toevoeging van informatie aan de ruwe tekst of spraak. De handmatig toegevoegde informatie, zoals transcripties, woordsoorten, constituenten, 'dialogue acts', onderwerpclassificaties, etc. is van belang bij het ontwikkelen en testen van Nederlandse taaltechnologische applicaties. In deze sectie worden alleen tekstuele corpora beschreven. Spraakcorpora worden beschreven in sectie 10. Multimedia corpora waarin documenten in uiteenlopende formaten zijn opgenomen (tekst, spraak, plaatjes, video, of een combinatie)

spelen tot nu slechts een bescheiden rol in onderzoek met een taaltechnologische dimensie. Een uitzondering is het terrein van multimedia retrieval waarbij audio- en/of beeldmateriaal geïndexeerd wordt aan de hand van het aanwezige taalmateriaal, zoals Teletekst bij video, en onderschriften bij krantenfoto's. Anders dan bij textretrieval is op dit terrein nog niet of nauwelijks sprake van een standaard evaluatiemethodologie.

- **Het Eindhoven corpus**

  Omschrijving: Zie sectie 8.3.3 voor meer informatie over dit corpus. Het Eindhoven corpus werd samengesteld en geannoteerd aan de TU Eindhoven door Uit den Boogaart om een nauwkeurig beeld te krijgen van in Nederland voorkomend taalgebruik door frequentietellingen van woorden. Het corpus bevat tekst afkomstig uit dagbladen, opiniebladen, gezinsbladen, romans en novellen, populair wetenschappelijke teksten, gesproken taal (alleen transcriptie beschikbaar) en 'ambtenarees'. De omvang van het corpus is ruim 53.000 zinnen bestaande uit bijna 800.000 woorden.

  Annotatie: Het corpus is volledig geannoteerd voor woordsoorten en is daarom uitermate geschikt voor het testen (en trainen) van part-of-speech taggers en andere woordsoort disambigueringsprogramma's. In de versie die verder ontwikkeld is door de VU Amsterdam in 1989 zijn de woordsoorten gecodeerd met driecijferige codes en wordt onderscheid gemaakt tussen de standaard woordsoorten, maar ook tussen kenmerken als persoon en enkelvoud/meervoud. De annotatie van het corpus is verder ontwikkeld door [138].

  Beschikbaarheid: Op verschillende instituten in Nederland is een versie van het corpus aanwezig; het is onduidelijk of er copyright op het corpus rust. Waarschijnlijk is dit niet het geval voor wetenschappelijk gebruik.

  State of the art internationaal: Klassieke voorbeelden van vergelijkbare corpora die veel gebruikt zijn voor het trainen van Engelse part-of-speech taggers zijn het Brown corpus [57] en de Penn-treebank [93].

- **Het Wizard of Oz corpus**

  Omschrijving: Het Wizard of Oz corpus werd in 1995 samengesteld op de Universiteit Twente voor de ontwikkeling van dialoogsystemen. Het corpus is tot stand gekomen door middel van een zogenaamd Wizard of Oz (WoZ) experiment, waarbij proefpersonen wordt verteld dat ze communiceren met een computerprogramma. In werkelijkheid echter, communiceren ze met een onzichtbaar persoon die probeert het gewenste gedrag van het computerprogramma te imiteren. Het geïmiteerde gedrag is dat van een automatisch systeem voor informatie over en het reserveren van theatervoorstellingen. Het WoZ corpus is geannoteerd met 'dialogue acts', woordsoorten en constituenten door [127] en is zowel geschikt voor het testen van parsers als voor het testen en ontwikkelen van dialoogstrategieën.

  Beschikbaarheid: Het Wizard of Oz corpus is verkrijgbaar via Hendri Hondorp (hendri@cs.utwente.nl) van de Parlevink groep van de Universiteit Twente.

  State of the art internationaal: Zie de verwijzingen in de sectie over het Eindhoven corpus voor referenties naar corpora die geannoteerd zijn met woordsoort- en constituenten-informatie.

- **Het CLEF corpus**

  Omschrijving: Het CLEF corpus is een testcorpus voor de evaluatie van information retrieval systemen. Het corpus is ontwikkeld aan de Universiteit Twente in samenwerking met NIST (Gaithersburg, VS), IZ Socialwissenschaften (Bonn), IEI-CNR (Pisa), UNED (Madrid) en Universiteit Hildesheim. Het corpus bestaat uit drie onderdelen: tekstuele documenten, zoekvragen ('queries'), en relevantiebepalingen (de 'goede antwoorden'). De documentcollectie bestaat uit bijna 200.000 krantenartikelen die in 1994 en 1995 werden gepubliceerd door het NRC Handelsblad en het Algemeen Dagblad. De zoekvragen zijn opgesteld als natuurlijke taaluitingen waarin de gezochte informatie wordt omschreven; deze dienen als basis voor het formuleren van queries voor zoeksystemen. Op dit moment, in 2001, zijn er 50 zoekvragen beschikbaar, maar uiteindelijk zullen dat er 150 zijn in 2003. De relevantiebepalingen zijn tot stand gekomen door de inzet van potentiële gebruikers van zoeksystemen en bestaan voor elke zoekvraag uit een lijst van de goede (gezochte) documenten. Het Nederlandse CLEF corpus is onderdeel van een groter meertalig testcorpus voor information retrieval, waarvan ook krantenmateriaal in het Duits, Engels, Frans, Italiaans en Spaans deel uitmaakt. De 50 zoekvragen zijn beschikbaar voor alle zes talen (plus nog een aantal talen waarvoor geen documenten zijn) en kan naast experimenten voor elke taal afzonderlijk worden gebruikt voor het testen van systemen voor het zoeken in anderstalige informatie (d.w.z. het zoeken van informatie in alle zes talen met behulp van een zoekvraag in een taal naar keuze: 'cross-language information retrieval').

  Annotatie: Door de 50 zoekvragen en bijbehorende relevantiebepalingen is het corpus geschikt voor het testen van zoeksystemen, filtersystemen, etc. die zich speciaal op de Nederlandse taal richten. Daarnaast bevat het corpus handmatige annotaties van de uitgever van de documentcollectie (PCM Landelijke Dagbladen), zoals handmatig toegevoegde trefwoorden en classificatie naar onderwerp (bijv. sport, binnenland, financieel, etc.). Dit maakt het corpus tevens geschikt voor het testen van automatische classificatiesystemen.

  Beschikbaarheid: Copyrighthouder van de documentcollectie is PCM Landelijke Dagbladen / het Parool. Het corpus wordt voor wetenschappelijke doeleinden beschikbaar gesteld enkel aan deelnemers van de officiële CLEF evaluatie [109]. Meer informatie is te vinden op `http://parlevink.cs.utwente.nl/projects/clef.html` en `http://www.clef-campaign.org`

  State of the art internationaal: Het Nederlandse corpus is samengesteld volgens de methode die al enkele jaren succesvol wordt toegepast in de Text Retrieval Conferenties (TREC, `http://trec.nist.gov`). TREC wordt georganiseerd door NIST: het Amerikaanse National Institute for Standards and Technology dat tevens een van de mede-organisatoren van CLEF is.

- **Het Teletekst-corpus voor het NOS-Journaal**

  Omschrijving: Ten behoeve van de ontwikkeling van spraakherkenning voor het Nederlands die kan worden ingezet voor spoken document retrieval zijn voor de ontwikkeling van de taalmodellen behalve spraakcorpora ook tekstcorpora nodig. In het project DRUID (`http://dis.tpd.tno.nl/druid/`) is daarvoor onder meer een corpus aangelegd van de Nederlandstalige ondertitels voor het NOS-Journaal die via Teletekst worden uitgezonden [45].

Formaat: Platte tekst of xml, beide met tijdcodes.

Beschikbaarheid: Het copyright op Teletekst berust bij de NOS. Het materiaal kan met simpele middelen worden opgevangen van het kabelsignaal. De Parlevink-groep van de Universiteit Twente verzamelt systematisch het Teletekst-materiaal van de NOS-journaals sinds Augustus 1998. Dat corpus kan voor onderzoeksdoeleinden ook door anderen worden gebruikt. Voor inlichtingen: Thijs Westerveld (westerve@cs.utwente.nl) van de Parlevink groep van de Universiteit Twente.

State of the art internationaal: Teletekst-bestanden worden in andere landen ook wel gebruikt voor onderzoeksdoeleinden. De onderzoeksdoelen variëren van de ontwikkeling van taalmodellen voor spraakherkenning tot leeralgoritmen voor de automatische generatie van Teletekst.

- **Het autocue-corpus voor het NOS-Journaal**

  Omschrijving: Naast een Teletekst-corpus is er binnen DRUID ook een corpus aangelegd van de Nederlandstalige autocue files die gebruikt worden door de nieuwslezers van het NOS-Journaal [136]. Het corpus verschilt van het Teletekst-corpus onder meer doordat er ten opzichte van hetgeen er uitgesproken is nauwelijks sprake is van afwijkingen, terwijl de Teletekst is 'ingedikt' omdat er maar ruimte is voor een beperkt aantal karakters per ondertitel.

  Formaat/annotatie: De bestanden zijn opgeslagen in XML-formaat, met annotaties voor tijdcodes en camerabewegingen.

  Beschikbaarheid: Het copyright op het autocue-corpus berust bij de NOS. Voor onderzoeksdoeleinden is het vrij te gebruiken mits de NOS daarvoor toestemming heeft verleend. Voor advies en inlichtingen: Franciska de Jong (fdejong@cs.utwente.nl) van de Parlevink groep van de Universiteit Twente.

  State of the art internationaal: Het is niet bekend of elders ook gebruik gemaakt wordt van autocue-corpora.

- **Het krantenfoto-corpus**

  Omschrijving: Ten behoeve van de ontwikkeling en evaluatie van cross-modal retrieval, waarbij vanuit tekstuele zoekvragen beeldmateriaal gezocht kan worden of andersom, is er in het kader van het project LIP (`http://parlevink.cs.utwente.nl/Projects/lip.html`) een corpus aangelegd van krantenfoto's met bijbehorende onderschriften [144]. Het corpus bestaat uit de krantenfoto's uit de online versie van het Reformatorisch Dagblad (`http://www.reformatorischdagblad.nl`) sinds december 1997. De foto's zijn onderverdeeld in categorieën als voorpagina, binnenland, buitenland en economie.

  Formaat: Html

  Beschikbaarheid: Het copyright op het materiaal berust bij het Reformatorisch Dagblad, bij de verschillende persbureau's die de foto's hebben aangeleverd en bij de fotografen. Het archief van het RD (inclusief de foto's) is te vinden op `http://www.reformatorischdagblad.nl`). Het corpus kan voor onderzoeksdoeleinden ook door anderen worden gebruikt. Voor inlichtingen: Thijs Westerveld (westerve@cs.utwente.nl) van de Parlevink groep van de Universiteit Twente.

State of the art internationaal: Het is onbekend of er vergelijkbare corpora zijn voor andere talen. De (on)mogelijkheden van cross-modal retrieval wordt slechts op bescheiden schaal onderzocht.

9 Spraaktechnologie modules

9.1 Prosodiegeneratie en -herkenning

Deze sectie handelt over taalspecifieke prosodische componenten ten behoeve van spraaksynthese, spraakherkenning en gespreksanalyse. Men onderscheidt twee soorten componenten:

Prosodiegeneratie Bij spraaksynthese wil men natuurlijke en expressief klinkende spraak verkrijgen. Daartoe is het onontbeerlijk dat de toonhoogte een geschikt verloop kent, dat er op de juiste plaatsen pauzes tussen woorden gelaten worden, dat de klanken de gepaste lengtes hebben, en dat het volumeverloop in orde is. De generatie van deze prosodische variabelen gebeurt vaak in twee stappen. In een eerste stap genereert men een fonologische beschrijving van prosodie in termen van accenten, prosodische grenzen, toonhoogtebewegingen, etc. In een tweede stap wendt men alle beschikbare informatie aan voor de generatie van de uiteindelijke toonhoogte-, pauze-, klankduur- en volumepatronen.

Prosodieherkenning Bij spraakherkenning en gespreksanalyse wil men uit het akoestisch signaal prosodische elementen halen die kunnen bijdragen tot een betere herkenning, of tot een betere identificatie van de sprekerintentie (b.v. het onderscheid tussen een bevestigend en een vragend *ja*, het onderscheid tussen het woord *punt* en het leesteken *punt*), etc. De extractie van die prosodische elementen noemt men prosodieherkenning.

9.1.1 State of the art internationaal

Er bestaan verschillende commerciële spraaksynthesizers voor alle belangrijke talen, waaronder het Nederlands, en deze bevatten alle een prosodiegenerator. Nochtans vertonen deze prosodiegeneratoren een aantal belangrijke tekortkomingen. Vooreerst genereren ze doorgaans niet veel meer dan een neutraal en weinig expressief toonhoogteverloop. Verder nemen ze doorgaans enkel het frase- of zinsniveau in rekening, waardoor ze weinig geschikt zijn voor het voorlezen van volledige teksten, laat staan voor het voorlezen van gestructureerde teksten zoals HTML en Word-documenten. Daarnaast bevat de prosodie ook nog geregeld storende fouten, met name bij prosodische grenzen en accenten. M.a.w., als men het aantal toepassingen van spraaksynthese wil vergroten, dan zal men veel betere prosodiegeneratoren dienen te ontwikkelen.

Op het vlak van de prosodieherkenning is de situatie vrij eenvoudig. De meeste spraakherkenners ingebouwd in dicteersoftware maken gebruik van een zeer eenvoudige prosodieherkenner - vaak uitsluitend gebaseerd op de detectie van pauzes - voor het herkennen van leestekens in een dictaat. Verder is er heel wat onderzoek in Nederland, Duitsland, Japan en de Verenigde Staten naar het gebruik van prosodieherkenners in de context van dialoogsystemen, maar de resultaten hiervan zijn tot op heden eerder bescheiden van aard.

Voor meer informatie:

- M. Ostendorf, University of Washington
  `http://ssli.ee.washington.edu/people/mo/cover.html`
- E. Shriberg en A. Stolcke, International Computer Science Institute (ICSI)
  `http://www.speech.sri.com/people/`

- Julia Hirschberg, AT& T Research Labs
  `http://www.research.att.com/~julia/`
- Reinhard Karger, M.A., German Research Center for Artificial Intelligence (DFKI GmbH)
  Email: karger@dfki.uni-sb.de
- M. Swerts, IPO Center for User-System Interaction, Eindhoven en CNTS, Antwerpen
  `http://www.ipo.tue.nl/homepages/mswerts/index.html`
- Keikichi Hirose, University of Tokyo, Japan
  `http://www.gavo.t.u-tokyo.ac.jp/index-e.html`
- Colin Wightman, Minnesota State University, Mankato
  `http://www.ee.nmt.edu/~cww/`

Uit het voorgaande volgt dat er nood is aan verder onderzoek naar de ontwikkeling van betere prosodische componenten. Dit onderzoek blijkt bovendien in hoge mate gebruik te maken van automatisch lerende systemen die leren uit data. In het licht daarvan is het nodig om niet alleen een inventaris te maken van de bestaande prosodiegeneratoren en -herkenners voor het Nederlands, maar ook van de beschikbare infrastructuur voor de verdere ontwikkeling van dergelijke prosodische componenten.

9.1.2 Specifieke evaluatiecriteria

Voor beide componenten gelden de standaard evaluatiecriteria van beschikbaarheid, bruikbaarheid, documentatie en kwaliteit.

Aangezien de prosodiegenerator steeds een onderdeel vormt van een spraaksynthesizer waarvan de kwaliteit moeilijk op een objectieve manier te evalueren is, kan ook de kwaliteit van een prosodiegenerator moeilijk objectief worden gemeten.

Wanneer de prosodieherkenner wordt aangewend als onderdeel van een spraakherkennings- of dialoogsysteem, dan kan zijn kwaliteit worden geschat via een evaluatie van het totaalsysteem met en zonder de prosodieherkenner.

9.1.3 Inventaris beschikbare software

Prosodiegeneratoren voor het Nederlands Er bestaan reeds verschillende spraaksynthesizers voor het Nederlands (zie ook sectie 9.2 voor een gedetailleerdere beschrijving), en alle hebben een ingebouwde prosodiegenerator. Er is dikwijls nochtans weinig informatie beschikbaar omtrent de precieze eigenschappen van die generator.

- **Fluent Tekst naar Spraak**

  Omschrijving: De prosodiegenerator berekent enkel een geschikt toonhoogteverloop.

  Beschikbaarheid: te koop

  Meer informatie: `http://www.fluency.nl`

- **L& H TTS-3000**

  Omschrijving: De prosodiegenerator berekent een geschikt toonhoogteverloop, alsook aangepaste klank- en pauzeduren.

Beschikbaarheid: te koop

Meer informatie: `http://www.lhsl.com`

- **L& H RealSpeak**

  Omschrijving: RealSpeak is een corpusgebaseerde spraaksynthesizer waarin de prosodiegenerator prosodische elementen (in hoofdzaak een fonologische beschrijving van prosodie) aanbrengt die dan worden aangewend ter identificatie van de meest geschikte elementaire bouwstenen in de segmentdatabank. Er wordt gebruik gemaakt van het zogenoemde 'unit-selection' principe (zie sectie 9.2). De manipulatie van de prosodie blijkt slechts in beperkte mate mogelijk.

  Beschikbaarheid: te koop

  Meer informatie: `http://www.lhsl.com`

- **IPO**

  Omschrijving: Aan het IPO loopt onderzoek naar de toepasbaarheid van verschillende 'discourse structure' theorieën voor prosodische structurering van teksten, met de bedoeling om uiteindelijk een algoritme te ontwikkelen dat informatie over 'discourse structure' als input neemt en daarna een geschikte prosodische realisatie van deze structuur genereert. Men verwacht dat de kwaliteit van de synthetische spraak zal verbeteren als de structuur van de tekst op een juiste manier prosodisch geannoteerd wordt (met toonhoogte, spreeksnelheid, plaats van pauzes, duur).

  Beschikbaarheid: ??

  Meer informatie: `http://www.ipo.tue.nl/ipo/sli/research.html`
  `http://www.ipo.tue.nl/homepages/ovherwij/work.html`

- **IPO/KUN**

  Omschrijving: IPO en KUN hebben samen onderzoek verricht naar prosodiegeneratie ter verbetering van niet-commerciële Noord-Nederlandse spraaksynthese met gebruikmaking van taalkundige kennis.

  Beschikbaarheid: ??

  Meer informatie: `http://www.ipo.tue.nl/ipo/sli/research.html`
  `http://lands.let.kun.nl/staff/kerkhoff.en.html`

  Zie ook [96].

- **CNTS/KUB**

  Omschrijving: CNTS en KUB verrichten samen empirisch onderzoek (PROSIT-project) naar de generatie van een natuurlijk klinkende prosodie op basis van: (a) robuuste analyse van tekst met behulp van technieken uit *information retrieval* en *information extraction*, en (b) geavanceerde zelflerende en meta-lerende systemen.

  Beschikbaarheid: ??

  Meer informatie: `http://ilk.kub.nl/prosit/`

- **ELIS/L& H**

Omschrijving: ELIS en L& H verrichten samen onderzoek (PROMOTEX-project) naar de datagedreven ontwikkeling, van prosodische modellen (toonhoogte, klank- en pauzeduren) voor het voorlezen van gestructureerde teksten. Er wordt ondermeer gebruik gemaakt van (recurrente) neurale netwerken waarvan de inputs, inclusief de prosodische features, automatisch worden gegenereerd uit de databanken.

Beschikbaarheid: niet beschikbaar

Meer informatie: `http://chardonnay.elis.rug.ac.be/en/research/`

Prosodieherkenners voor het Nederlands

- **Parlevink groep, Universiteit Twente**

  Er is werk verricht aan de Universiteit Twente door M. Huijbregts die geprobeerd heeft een prosodietool, ontwikkeld door F. Beaugendre toentertijd aan het IPO, aan te passen zodat deze bruikbaar was voor het Nederlands. Het werk van Huijbregts is beschreven in de scriptie: "Het gebruik van prosodie bij segmentatie van Nederlandse spraak" (2001) aan de Universiteit Twente.

  Een ander gedeelte van het onderzoek bestond uit de training en het testen van een algoritme dat op basis van prosodische kenmerken, pauzes en F0, een segmentatie op frase- en zinsniveau genereert. Het algoritme is getraind op een spraakcorpus dat geannoteerd is met frase-, zins- en topicgrenzen.

  Meer informatie: `http://www.cs.utwente.nl/~ordelman/`
  `http://parlevink.cs.utwente.nl/`

- **IPO**

  Van 1993 tot 1995 is aan het IPO onderzoek verricht door naar het automatisch herkennen en classificeren van toonhoogtebewegingen. De bedoeling was te komen tot deze classificatie op basis van akoestische kenmerken, zoals pitch, luidheid en positie van vowel onset. Het resultaten varieerden van 50 procent bij gedetailleerde herkenning tot boven de 90 ingeval van herkenning van stijgende of dalende intonatie. Dit onderzoek is uitgevoerd door Louis ten Bosch in het kader van een NWO project.

  Meer informatie:

  Referenties [17, 18, 19].

Prosodieherkenners voor andere talen Wereldwijd zijn veel onderzoekers bezig met prosodieherkenning. Hieronder volgt een lijstje van links naar referenties van publicaties, onderzoekers en projecten. Deze lijst in niet volledig, het geeft slechts een indicatie.

- Mari Ostendorf: `http://ssli.ee.washington.edu/people/mo/cover.html`
- Julia Hirschberg: `http://www.research.att.com/~julia/`
- Elmar Noeth: `http://faui58f.informatik.uni-erlangen.de/HTML/English/Persons/MA/noe/noe.html`

9.1.4 Infrastructuur voor ontwikkeling van prosodiecomponenten voor het Nederlands

Zoals reeds vermeld is er nood aan meer onderzoek naar de ontwikkeling van prosodische componenten voor het Nederlands, meer in het bijzonder, datagedreven onderzoek. Daartoe dient de volgende infrastructuur ter beschikking te zijn:

- **Standaarden voor prosodiebeschrijving**

  Wil men de prosodie van het Nederlands op een afdoende wijze kunnen beschrijven, dan dient men vooreerst te beschikken over een beschrijvingssysteem. Dit systeem kan een fonologische beschrijving in termen van accenten, woordprominentie, prosodische grenzen, etc. opleveren, maar het kan ook een parametrische beschrijving van bijvoorbeeld het toonhoogteverloop in functie van de tijd beogen.

- **Prosodisch gelabelde spraakcorpora**

  Zie hiervoor het gedeelte over spraakcorpora in 10

- **Software voor prosodische labeling van Nederlandse spraak.**

  Als er onvoldoende gelabelde data bestaat, dan kan men zijn toevlucht nemen tot automatisch gelabelde data, op voorwaarde dat men beschikt over software voor prosodische labeling van Nederlandse data. Van deze software verwacht men dat ze een significante taalafhankelijke component bezit (b.v. de extractie van kenmerken die voor labeling van belang zijn).

- **Software voor het aanleren van prosodiemodellen voor het Nederlands**.

  Dit is software waarmee men op basis van prosodisch gelabelde corpora prosodische modellen voor prosodiegeneratie en -herkenning kan aanleren.

  Verwacht wordt dat deze software in hoge mate taalonafhankelijk is, en dus niet speciaal voor het Nederlands dient te worden ontwikkeld.

- **Software voor het uittesten van prosodiemodellen voor het Nederlands**

  Wil men de verkregen prosodische componenten op een degelijke manier kunnen evalueren, dan dient men ze te kunnen inpluggen in een spraaksynthesizer (voor prosodiegeneratie) of een spraakherkenner (voor prosodieherkenning). Dit veronderstelt de beschikbaarheid van open en modulair opgebouwde spraaksynthesizers en spraakherkenners voor het Nederlands, zie ook 9.11.

9.1.5 Inventaris van standaarden voor prosodiebeschrijving

Men kan hier gaan kijken naar systemen die reeds voor andere talen werden ontwikkeld, en nagaan in welke mate ze kunnen of moeten worden aangepast om bruikbaar te zijn voor het Nederlands, en in welke mate ze volstaan voor de beschrijving van alle prosodische aspecten die men wenst te bestuderen.

Specifieke evaluatiecriteria: werd de beschrijving reeds toegepast op het Nederlands, zijn er aanwijzingen omtrent de mate waarin een aanpassing aan het Nederlands noodzakelijk is?

- **INTSINT**

  Omschrijving: Deze beschrijving gaat ervan uit dat toonhoogtecontouren adequaat beschreven kunnen worden met een set van tonale symbolen (Top, Mid, Bottom, Higher, Same, Lower, Upstep, en Downstep). Elk van deze symbolen karakteriseert een punt op de F0-curve. Het idee achter het INTSINT systeem is dat de F0-waarden van de doelposities van de toonhoogte op twee manieren geprogrammeerd worden: als absolute tonen (Top, Mid, Bottom) verwijzend naar het algemene toonhoogtebereik van de spreker, en als relatieve tonen (de andere labels) verwijzend naar de waarden van de voorafgaande doelpositie.

  Taalafhankelijkheid: werd reeds met succes toegepast op meerdere talen (Bambara, Bulgaars, Catalaans, Tsjechisch, Nederlands, Engels, Estonisch, Frans, Duits, Hongaars, Italiaans, Kikongo, Occitan, Roemeens, Sloveens, Spaans, Zweeds en Swahili).

  Meer informatie: `http://www.lpl.univ-aix.fr/projects/multext/MUL7.html`

- **TILT**

  Omschrijving: TILT is een systeem voor de parametrische beschrijving van toonhoogtecontouren. Een labeling in TILT bestaat uit toekenning van de vier basisintonatie aspecten aan een uiting: toonhoogte accent, grenstonen, verbindingstonen en stilte. Elk van deze aspecten wordt beschreven met een aantal parameters, zoals b.v. F0 in begin, duur, stijging en daling van toonhoogte etc.

  Taalafhankelijkheid: Het principe is taalonafhankelijk.

  Meer informatie: `http://www.cstr.ed.ac.uk/~awb/papers/ESCA-int97/athens97.html`
  `http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html`

- **ToBi**

  Omschrijving: ToBI (for Tones and Break Indices) is een systeem voor het transcriberen van intonatiepatronen en andere aspecten van prosodie voor het Engels. Het is ontwikkeld door een gevarieerde groep van mensen (elektrotechnici, psychologen, linguïsten, fonetici, etc) die onderzoek doen op het gebied van spraak. Hun gemeenschappelijke doel was het maken van een standaard voor het transcriberen van prosodische elementen zodat prosodisch geannoteerde spraakdata kunnen worden hergebruikt door verschillende onderzoekscentra met eigen doelen. ToBi is over het algemeen een breed geaccepteerde beschrijving van intonatiepatronen.

  Taalafhankelijkheid: Het principe van de labeling of transcriptie van intonatiepatronen is taalonafhankelijk. ToBi is specifiek ontwikkeld op het Amerikaansengels, er bestaat ook een Duitse variant, GtoBi, en een Nederlandse, ToDi.

  Meer informatie: `http://ling.ohio-state.edu/phonetics/ToBI/main.html`

- **ToDi**

  Omschrijving: Dit is de Nederlandse variant van ToBI, maar met dat verschil dat er geen 'break indices' werden voorzien. ToDi hanteert namelijk één enkele grens tussen intonatiefrases, waarvoor het ook bedoeld is.

  Taalafhankelijkheid: ToDi is taalafhankelijk te noemen.

  Meer informatie: `http://lands.let.kun.nl/todi/todi/home.htm`

- **IPO**

  Omschrijving: Dit is een reeds vrij oud systeem voor de transcriptie van Nederlandse intonatie, ontwikkeld op basis van de IPO intonatiegrammatica.

  De theorie van de intonatie is gebaseerd op het modelleren van F0-contouren in de vorm van een gestroomlijnde representatie die equivalent is met het originele contour.

  Taalafhankelijkheid: Is taalafhankelijk.

  Meer informatie: Zie [124].

- **KUN**

  Omschrijving: De KUN beschrijving is meer taalkundig van aard ten opzichte van ToDi, en probeert waar mogelijk te abstraheren en generaliseren om tot een beschrijving te komen die zowel verklarend adequaat als economisch is. Een van de hulpmiddelen hierbij is dat het generatief is, en een onderliggende structuur transformeert naar een oppervlakte beschrijving van de intonatie. Deze autosegmentele beschrijving van de Nederlandse intonatie is gebaseerd op de toongebaseerde analyse van Gussenhoven.

  Taalafhankelijkheid: Is taalafhankelijk.

  Meer informatie: Prof. C. Gussenhoven, KUN

9.1.6 Inventaris van software voor prosodische labeling

Bij de inventarisatie van dergelijke software is er gekeken naar systemen waarvan men kan vermoeden dat ze een goede basis kunnen vormen voor de ontwikkeling van labelingsoftware voor het Nederlands.

- **MULTEXT**

  Omschrijving: Binnen het Europees project MULTEXT heeft CNRS (AIX) een geïntegreerde set van tools voor de prosodische annotatie van spraak ontwikkeld. Met deze tools zijn onder andere de volgende taken uit te voeren:

  - het vertonen en beluisteren van het spraaksignaal en het oplijnen van orthografische labels;
  - het berekenen, vertonen en bewerken van een F0-curve;
  - het berekenen, vertonen en bewerken van doelposities teneinde de F0-curve volgens INTSINT te modelleren, en ze vervolgens door middel van PSOLA-resynthese hoorbaar te maken;

  Beschikbaarheid: Alle Multext resultaten zijn vrij beschikbaar en publiekelijk toegankelijk voor niet-commerciële en niet-militaire doeleinden.

  Meer informatie: `http://www.lpl.univ-aix.fr/projects/multext/MUL7.html`

- **Festival**

  Omschrijving: Festival is ontwikkeld door CSTR (Edinburgh). Het is bevat een volledige multilinguale spraaksynthesizer met verscheidene API's, maar ook een ontwikkel- en onderzoeksomgeving ten behoeve van spraaksynthesetechnieken. Festival

laat ondermeer toe om het TILT-systeem voor de beschrijving van intonatiecontouren te gebruiken. Het is geschreven in C++. Het is opgenomen in de CSLU Toolkit.

Beschikbaarheid: beschikbaar t.b.v. onderzoeksactiviteiten

Meer informatie: `http://www.cstr.ed.ac.uk/projects/festival/`

- **Overige prosodieschema's:** `http://www.dfki.de/mate/d11/annex.html`

  PROSPA , TEI , SAMPROSA , TSM, KIM, PROZODIAG, Göteborg

9.2 Spraaksynthese

Spraaksynthese is het omzetten van geschreven tekst in gesproken taal. Volledige spraaksynthese betekent het direct omzetten van geschreven naar gesproken tekst, zonder dat de afzonderlijke woorden waaruit de tekst bestaat ingesproken dienen te worden. Dit laatste is beter bekend onder de term 'tekst-naar-spraak'.

Mogelijke toepassingsgebieden voor tekst-naar-spraaksystemen zijn telecommunicatie (information retrieval, unified messaging, e-mail reading), taalleren, hulpmiddelen voor gehandicapten, multimedia mens-machine-interactie , etc.

Spraaksynthese is in te delen in:

1. regelgebaseerde synthese en
2. concatenatieve synthese.

Regelgebaseerde synthese Om praktische en historische redenen zijn regelgebaseerde synthesesystemen altijd formant synthesizers. Formant synthesizers beschrijven spraak door middel van een aantal (soms wel 60) parameters, waarvan de meeste gerelateerd zijn aan formant en anti-formant frequenties en bandbreedtes, samen met glottale golfvormen. De Klatt synthesizer [82] is het vroegste voorbeeld van een regelgebaseerde synthesizer.

De spraak die uit de formant synthesizers komt is meestal verstaanbaar, maar laat aan natuurlijkheid nogal wat te wensen over. Regelgebaseerde synthesizers zijn echter wel in trek, omdat het bijvoorbeeld relatief makkelijk is om van de ene naar de andere stem over te schakelen, bovendien maakt regelgebaseerde synthese het mogelijk om te gaan met de articulatorische aspecten van veranderen van spraakstijl.

Concatenatieve synthese Concatenatieve spraaksynthese is gebaseerd op het aan elkaar plakken van elementaire segmenten die uit opgenomen spraak geknipt zijn. Deze segmenten kunnen allerlei eenheden zijn, bijv. allofonen, difonen, etc.. Ook in deze methode kan de duur en de toonhoogte van de geconcateneerde segmenten worden aangepast. Concatenatieve synthese en vooral difoonsynthese heeft als voordeel boven formantsynthese dat het eenvoudiger te realiseren is, terwijl de spraakkwaliteit toch beter is. Echter, doordat opgenomen spraak wordt gebruikt, zijn dergelijke systemen niet erg flexibel: weliswaar kunnen toonhoogte en duur worden aangepast, maar aspecten als stemkwaliteit en timbre kunnen niet worden gemanipuleerd. Tot nog toe is spraaksynthese met gebruik van difonen het meest populair geweest.

De laatste jaren wordt er steeds meer gebruik gemaakt van een techniek voor concatenatieve synthese waarbij geen vaste eenheden (zoals difonen) worden gebruikt, maar waarbij de lengte van de eenheden variabel is. Om deze eenheden te selecteren is zgn. '*unit-selectie*'

nodig. Unit-selectie is het on line selecteren van akoestische eenheden (van verschillende lengte) uit een grote spraakdatabase met fonetische sequenties in verschillende prosodische contexten. Door gebruik te maken van grotere eenheden wordt het aantal grenzen tussen eenheden gereduceerd en ook de hoeveelheid benodigde modificaties aan het geluid. De spraakdatabase is een verzameling opgenomen spraak die gesegmenteerd is, zodanig dat van elk segment (foon: allofoon of difoon) de begin en eindpositie bekend is, en die geannoteerd is met relevante prosodische kenmerken. Op deze manier kunnen eenheden van verschillende grootte worden geselecteerd, bijvoorbeeld fonen, syllabes, woorden of zelfs hele frasen. Het belangrijkste verschil met de traditionele difoonconcatenatie is dat het unit-selectie algoritme probeert de grootst mogelijke eenheid te selecteren. In vergelijking met frase concatenatie is unit-selectie meer flexibel, in de zin dat woorden die niet als zodanig in de database staan, kunnen worden opgebouwd uit kleinere eenheden.
Klabbers, 2001 - persoonlijke communicatie

Voor meer informatie, zie [51, 64, 68, 125].

9.2.1 State of the art internationaal

Er is nog een lange weg te gaan voordat we de perfecte tekst-naar-spraak hebben bereikt. Maar de laatste jaren is zowel de verstaanbaarheid als de natuurlijkheid van de beschikbare systemen sterk verbeterd. Dit is een trend die zich waarschijnlijk zal voortzetten. Echter, op dit moment is voor commerciële doeleinden de natuurlijkheid nog niet goed genoeg, met name de prosodie laat nog te wensen over.

Dingen waaraan in de toekomst aandacht besteed zal moeten worden zijn bijvoorbeeld:

- Hoe verkrijg je een optimale set spraaksegmenten voor concatenatieve spraaksynthese?
- Hoe kun je natuurlijke intonatie en duren genereren op basis van abstracte prosodische patronen?
- Hoe kan worden omgegaan met spreker en spreekstijl effecten?
- Hoe kan een adequate prosodische beschrijving op symbolisch niveau worden gegenereerd op basis van tekstanalyse?

9.2.2 Evaluatiecriteria

Bij subjectieve evaluatie van tekst-naar-spraaksystemen worden meestal de volgende criteria gebruikt:

- Algemene kwaliteit
- Verstaanbaarheid
- Luistergemak
- Aangenaamheid

Objectieve evaluatie van tekst-naar-spraaksystemen kan worden gedaan op veel verschillende niveaus: segmenteel, prosodisch en stemkwaliteit. Segmentele tests worden het meest uitgevoerd, omdat de segmenten (consonanten en vocalen) het belangrijkst zijn voor het herkennen

van een woord: wanneer de segmenten geïdentificeerd kunnen worden, kan het woord herkend worden, ook wanneer duur en prosodie niet helemaal kloppen. Segmentele tests worden gedaan op woordniveau (Diagnostic Rhyme Test / Modified Rhyme test, de SAM standaard segmentele test en de difoontest zijn de meest gebruikte tests) of op zinsniveau (bijvoorbeeld door gebruik te maken van SUS (Semantically Unpredictable Sentences)).

In [64] worden enkele algemene aanbevelingen gedaan waaraan moet worden voldaan wil men bijvoorbeeld twee systemen met elkaar vergelijken.

In het blad Infovisie Magazine van september 2000 wordt gerapporteerd over een evaluatie van vijf verschillende Nederlandse tekst-naar-spraaksystemen (Fluent Dutch text-to-speech, Infovox 330, Orpheus, Speech Connection en Eurovocs Soft). Aan het onderzoek werkten 119 vrijwilligers mee van Nederlandse en Belgische nationaliteit. De systemen werden geëvalueerd aan de hand van een aantal subjectieve criteria (aangenaamheid, luistergemak, etc.) en er werd een objectieve verstaanbaarheidstest gedaan.

Zie: `http://www.esat.kuleuven.ac.be/teo/docarch/iv/sep00/im143.htm#4`

Meer informatie over evaluatie van tekst-naar-spraaksynthese software in [130] en [64].

9.2.3 Inventarisatie beschikbare software

Inventarisatie beschikbare software voor regelgebaseerde synthese

- **TruVoice Lernout en Hauspie**

  Omschrijving: De TruVoice tekst-naar-spraakserie van Lernout en Hauspie maakt gebruik van formant synthese, dit is een vorm van regelgebaseerde synthese, waarbij de het spraakkanaal wordt beschreven met behulp van een mathematisch model. Hierbij wordt de tekst eerst voorbewerkt op een manier die zeer goed werkt voor applicaties die e-mails, namen, of adressen voorlezen.

  Beschikbaarheid: Te koop

  Meer informatie: `http://www.lhs.com/ssyn/ttstvhost.asp`

Inventarisatie beschikbare software voor concatenatieve synthese

Difoonsynthese

- **Fluent Tekst naar Spraak**

  Omschrijving: Dit systeem synthetiseert natuurlijk klinkende spraak. Om dit te bereiken heeft de synthesizer een specificatie van spraakgeluiden, duren en toonhoogte nodig. Er zijn regels geschreven die deze informatie van een fonetische transcriptie afleiden. Voor de volledige tekst-naar-spraakconversie is er additionele software ontwikkeld die geschreven tekst naar een fonemische transcriptie omzet. De software gebruikt een grote woordenlijst, uitspraakregels en regels die numerieke informatie en punctuatie interpreteren.

  Beschikbaarheid: te koop

  Meer informatie: `http://www.fluency.nl/`

- **KUN-TTS**

  Omschrijving: KUN-TTS is voorgekomen uit het ASSP-project. Hoewel het de laatste jaren met beperkte middelen is onderhouden (uitbreiding van lexicon, toevoegen van

nieuw algoritme voor segmentduren) is het in meerdere opzichten niet meer up-to-date. Tevens zijn de verschillende onderdelen ten opzichte van elkaar modulair opgebouwd maar vaak verweven in elkaar wat onderhoud moeilijker maakt.

Beschikbaarheid: Alleen beschikbaar voor Polyglot-partners.

Meer informatie: J. Kerkhoff, KU Nijmegen

- **L& H TTS3000**

  Omschrijving: De Lernout & Hauspie's TTS/3000 tekst-naar-spraakproductlijn converteert willekeurige tekst naar redelijk natuurlijk klinkende synthetische spraak. De onderliggende synthesetechnologie is geparametriseerde segmentconcatenatie, waarbij de segmenten di- tri- en tetrafonen kunnen zijn. Gedetailleerde linguïstische analyse van de invoertekst levert een in hoge mate correcte uitspraak, gecombineerd met een geavanceerde, eveneens op regels gebaseerde prosodie.

  TTS/3000 is beschikbaar in verschillende versies voor Windows platform. Daarnaast is TTS/3000 beschikbaar voor het in de telefoniewereld populaire Antares platform.

  Beschikbaarheid: te koop

  Meer informatie: `http://www.lhs.com/ssyn/tts3000m.asp`

- **Spengi**

  Omschrijving: Spengi (SPeech ENGine) is de benaming voor het spraaksynthesesysteem dat op het IPO is ontwikkeld en nog steeds in ontwikkeling is. Spengi is een phonetics-to-speech-systeem, en verwacht dus een fonetische transcriptie als invoer. De synthese is gebaseerd op difonen en de kwaliteit van de spraak is *state of the art*, mede door goede prosodische beregeling en geavanceerd gebruik van PSOLA technieken. Spengi is beschikbaar als een API en kan daardoor makkelijk in bijvoorbeeld een C-programma worden geïntegreerd. Verder zijn er twee front-end applicaties beschikbaar voor demonstratie- en onderzoeksdoeleinden: Ipologue is een conventioneel command-line programma (DOS en UNIX); Calipso is een Windows programma. Beide applicaties kunnen gebruik maken van grafeem-foneemomzetters die ontwikkeld zijn aan de KUN en aan de KUB. Calipso is verder nauw geïntegreerd met het signaalbewerkingsprogramma GIPOS.

  Beschikbaarheid: Er zijn momenteel twee difoondatabases beschikbaar voor het Nederlands: een vrouwenstem en een mannenstem.

  Meer informatie: `http://sirius.ure.cas.cz/dpt210/cost258/terken.html`

- **Infovox 330**

  Omschrijving: Infovox 330 is een meertalig spraaksyntheseprogramma dat beschikbaar is voor Windows 95 en Windows NT3.51 of hoger. Het programma kent momenteel vijf talen: Nederlands, Frans, Engels, Duits en Zweeds. Per taal maakt de Infovox 330 gebruik van een bibliotheek met vooraf opgenomen menselijke spraakvoorbeelden om een zo natuurgetrouw mogelijke spraaksynthese te produceren. De gebruiker kan de inhoud van deze woordenboeken naar eigen smaak en behoefte wijzigen. Infovox 330 werkt volgens de Microsoft SAPI-standaard. De gebruiker kan spraakparameters zoals spreeksnelheid (tot maximaal 400 woorden per minuut), intonatie en geluidsvolume

instellen. Op de website van Telia Promotor is een demoversie van dit product beschikbaar. De minimale systeemvereisten zijn een Pentium PC met 32 Mb werkgeheugen en een Windows-compatibele 16-bits geluidskaart.

Infovox 330 kan niet zelfstandig worden gebruikt, het wordt als tekst-naar-spraakmodule gekoppeld aan andere software.

Beschikbaarheid: te koop

Meer informatie:

`http://www.promotor.telia.se/infovox/product.htm`
`http://www.kompagne.nl/htm/infovox.htm`

- **Eurovocs Soft**

  Omschrijving: Eurovocs Soft is een spraaksyntheseprogramma voor Windows 95/98, Windows NT en Windows 2000, gebaseerd op spraaktechnologie van Lernout & Hauspie. Het programma wordt aangeboden in een ééntalige, Nederlands, en in een zestalige versie. Het geluidsvolume, de toonhoogte, de spreeksnelheid en de stemkeuze behoren tot de basisinstellingen. De gebruiker kan elk van deze instellingen in negen stappen bijregelen. Dit programma is compatibel met de SAPI-standaard van Microsoft. Minimale configuratie: PC met Pentium processor, 32 Mb RAM-geheugen, een 16-bits SoundBlaster-compatibele geluidskaart en ongeveer 2,5 Mb vrije ruimte op de harde schijf.

  Eurovocs kan niet zelfstandig worden gebruikt, het wordt als tekst-naar-spraakmodule gekoppeld aan andere software.

  Beschikbaarheid: te koop

  Meer informatie: `http://www.kompagne.nl/htm/eurovocs.htm`

- **Orpheus**

  Omschrijving: Orpheus is een meertalig spraaksyntheseprogramma dat beschikbaar is voor Windows 95/98 en Windows NT4.0/2000. Het kent ongeveer 40 talen waaronder Nederlands, Frans, Engels, Amerikaans, Duits, Spaans, Castiliaans, Italiaans. Aan een Orpheus-versie met Microsoft SAPI-compatibiliteit wordt gewerkt. De gebruiker kan spraakparameters zoals spreeksnelheid, intonatie en stem instellen. De aanbevolen systeemvereisten zijn een Pentium II PC met 20 Mb vrije harde schijfruimte, 32 Mb (voor Windows 95/98) of 64 Mb (voor Windows NT/2000) werkgeheugen en een SoundBlaster compatibele 16-bits stereo-geluidskaart.

  Beschikbaarheid: te koop

  Een gratis demoversie, die gedurende 35 minuten te gebruiken is, kan via de website van Dolphin verkregen worden.

  Meer informatie: `http://www.dolphinuk.co.uk/news/jan2001_bul.htm`

- **MBROLA**

  Omschrijving: MBROLA is een difoongebaseerde spraaksynthese module. Het neemt een lijst van fonemen als input, samen met prosodische informatie (duur van de fonemen en een beschrijving van de toonhoogte), en produceert spraak van 16 bits (linear), met een sample frequentie die gelijk is aan die van de difoondatabase die gebruikt wordt.

Het is hiermee geen tekst-naar-spraakmodule, maar een difoonspraakmodule. Er zijn drie difoondatabases voor het Nederlands beschikbaar, nl1, nl2 en nl3, waarvan de eerste twee een mannenstem bevat en de laatste een vrouwenstem.

Beschikbaarheid: MBROLA is vrij verkrijgbaar voor niet-commerciële en niet-militaire doeleinden.

Meer informatie: `http://tcts.fpms.ac.be/synthesis`
`http://tcts.fpms.ac.be/synthesis/mbrola.html`

Toepassingen die zijn gemaakt met MBROLA:

1. Babel - technologies: SDK - Nederlandse vrouwenstem
   Te koop: `http://www.babeltech.com/`
2. Archangelis: Plug-in die stemmen toevoegt aan een webpagina
   gratis te downloaden: `http://www.archangelis.com/`
3. WinEuler (Multitel TCTSlab): TTS systeem voor gebruikers, ontwikkelaars, etc.
   Gratis te downloaden
4. MBRDICO (Multitel TCTSlab):
5. Fluency (`http://www.fluency.nl/`)

Synthese door middel van 'unit selection'

- **L& H RealSpeak**

  Omschrijving: RealSpeak is een tekst-naar-spraakpakket dat computer tekst leest en omzet naar natuurlijk klinkende en verstaanbare spraak. De technologie is gebaseerd op concatenatie van fragmenten van menselijke spraak. RealSpeak bevat niet alleen difonen, maar ook syllaben en lagere foneemsequenties. RealSpeak maakt gebruik van diepe linguïstische processing voor de juiste uitspraak en prosodische regels voor de juiste intonatie.

  Beschikbaarheid: te koop

  Meer informatie: `http://www.lhsl.com/ssyn/default.asp`

- **BOSS**

  Omschrijving: BOSS (Bonn Open Synthesis System) is een open source software architectuur voor zowel difoonsynthese als synthese op basis van unit selectie. BOSS heeft een modulaire structuur waardoor het toevoegen of weghalen van modules eenvoudig is. Door de strikte scheiding van algoritmen en data is het Boss systeem taalonafhankelijk een taalonafhankelijk systeem, de unit selectie database die wordt gebruikt is taalafhankelijk. Experimenten hebben aangetoond (zie [80]) dat een werkend systeem voor het Nederlands gemaakt kan worden zonder aanpassingen aan de source code.

  Beschikbaarheid: Open source, software verkrijgbaar via Universiteit van Bonn.

  Meer informatie:

  `http://www.ikp.uni-bonn.de/~kst/boss_ii.htm`

  Zie ook [80]

9.3 Spraakherkenning

Automatische spraakherkenning (ASH) is het automatisch omzetten van menselijke spraak (akoestisch) in woorden (tekst). Deze woorden kunnen dienen als het uiteindelijke resultaat, maar kunnen ook als invoer dienen van een module die bijvoorbeeld probeert te begrijpen wat de woorden betekenen (tekstbegrip). In dit hoofdstuk over spraakherkenning wordt tekstbegrip niet behandeld, dit komt aan de orde in het hoofdstuk Semantische en Pragmatische analyse. We zullen ons hier beperken tot de zgn. spraak-naar-tekstomzetting.

Automatische spraakherkenning kent vele toepassingen, variërend van het automatiseren van allerlei telefonische diensten die informatie verstrekken over het weer, het telefoonverkeer, de beurs en het openbaar vervoer tot het bedienen van machines door middel van de stem of telefonisch winkelen en telefonisch bankieren. Ook het dicteren van bijvoorbeeld medische dossiers is een belangrijk toepassingsgebied, alsook het ontsluiten van spraakarchieven (spoken document retrieval).

Er zijn veel verschillende spraakherkenningsystemen. Deze kunnen onder andere verschillen in de volgende dimensies:

- sprekerafhankelijke vs. sprekeronafhankelijke
- losse woorden / commando's vs. continue spraak
- groot lexicon vs. klein lexicon
- etc.

De meeste automatische spraakherkenningsystemen bevatten de drie volgende, taalafhankelijke componenten:

- akoestische decodering
- lexicon
- taalmodellen

Figuur 1 laat zien hoe het herkenproces verloopt voor een standaardherkenner. Allereerst wordt het binnenkomende signaal voorbewerkt: op vaste tijdsintervallen (van 10 à 25 ms.) wordt het signaal geanalyseerd en omgezet in reeksen van getallen die de spectrale eigenschappen van het signaal representeren. Op basis van deze getallen en drie ASH modules (akoestische modellen, het lexicon en de taalmodellen) wordt geprobeerd om de onbekende uiting te herkennen, dwz. om de spraak om te zetten naar tekst. De drie ASH modules zullen hieronder in meer detail worden beschreven. Voor een overzicht, zie [121].

Akoestische decodering Voor akoestische decodering kan gebruik gemaakt worden van Hidden Markov Modellen (HMM-en) en neurale netwerken. Beide worden hier kort besproken.

Hidden Markov Modellen

De meest gebruikte manier om spraak te modelleren is door gebruik te maken van Hidden Markov Modellen. Het gebruik van HMM-en is gebaseerd op het idee dat spraak een opeenvolging is van stationaire akoestische events met verschillende duren. Een HMM bestaat uit een opeenvolging van states die worden verbonden door transities. De states staan voor de alternatieven van het stochastische proces, en de transities bevatten probabilistische en andere

Figuur 1: Spraakherkenning

data op basis waarvan wordt besloten welke state de volgende state wordt. Elke state bevat statistische gegevens over alle samples van een deel van het woord of foneem waarvoor het model moet staan. Deze statistische gegevens beschrijven de gemiddelden en standaard deviaties voor alle parameters. Een vergelijking van het binnenkomende signaal met alle states in een HMM levert een score op, die weergeeft hoe waarschijnlijk het is dat het betreffende model matcht met de input.

Meer informatie over HMM-en is te vinden in [111].

Neurale netwerken

De laatste jaren wordt naast spraakmodellering door middel van HMM-en, modellering van spraak met behulp van neurale netten toegepast. De ontwerpers van deze methode zijn geïnspireerd door de wijze van informatieverwerking in de hersenen van de mens.

Een neuraal net bestaat meestal uit een aantal lagen, bestaande uit cellen. Op de inputlaag wordt het spraaksignaal gerepresenteerd, bijvoorbeeld via de binaire representatie van een gecodeerd spectrum. De inputlaag wordt gevolgd door een aantal verborgen lagen, en tenslotte volgt de outputlaag, waarop het uitgangssignaal wordt gerepresenteerd. De verbindingen tussen de cellen, en het aantal lagen is verschillend voor ieder neuraal net. Voor spraak wordt vaak één verborgen laag gebruikt, waarbij de cellen allemaal met alle cellen uit de volgende laag verbonden zijn, maar niet met cellen uit de eigen laag. De outputwaarde van elke cel is een gewogen functie van de verschillende inputwaarden, gecombineerd met een bepaalde drempel. De vorm van de weegfunctie (bijvoorbeeld een stap- of S-vorm) kan verschillen bij diverse netarchitecturen.

Tijdens de trainingsfase moeten de weegfactoren zodanig worden geoptimaliseerd dat de gewenste relatie tussen input en output ontstaat. Daartoe wordt elk woord een groot aantal keer aangeboden aan het net en worden via een iteratieve procedure de weegfactoren bijgesteld (bijv. backward error propagation algoritme).

Er bestaan verschillende types neurale netwerken waarvan de meest frequent gebruikte de 'multi-layer perceptron' of het 'feed forward' neuraal netwerk is.

In normale realistische situaties, waar altijd sprake is van achtergrondlawaai of van meerdere stemmen, blijken neurale netwerken een veel betere performance te geven dan de traditionele technieken. Zij leveren echter weinig kennis op, want het is niet mogelijk te achterhalen hoe het neuraal net precies zijn werk doet.

Vaak worden hybride systemen gebruikt waarin gebruik wordt gemaakt van een combinatie van HMM-en en neurale netten. Zie onder anderen de volgende referenties: [13, 24, 25, 100, 115].

Lexicon Uiteraard herkent de computer alleen de woorden die in zijn woordenboek staan. Het woordenboek bevat van ieder woord twee vormen: ten eerste de orthografische vorm, oftewel het woord zoals het wordt geschreven, en ten tweede het woord zoals het wordt uitgesproken. Dit is een reeks van foneemsymbolen, en heet daarom een foneemtranscriptie. Ieder foneemsymbool staat voor een bepaalde klank.

Er zijn twee verschillende manieren om een lexicon te bouwen.

1. Neem bestaande data (zie de opsommingen in 9.3.3)
2. Neem een grafeem-foneemomzetter, welke op basis van de orthografische vorm automatisch de juiste foneemtranscriptie oplevert (zie ook de opsomming in sectie 9.3.3).

Of gebruik een combinatie van beide methoden, door bijvoorbeeld zoveel mogelijk woorden op te zoeken in bestaande lexica, en voor de onbekende woorden (bijv. eigennamen) gebruik te maken van een grafeem-foneemomzetter.

De zo juist beschreven manier kan gebruikt worden om een basislexicon te maken. Deze basislexica worden vaak geoptimaliseerd voor een bepaald domein of toepassing met lexiconadaptatie (zie [120]. Lexiconadaptatie komt aan de orde in sectie 9.9

Taalmodellen Naast de neurale netten en/of HMM-en voor de akoestische decodering en het lexicon worden tijdens de herkenning taalmodellen gebruikt. Een taalmodel representeert de kennis van de taal die moet worden herkend. Zij worden tijdens de herkenning gebruikt om te disambigueren en om een keuze te maken uit de vele foon- en woordhypothesen die door de herkenner opgeleverd worden. Het taalmodel legt restricties op aan de manier waarop woorden kunnen worden gecombineerd tot zinnen. Er zijn verschillende soorten taalmodellen:

- finite-state taalmodel: een finite-state network wordt gebruikt om weer te geven welke woordvolgordes toegestaan zijn.
- n-gram taalmodel: hierbij zijn alle woordvolgordes in principe mogelijk, maar de waarschijnlijkheid van een woord hangt af van zijn $n-1$ voorgangers.
- grammatica gebaseerd taalmodel: deze modellen zijn gebaseerd op stochastische contextvrije grammatica's of andere 'phrase structure grammars'.

De n-gram modellen worden verreweg het meeste gebruikt.

9.3.1 State of the art internationaal

Vooropgesteld moet worden dat de *state of the art* heel erg afhankelijk is van de specifieke toepassing waarvoor de spraakherkenner gebruikt zal gaan worden. Bijv. voor toepassingen met een klein lexicon is hele woord modellering geschikt, voor systemen met een lexicon dat duizenden woorden bevat zijn akoestische woordmodellen totaal niet geschikt.

Er is de laatste jaren veel vooruitgang geboekt in performance wanneer deze gemeten wordt in Word Error Rate (zie sectie 9.3.2). De WER daalt elke twee jaar met een factor twee. Verder worden systemen steeds meer sprekeronafhankelijk en werken ze met steeds grotere vocabulaires. Deze vooruitgang is niet in de laatste plaats mogelijk gemaakt door de opkomst van HMM-en, en het uitbreiden van de capaciteit van onze computers.

Er is wereldwijd veel werk verricht aan het bouwen van spraakcorpora voor het ontwikkelen, trainen en evalueren van spraakherkenningsystemen. Tenslotte zijn er steeds meer standaarden ontwikkeld voor het testen van spraakherkenners, waardoor het mogelijk wordt verschillende spraakherkenners met elkaar te vergelijken.

Er worden meer en meer technieken toegepast om spraakherkenners meer bestand te maken tegen onverwachte invoer, zoals ruis en spraak van niet-moedertaal sprekers. Deze technieken zullen worden besproken in secties 9.5 en 9.6

9.3.2 Evaluatiecriteria

Voor evaluatie van spraakherkenners worden de volgende twee maten gebruikt (beide op woordniveau):

Word Error Rate: WER = 100% * (S+I+D) / N

Word Accuracy: WAcc = 100% * (N-(S+D)) / N

Waarbij:N = # woorden, S = # substituties, I = # inserties, en D = # deleties. 'Word Error Rate' is de meest gebruikte maat. De maat 'Word Accuracy' heeft als nadeel dat de inserties er niet in verdisconteerd zijn.

Ook voor de hierboven besproken drie onderdelen van de spraakherkenner bestaan aparte evaluatiecriteria:

- Modellen:
  - Foon Error Rate: het relatief aantal fonen dat geïnserteerd, gedeleerd of gesubstitueerd wordt in een (vrije) foonherkenning (analoog aan WER).
  - Foon Accuracy: het relatief aantal fonen dat correct herkend wordt in een vrije foonherkenning (analoog aan WAcc).
- Lexica:
  - Dekking: Hoeveelheid woorden uit een testset die in het lexicon voorkomt. Direct hieraan gerelateerd is het aantal Out-of-Vocabulary (OoV) woorden: woorden die wel in de testset voorkomen, maar die niet gevonden kunnen worden in het lexicon.
  - Verwarring: Fonetische gelijkenis tussen de woorden in het lexicon.
- Taalmodellen:
  - Perplexiteit.

Belangrijk is dat voor een eerlijke evaluatie van een spraakherkenningsysteem een benchmark beschikbaar moet zijn. Alleen wanneer de verschillende spraakherkenners zijn getest op een zelfde, onafhankelijke testset, kunnen de resultaten onderling op een eerlijke manier worden vergeleken. Voor het Amerikaansengels zijn dergelijke benchmarks sinds lange tijd beschikbaar (bijv. Darpa Resource Management, Wallstreet Journal, Switchboard, en Call Home). Voor het Nederlands is in het SQALE project door TNO TM een benchmark corpus gecreëerd dat het mogelijk maakt verschillende systemen met elkaar te vergelijken, zie ook hoofdstuk 10. Het corpus bestaat uit zinnen voorgelezen uit kranten in het Nederlands, Engels, Frans en Duits.

9.3.3 Inventarisatie beschikbare software

Akoestische decodering

Onderstaande is een opsomming van de meest bekende software modules voor het trainen van akoestische modellen. Deze modules zijn op zich niet specifiek voor het Nederlands, maar een aantal parameters moeten wel specifiek voor het Nederlands ingesteld worden, in elk geval is bijvoorbeeld de gebruikte lijst met fonemen (in geval van foneemmodellen) of de lijst met woorden (in geval van woordmodellen) taalspecifiek.

Hidden Markov Modellen

- **Philips SpeechPearl**

  Omschrijving: SpeechPearl is een sprekeronafhankelijk spraakherkenningspakket voor continue spraak. Met dit pakket kunnen ook taalmodellen getraind worden.

  Beschikbaarheid: te koop

  Meer informatie: `http://www.speech.philips.com/`

- **Philips SpeechMania**

  Omschrijving: SpeechMania is een softwareplatform voor natuurlijke spraakherkenning en natuurlijke taalverwerking. SpeechMania kan worden gebruikt voor het maken van bijvoorbeeld applicaties voor telefoondialogen.

  Beschikbaarheid: te koop

  Meer informatie: `http://www.speech.philips.com/`

- **HTK**

  Omschrijving: De Hidden Markov Toolkit (HTK) is een toolkit voor het bouwen en bewerken van HMM-en. Hoewel HMM-en ook voor andere doeleinden kunnen worden gebruikt, wordt HTK voornamelijk toegepast voor spraakherkenning. HTK is een verzameling library modules en programma's in C, die kunnen worden gebruikt voor spraakanalyse, trainen van HMM-en, spraakherkenning en analyse van de resultaten. Zowel continuous density HMM-en als discrete verdelingen worden ondersteund.

  Beschikbaarheid: gratis te downloaden

  Meer informatie: `http://htk.eng.cam.ac.uk/`

- **ESAT/PSI herkenner**

  Omschrijving: Is een large-vocabulary sprekeronafhankelijke spraakherkenner, met een modulaire opbouw waarbij alle modules gescheiden zijn, en met elkaar communiceren via een verzameling eenvoudige interface protocollen.

  Meer informatie: `http://www.esat.kuleuven.ac.be/~spch/research/Recog/index.html`

- **SpeechWorks 6.5SE**

  Omschrijving: SpeechWorks is een software pakket waarmee 'large vocabulary' netwerkgebaseerde spraakherkenningsdiensten, zoals informatievoorziening, communicatie en transactiediensten kunnen worden gemaakt. Voor open source is er VoiceXML-gebaseerde OpenSpeech waarmee integratie in andere applicaties mogelijk is.

  Beschikbaarheid: te koop voor het Nederlands

  Meer informatie: `http://www.speechworks.com/products/speechrec/speechworks65se.cfm`

- **Nuance7.0**

  Omschrijving: Nuance 7.0 is een software pakket voor 'large vocabulary' spraakherkenning via de telefoon. Het is beschikbaar in verschillende programmeertalen, zoals Java, C, C++, ActiveX waardoor combinaties met andere applicaties te maken zijn.

  Beschikbaarheid: te koop voor het Nederlands

  Meer informatie: `http://www.nuance.com/products/products.html`

- **ISIP ASR Toolkit**

  Omschrijving: De ISIP ASR Toolkit is een volledig functioneel spraakherkenningsysteem dat beschikbaar is voor onderzoek. Het is onder andere mogelijk akoestische modellen te trainen. Het basissysteem is geschreven in GNU C++, verschillende onderdelen van het systeem zijn geschreven in Perl, en de grafische interface in tcl/tk.

  Beschikbaarheid: gratis te downloaden

  Meer informatie: `http://www.isip.msstate.edu/projects/speech/software/index.html`

- **Lernout & Hauspie ASR 1000/T en 1000/M**

  Omschrijving: Automatische spraakherkenningssoftware voor continue spraakherkenning, geïsoleerde woordherkenning, keyword spotting of continue cijferherkenning. De systemen zijn sprekeronafhankelijk en foneemgebaseerd. ASR1000/T is ontwikkeld voor de telefoon- en telecommunicatiemarkt. ASR1000/M is ontwikkeld voor de computer en multimedia markt.

  Beschikbaarheid: te koop

  Meer informatie: `http://www.lhs.com/`

- **Lernout & Hauspie ASR 200/A**

Omschrijving: ASR software, ontwikkeld voor de auto-industrie markt, bevat geïsoleerde woord herkenning, keyword spotting en alfabetherkenning. De herkenner is woordgebaseerd en kan gebruikt worden via de telefoon.

Beschikbaarheid: te koop

Meer informatie: `http://www.lhs.com/`

- **Lernout & Hauspie Consumer ASR1600**

  Omschrijving: ASR software, geoptimaliseerd voor RISC-gebaseerde platforms (`http://www.risc.uni-linz.ac.at/`). Geschikt voor continue spraakinput, hands-free gebruik, continue cijfer of karakter input.

  Beschikbaarheid: te koop, beschikbaar voor acht talen.

  Meer informatie: `http://www.lhs/com/`

- **Lernout & Hauspie Consumer ASR300**

  Omschrijving: Deze ASR software heeft lage systeem vereisten, het is beschikbaar voor DSP en RISC-gebaseerde platforms. De software is geschikt voor natuurlijk command& control, hands-free gebruik voor Personal Digital Assistants, mobiele telefoons en spelletjes.

  Beschikbaarheid: te koop.

  Meer informatie: `http://www.lhs.com/`

- **Lernout & Hauspie Consumer ASR100**

  Omschrijving: De software is geoptimaliseerd voor goedkopere hardware. Door het uitspreken van een naam is gemakkelijke en snelle toegang tot een persoonlijk adres- of telefoonboek mogelijk.

  Beschikbaarheid: te koop.

  Meer informatie: `http://www.lhs.com/`

Neurale netwerken Onderstaande opsomming bevat een aantal systemen waarmee neurale netten voor spraakherkenning kunnen worden getraind. Deze lijst is zeker niet volledig.

- **AbbotToolkit**

  Omschrijving: De AbbotToolkit is een toolkit voor large-vocabulary sprekeronafhankelijke spraakherkenning, ontwikkeld door Cambridge University. De onderdelen van de toolkit zijn: akoestische feature extractie, akoestische modellen, uitspraakmodellen en taalmodellen. Extra features zijn onder anderen sprekeradaptatie en automatische uitspraakgeneratie. Abbot gebruikt neurale netten voor de akoestische modellering, en HMM-en voor taalmodellen. De toolkit bestaat uit een verzameling UNIX tools, gebaseerd op een C-library. De Nederlandse modellen zijn gemaakt door TNO TM.

  Beschikbaarheid: Te koop via SoftSound

  Meer informatie: `http://www.softsound.com/AbbotToolkit.html`

- **CSLU toolkit**

  Omschrijving: De CSLU toolkit is een verzameling tools voor het ontwikkelen van interactieve spraaksystemen. De toolkit bevat spraakherkenning, natuurlijke taalverwerking, spraaksynthese en visuele animatie. Met deze toolkit kunnen hybride neurale netten worden getraind.

  Beschikbaarheid: te koop

  Meer informatie: `http://cslu.cse.ogi.edu/toolkit/`
  `http://cslu.cse.ogi.edu/tutordemos/nnet_training/tutorial.html`

Uitspraaklexicon

Bestaande lexica Hieronder worden enkele lexica genoemd waarin naast orthografie ook een fonetische transcriptie opgenomen is.

- **CELEX**

  Omschrijving: Elektronische database met informatie over Nederlandse woorden (381.000 woordvormen). De database bevat informatie over orthografische kenmerken, fonologische, morfologische en syntactische eigenschappen van lemmata, over de frequentie van voorkomen op basis van het INL corpus, en over syntactische en semantische subcategorisaties.

  Beschikbaarheid: Beschikbaar op cd-rom via LDC

  Meer informatie: `http://www.kun.nl/celex/`
  `http://morph.ldc.upenn.edu/Catalog/`

- **FONILEX**

  Omschrijving: Een uitspraaklexicon voor het Nederlands in Vlaanderen (200.000 woordvormen). Fonilex is ontleend aan CELEX Deze database is gemaakt tijdens het Fonilex onderzoeksproject, gesubsidieerd door de Vlaamse regering (IWT), van januari 1995 tot december 1997.

  Beschikbaarheid: Beschikbaar voor onderzoeks- en commerciële doeleinden.

  Meer informatie: `http://bach.arts.kuleuven.ac.be/fonilex/`
  `http://www.ccl.kuleuven.ac.be/about/FONILEX.html`

- **CGN**

  Omschrijving: Binnen het CGN 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.

  Beschikbaarheid: Referenties naar het lexicon zijn onderdeel van de 3^{e} release van het corpus.

  Meer informatie: `http://www.elis.rug.ac.be/cgn/`
  `http://lands.let.kun.nl/cgn/`

- **Onomastica**

  Omschrijving: Onomastica bevat fonetische transcripties van eigennamen, zoals plaatsnamen, straatnamen, familienamen, voornamen en productnamen. De database is ontwikkeld in het Onomastica project, gesubsidieerd door het LRE programma.

  Beschikbaarheid: Niet beschikbaar.

  Meer informatie: Zie [83].

- **Speri-Data AG**

  Omschrijving: Deze database bevat fonetische transcripties van 12000 Nederlandse woorden gebruikt in het dagelijks leven.

  Beschikbaarheid: Distributie door ELRA.

Grafeem-foneemomzetters Voor grafeem-foneemconversie zie ook sectie 9.3

- **TreeTalk**

  Omschrijving: Ondertussen is er in Tilburg in het ILK project een grafeem-naar-foneem conversie programma voor het Nederlands gemaakt (met on line demo). Dit programma is onderdeel van het spraaksynthese programma TreeTalk. Het programma is getraind met behulp van de CELEX database. (d.m.v. memory based learning algorithms).

  Meer informatie: `http://ilk.kub.nl/g2p-www-demo.html`

- **CNTS**

  Omschrijving: Het Centrum voor Nederlandse Taal en Spraak maakt computerprogramma's die een tekst automatisch omzetten in fonetisch schrift, zodat een spraaksyntheseprogramma het kan uitspreken. Kenmerkend voor de aanpak van het CNTS is het gebruik van leeralgoritmen. De computer "zoekt" zelf een optimale oplossing voor een probleem zoals het omzetten van een geschreven tekst naar een uitspraak ervan.

  Beschikbaarheid: onbekend

  Meer informatie: Walter Daelemans en Steven Gillis (CNTS, Antwerpen)

- **Niros / KunTTS**

  Omschrijving: Deze grafeem-foneemomzetter is onderdeel van het volledige tekst-naar-spraaksysteem van de KUN (Zie ook paragraaf 9.1.3). Deze grafeem-foneemomzetter heeft een aantal lexica te beschikking (Celex en Onomastica) en heeft een regelgebaseerde module voor onbekende woorden. De grafeem-foneemomzetter kan ook alleen regelgebaseerd werken.

  Beschikbaarheid: Beschikbaar voor onderzoeksdoeleinden.

  Meer informatie: Joop Kerkhoff (KUN)

Taalmodellen

- **CMU**

  Omschrijving: De CMU-Cambridge Statistical Language Modeling toolkit is een verzameling UNIX tools waarmee statistische taalmodellen kunnen worden gebouwd en getest. Deze toolkit is onlangs herschreven door Philip Clarckson en Roni Rosenfeld, waardoor hij meer functionaliteit heeft, en efficiënter werkt. Met deze versie kunnen niet alleen bi- en tri-grammen getraind worden, maar alle soorten n-grammen. Ook discounting schemes worden ondersteund door deze toolkit.

  Beschikbaarheid: gratis te downloaden

  Meer informatie:

  `http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html`

  Zie ook [32].

9.3.4 Inventarisatie herkenningssoftware voor het Nederlands

Inventarisatie transcriptiesoftware Binnen het DRUID-project (`http://dis.tpd.tno.nl/druid/`) is een spraakherkenner voor het Nederlands ontwikkeld die kan worden ingezet bij de ontsluiting van multimedia-archiven met spraakdata (audi0 en/of video). De gesproken fragmenten worden met de spraakherkenningstechnologie omgezet naar tekst. De gebruikte akoestische modellen zijn ontwikkeld door TNO TM (op basis van de neurale netwerktechnologie van Abbot, zie ook sectie 9.3.3). De taalmodellen zijn ontwikkeld door de Universiteit Twente. De herkenningstechnologie werkt sprekeronafhankelijk en heeft een lexicale coverage van 65.000 woorden.

De kwaliteit van de herkenning kan worden gevolg aan de hand van twee on-line demonstraties:

- Radio1 Journaal: `http://speech.tm.tno.nl:1080/radio1/`
- 8 uur Journaal: `http://wwwhome.cs.utwente.nl/~ordelman/research/demo/druid_sdr_demo.html`

Voor meer informatie: hltgroup@cs.utwente.nl

Inventarisatie dicteersoftware Onderstaande is een lijst van enkele (commerciële) dicteersystemen, ook wel 'speech to text' genoemd, beschikbaar voor het Nederlands. Zie verder: `http://home.hccnet.nl/e.ley/speech.html` en `http://spraakherkenning.pagina.nl/`

- **Philips FreeSpeech2000**

  Omschrijving: Philips kwam in 1998 met FreeSpeech op de markt. Inmiddels is FreeSpeech 2000 uitgekomen. In tegenstelling tot de andere bedrijven, waar het kiezen tussen alle versies software soms moeilijk is, komt Philips met 1 versie, bedoeld voor zowel de consument als de zakelijke gebruiker. Wel kun je een keuze maken tussen een hoofdmicrofoon of de zogenaamde SpeechMike; een combinatie van een microfoon en een trackball, die je in je hand houdt tijdens het dicteren.

Met FreeSpeech 2000 kun je: dicteren, tekst opmaken en heel Windows met je stem besturen. Met FreeSpeech 2000 is het verder mogelijk om in meerdere talen (waaronder Engels, Frans en Duits) dicteren.

Beschikbaarheid: te koop

Meer informatie: `http://www.speech.philips.com/ds/`

- **L& H VoiceXpress**

  Omschrijving: Voice Xpress is verkrijgbaar in verschillende pakketten, van professioneel gebruik tot incidenteel gebruik. Het pakket is te incorporeren in Microsoft applicaties.

  Beschikbaarheid: te koop

  Meer informatie: `http://www.lhsl.com/voicexpress/`

- **Dragon Naturally Speaking**

  Omschrijving: Naast de dicteer modus heeft dit software pakket ook de mogelijkheid de gedicteerde tekst door middel van spraaksynthese ten gehore te brengen. Naturally Speaking kan in elk Windows programma geïncorporeerd worden.

  Beschikbaarheid: te koop

  Meer informatie: `http://www.synapseadaptive.com/NaturallySpeaking/professional.htm`

- **IBM ViaVoice**

  Omschrijving: Naast dicteersoftware voor Windows biedt ViaVoice ook pakketten aan voor integratie met Linux en Macintosh.

  Beschikbaarheid: te koop

  Meer informatie: `http://www-4.ibm.com/software/speech/desktop/`

- **MDT Talkkey**

  Omschrijving: MDT richt zich namelijk op de zakelijke markt, met gespecialiseerde spraakherkenning voor bijvoorbeeld de medische sector en de advocatuur. Daarnaast is ook Talkkey Home Office ontwikkeld ter besturing van Windows en waarmee in elke Windows applicatie gedicteerd kan worden.

  Beschikbaarheid: te koop

  Meer informatie:

  `http://www.mdt.nl`
  `http://home-1.worldonline.nl/~\sim~ebax/talkkey.html`

9.4 Foonstringbewerkingen

Steeds meer spraaktechnologische toepassingen maken gebruik van corpusgebaseerde methoden. Hiervoor is beschikbaarheid van een geannoteerd corpus noodzakelijk. Corpora kunnen op veel verschillende niveaus worden geannoteerd. Voor het bouwen van tekst-naar-spraaksystemen bijvoorbeeld, zijn fonetische transcripties en prosodische annotaties nodig die in tijd zijn opgelijnd met het spraaksignaal. Voor prosodische annotaties zie sectie 9.1.6.

Deze sectie gaat over automatische fonetische transcriptie en segmentering, over het oplijnen van twee foonstrings, en over het berekenen van de afstand tussen twee foonstrings.

Transcriptie & segmentering Voor spraaktechnologische toepassingen zijn meestal grote hoeveelheden data nodig. Het manueel transcriberen en segmenteren van een grote hoeveelheid spraak is zeer tijdrovend. Bovendien is een menselijke (handmatige) transcriptie subjectief, en is de kans op inconsistenties en fouten groot. Een automatische transcriptie is misschien niet foutloos, maar zeker efficiënter en consistenter, zie [91].

Omdat de term transcriptie op verschillende manieren gebruikt wordt, geven we hier een korte toelichting. Er bestaan verschillende soorten transcripties. Allereerst zijn er orthografische transcripties (reeksen grafemen), supra-segmentele / prosodische transcripties (reeksen prosodische symbolen, zie sectie 9.1.5), en segmentele transcripties (reeksen klanksymbolen). Hier behandelen we segmentele transcripties. Maar ook die bestaan er in vele soorten, zoals bijv. fonemische transcripties (reeksen fonemen) en fonetische transcriptie (reeksen fonetische symbolen). De belangrijkste verschillen tussen segmentele symbolen zijn de mate van detail, en de gebruikte verzameling klanksymbolen. De hoeveelheid detail in spraaktechnologie is meestal van fonemisch niveau. Maar omdat set van eenheden die in spraaktechnologie gebruikt wordt vaak niet precies overeen komt met de (standaard) foneeminventarisatie, en omdat soms ook subfonemische (allofonische) eenheden worden gebruikt, spreekt men in spraaktechnologie vaak van een foontranscriptie.

Een verzameling klanksymbolen die gebruikt wordt om segmentele transcripties te maken heet een fonetisch alfabet (PA: Phonetic Alphabet). Het internationaal meest bekende fonetisch alfabet is IPA (International Phonetic Alphabet). Omdat in IPA veel symbolen zitten die niet (standaard) op de meeste computers aanwezig zijn (geen ASCII), zijn daarnaast ook zgn. Computer Phonetic Alphabets (CPAs) ontwikkeld. Deze ontwikkeling is gebeurd in vele landen, en per land vaak nog op meerdere plaatsen, zodat er nu wereldwijd zeer veel CPAs bestaan. Ook voor het Nederlands worden verschillende CPAs gebruikt (zie bijv. Celex, Fonilex, en CGN).

Oplijning & afstandsberekening Met oplijning wordt hier bedoeld de oplijning van twee foonstrings, en met afstandsberekening de berekening van de afstand tussen twee foonstrings (op basis van een gemaakte oplijning). Deze technieken worden gebruikt voor verschillende doeleinden. Bijvoorbeeld voor het evalueren van transcripties, zowel handmatige als automatische transcripties. Bij handmatige transcripties kan zowel inter- als intratranscribent agreement berekend worden, en hiervoor is oplijning en afstandsberekening nodig. Beide technieken zijn ook nodig bij het evalueren van automatische transcripties, waarbij de automatische transcriptie meestal vergeleken wordt met een door experts gemaakte referentie transcriptie. Ook in de context van fonetisch zoeken in databases en voor de evaluatie van merknamen wordt afstandsberekening tussen twee foonstrings gebruikt. Als laatste toepassing noemen we hier het modelleren van uitspraakvariatie voor automatische spraakherkenning. Ook hierbij wordt veelvuldig gebruik gemaakt van oplijning, en zou afstandsberekening ook toegepast kunnen worden. Zie ook [79].

9.4.1 State of the art internationaal

Transcriptie & segmentering De laatste jaren is er een duidelijke toename waarneembaar in het onderzoek naar automatische transcriptie en segmentering van spraak. Zie bijv. [90,

98, 116, 141].

Met behulp van spraakherkenners kunnen segmentaties en fonetische transcripties automatisch worden verkregen, vooral wanneer een orthografische transcriptie van het spraaksignaal ook beschikbaar is.

Er worden veel verschillende technieken gebruikt voor automatische segmentering en labelling, voor een overzicht zie [98].

Uit alle publicaties kunnen de volgende conclusies getrokken worden: *Automatische segmentatie is goed mogelijk, gegeven een accurate transcriptie van het spraaksignaal.*

Automatische transcriptie is op dit moment nog te moeilijk: de transcripties die automatisch gegenereerd worden in een sprekeronafhankelijk systeem voldoen niet aan de kwaliteitseisen, zij zijn minder accuraat dan menselijke transcripties. Vanzelfsprekend zijn de kwaliteitseisen voor verschillende toepassingen verschillend. Voor sprekerafhankelijke automatische transcriptiesystemen geldt wel dat de automatisch gegenereerde transcripties van hetzelfde niveau zijn als die van een menselijke transcribent. Het gebruik van uitspraakvarianten helpt de performance van transcriptiesystemen te verbeteren.

Oplijning & afstandsberekening Het oplijnen en berekenen van de afstand tussen twee foneemstrings is minder triviaal dan het op het eerste gezicht lijkt. Voor een juiste oplijning moet bepaald worden welk symbool in de ene string overeenkomt met welk symbool in de tweede string. Om de oplijning te maken, wordt meestal gebruik gemaakt van *dynamic programming*, waarbij alle substituties, inserties en deleties een bepaalde strafwaarde krijgen. Deze strafwaarden worden vastgelegd in een matrix. Vaak worden simpele algoritmen gebruikt waarbij de straf voor een substitutie, insertie en deletie altijd 1 is (en 0 als de symbolen in beide strings identiek zijn). Echter, dit levert in een aantal gevallen een sub-optimale oplijning op. Zie onderstaand voorbeeld:

1) a. `/ A m s t @ R d A m /`

1) b. `/ A m s # @ t a: n # /`

Het symbool # in het bovenstaande voorbeeld stelt geen klank / segment voor (het representeert dus eigenlijk 'niets'), en wordt alleen gebruikt om de oplijning inzichtelijker te maken. Als 1a het uitgangspunt is, en 1b het resultaat, dan stelt iedere # een deletie voor van het corresponderende element in 1a.

In voorbeeld 1 worden twee deleties en drie substituties gevonden. Deze oplijning is echter niet optimaal, vnl. omdat de strafwaarde slechts 0 of 1 kan zijn. Daarom zijn andere oplijnalgoritmes ontwikkeld waarbij er meer gradaties zijn in de strafwaarden (dan alleen 0 of 1). Een mogelijkheid is om de symbolen in te delen in klassen (bijv. klinkers en consonanten), en een grotere strafwaarde te gebruiken voor verwisselingen tussen klassen, dan voor verwisselingen binnen een klasse [35]. Een andere mogelijkheid is om de afstanden tussen symbolen te baseren op zgn. verwarringsmatrices (die weer op verschillende manieren verkregen kunnen worden, bijv. ook m.b.v. een ASH) [128]. De afstand tussen twee symbolen (fonen) kan ook gebaseerd worden op de fonetische / fonologische features [37]. Dat resulteert voor het bovenstaande voorbeeld in de volgende oplijning:

2) a. `/ A m s t @ R d A m /`

2) b. `/ A m s # @ # t a: n /`

Het zal meteen duidelijk zijn dat deze 2e oplijning beter is dan de 1e oplijning. Ook in dit geval worden weer 2 deleties en 3 substituties gevonden. Voor een simpel algoritme, met alleen 0 en 1 als strafwaarde, zou de afstand van deze oplijning dan ook gelijk zijn aan die

van oplijning 1. Echter, meer geavanceerde algoritmes zullen een kleinere afstand vinden bij de 2e oplijning.

Bij een simpel algoritme, met als strafwaarden alleen 0 en 1, is alleen de symboolset taalafhankelijk. Bij de meer geavanceerde algoritmes zijn ook de kosten taalafhankelijk. Enkele van dergelijke meer geavanceerde algoritmes zijn al beschreven in de literatuur (zie hierboven), en kunnen vrij eenvoudig gereconstrueerd worden.

Voor meer informatie, zie [35, 37, 38, 97, 102, 145].

9.4.2 Evaluatiecriteria

Transcriptie & segmentering Het evalueren van automatische transcripties gebeurt meestal door deze te vergelijken met een manuele referentietranscriptie. Hetzelfde geldt voor segmentatie. De referentie transcriptie/ segmentatie is meestal een consensus transcriptie/segmentatie van twee of meer getrainde menselijke transcribenten. Vervolgens worden dan de volgende berekeningen gemaakt:

Voor transcriptie:

- het aantal verschillen (hoeveel symbolen wijken na een oplijning af van de manuele referentietranscriptie)
- de gemiddelde afstand tussen beide transcripties

Voor segmentatie:

- hoever (in milliseconden) wijkt de automatisch gezette grens gemiddeld af van de handmatig gezette grens.

Oplijning & afstandsberekening Afstandsberekening van foneemstrings is een evaluatiemethode voor transcriptie en segmentering, zie 9.4.2. De methode die gekozen wordt, bijvoorbeeld foneemklassen of strafwaardegradaties, om de afstanden te bepalen zorgen voor verschillende uitkomsten en zijn niet met elkaar te vergelijken. Het is dan ook lastig evaluatiecriteria op te stellen.

9.4.3 Inventarisatie beschikbare software

Transcriptie & segmentering

- **ATRANOS (Automatic TRAnscription and NOrmalization of Speech)**

  Omschrijving: Het doel van het ATRANOS project is de ontwikkeling van betere automatische transcriptiesystemen voor spraak in het algemeen, en voor het Nederlands in het bijzonder. In dit project wordt onder andere gewerkt aan de segmentatie van een audiostroom in homogene segmenten, en aan de normalisatie van transcripties, met als case-study het automatisch genereren van ondertiteling.

  Beschikbaarheid: Niet beschikbaar.

  Meer informatie: `http://atranos.esat.kuleuven.ac.be/`

Oplijning & afstandsberekening

- **Align**

  Omschrijving: Align is een programma dat twee foneemstrings optimaal met elkaar oplijnt. Deze tool is ontwikkeld in het kader van het promotie-onderzoek van Catia Cucchiarini en maakt gebruik van fonetische kenmerken om te berekenen hoe groot de afstand is tussen twee fonemen.

  Beschikbaarheid: Beschikbaar voor onderzoeksdoeleinden.

  Meer informatie: Catia Cucchiarini (KUN), Judith Kessens (KUN). Zie ook [51].

- **Levenshtein Demo (Peter Kleiweg - RUG)**

  Omschrijving: Dit is een eenvoudige demonstratie van een oplijnmethode (Levenshtein) waarbij gekozen kan worden voor een gewone oplijning en een oplijning gebaseerd op features. Bij de gewone oplijning moet de gebruiker zelf de kosten definiëren voor substituties en inserties, en deleties. Bij de oplijning aan de hand van features wordt gebruik gemaakt van fonetische / fonologische features van de symbolen. Er moet dan wel SAMPA gebruikt worden. Het tooltje toont als resultaat de matrix met de kosten en het optimale pad daar doorheen.

  Beschikbaarheid: kosteloos te downloaden

  Meer informatie: `http://odur.let.rug.nl/~kleiweg/lev/levenshtein.html`

- **NIST tools**

  Omschrijving: Op `http://www.nist.gov/speech/tools/` zijn enkele tools (aldistsm-1.2, seg_scr.v21) te vinden die gebruikt kunnen worden voor evaluatie en scoring van segmentaties en oplijningen.

  Beschikbaarheid: gratis te downloaden

  Meer informatie: `http://www.nist.gov/speech/tools/`

9.5 Robuuste spraakherkenning

Spraak is een signaal waarvan het energiespectrum continu in de tijd verandert. De eerste stap in een automatische spraakherkenner is dan ook het bepalen van een reeks opeenvolgende energiespectra uit de gesproken uiting. De reeks opeenvolgende energiespectra wordt vervolgens omgezet in een reeks waargenomen klankeigenschappen. Daarna wordt tijdens de herkenning de waargenomen reeks klankeigenschappen vergeleken met alle mogelijke reeksen van opeenvolgende klankmodellen en wordt de best passende reeks klankmodellen bepaald. De woorden die horen bij deze best passende reeks klankmodellen worden tenslotte gepresenteerd als de herkende reeks woorden.

Wanneer er extra geluid aanwezig is terwijl de spreker aan het woord is, of wanneer het spraakgeluid door een telefoon naar de automatische spraakherkenner wordt geleid, dan zal het door de herkenner waargenomen geluid niet precies meer overeenkomen met het in stilte waargenomen geluid of het geluid dat niet door een telefoon is gegaan. Met andere woorden, als gevolg van de opname- of spreekcondities treedt een verandering op in de reeks waargenomen energiespectra. Deze verandering leidt tot een verandering in de reeks waargenomen klankeigenschappen en dit heeft tot gevolg dat de reeks klankmodellen die als

beste past zal kunnen afwijken van de reeks klankmodellen die met de werkelijk gesproken klanken correspondeert. Het eindresultaat is dat de reeks herkende woorden zal kunnen afwijken van de werkelijk gesproken woorden. Het zal duidelijk zijn dat naarmate de invloed van de opnamecondities op de waargenomen reeks energiespectra groter is, de afwijking tussen hetgeen werkelijk gesproken en hetgeen als beste herkend is ook groter zal worden. Het doel van robuuste automatische spraakherkenning is nu om de invloed van de opnamecondities (zoals extra geluiden in de achtergrond of de invloed van het opnamekanaal) op de prestaties van de automatische spraakherkenner zoveel mogelijk te verkleinen. Met andere woorden: robuustheidstechnieken worden gebruikt om de betrouwbaarheid van het herkenresultaat in opnameomstandigheden die niet ideaal zijn toch zo goed mogelijk op het niveau te houden dat in ideale opnameomstandigheden wordt bereikt.

Er kunnen twee belangrijke soorten robuustheidstechnieken worden onderscheiden:

1. kanaalnormalisatietechnieken
2. ruisrobuustheidstechnieken

Kanaalnormalisatietechnieken zijn belangrijk in toepassingen waarbij spraaksignalen via een onvoorspelbaar transmissiekanaal worden waargenomen, zoals bij automatische spraakherkenning over de telefoon. De eigenschappen van het kanaal zijn in dergelijke toepassingen bepaald door de microfoon in de hoorn en door de verbinding tussen het toestel en de centrale. De spectrale doorlaatkarakteristiek van de microfoon en van de verbinding zijn tevoren niet aan de herkenner bekend. De doorlaatkarakteristiek varieert bovendien aanzienlijk van toestel tot toestel, en tussen verbindingen. Op deze wijze wordt er bij automatische spraakherkenning over de telefoon een bron van variatie geïntroduceerd die niet is gerelateerd aan de spraakklanken en die daarmee irrelevant (en potentieel hinderlijk) is voor de herkenning. Kanaalnormalisatietechnieken beogen om de effecten van deze variatiebron op de spectra van de spraakklanken zoveel mogelijk te verminderen.

Wanneer er achtergrondgeluiden aanwezig zijn tijdens de herkenning, dan "hoort" de herkenner in feite de combinatie van het achtergrondgeluid en de spraakgeluiden. Wanneer het achtergrondgeluid van tevoren bekend is, kunnen één of meerdere speciale modellen getraind worden die de herkenner helpen om spraak en achtergrondgeluid van elkaar te scheiden. Maar als het achtergrondgeluid onvoorspelbaar is (wat in de praktijk van mobiele telefonie erg vaak voorkomt), dan wordt een onbekende verstoring geïntroduceerd in de waarnemingen. Omdat de spraakklankmodellen die de herkenner gebruikt, getraind zijn met spraak zonder die verstoringen, ontstaat er een discrepantie tussen de condities gebruikt tijdens de training en het daadwerkelijke gebruik. Ruisrobuustheidstechnieken beogen de aanwezigheid van verstoorde waarnemingen als gevolg van de aanwezigheid van achtergrondgeluid zoveel mogelijk te verminderen.

9.5.1 State of the art internationaal

Er wordt in de hele wereld steeds meer onderzoek gedaan naar robuustheid van spraakherkenners. Spraakherkenners zijn steeds vaker robuust tegen allerlei soorten van sprekervariabiliteit en akoestische variabiliteit. Zo zijn de meeste huidige systemen sprekeronfhankelijk en zijn veel spraakherkenners tot op zekere hoogte bestand tegen akoestisch slechte signalen, veroorzaakt door zowel achtergrondruis als door vervormingen van het transmissiekanaal. Toch is in real-life applicaties de performance van spraakherkenners nog steeds veel slechter

in geval er sprake is van achtergrondlawaai of van telefoonspraak. Ook het herkennen van spraak met andere menselijke spraak op de achtergrond is een groot probleem. Op dit gebied zijn totnogtoe nog niet erg goede resultaten behaald.

Er komen steeds meer systemen die op basis van een kleine hoeveelheid binnenkomende spraak on line adaptatie doen aan de omstandigheden waarin het geluid werd opgenomen.

Technieken om de robuustheid van automatische spraakherkenners te verbeteren, kunnen worden toegepast op verschillende niveau's:

1. het niveau waar de features worden geëxtraheerd
2. het niveau van de akoestische modellen
3. het niveau van het zoekalgoritme

ad 1.

A. Cepstral mean normalisatie

Deze techniek werkt vooral goed voor het onderdrukken van langzaam variërende akoestische omgevingen (bijv. transmissiekanalen). Bij cepstral mean normalisatie wordt van het inkomend signaal voor iedere cepstrale waarde het gemiddelde per tijdvenster vastgesteld, welke vervolgens wordt afgetrokken van de waarde voor ieder tijdsframe. De achterliggende gedachte is de aanname dat het gemiddelde van het cepstrum de kanaalvervorming weergeeft (dit kan echter alleen bij een long term cepstrum). Zie bijvoorbeeld [58]. Short-term toepassingen van cepstral mean normalisatie gaan er vaak vanuit dat de kanaalvervorming maar langzaam verandert ten opzichte van de spraak. Een groot voordeel van deze techniek is dat deze in real-time kan worden toegepast, in dat geval wordt gebruik gemaakt van een kort tijdvenster ([7]).

B. Spectral subtraction

Spectral subtraction wordt vooral toegepast om om te gaan met "additive" ruis in spraak. Hierbij wordt het spectrum van de ruis gemeten in een periode wanneer alleen ruis aanwezig is (vastgesteld door een Voice Activity Detector) en dit spectrum wordt vervolgens afgetrokken van het binnenkomende signaal, zodat een schatting ontstaat van de schone spraak. Het voordeel van deze methode is dat hij relatief eenvoudig is, en dat er alleen een schatting van de noise nodig is, zonder aannames over het signaal zelf. Een nadeel is dat als artefact van deze methode vaak een zgn. musical noise wordt gegenereerd, doordat de schatting van de ruis altijd de echte waarde over- of onderschat. Voor het verbeteren van verstaanbaarheid en kwaliteit van breedband achtergrondruis is spectral subtraction ineffectief gebleken. (`http://cslu.cse.ogi.edu/nsel/wan_manuscript/node10.html`)

Voor meer informatie, zie [15] en [9] en [142].

C. Noise masking

Maskering is een psychologisch fenomeen waarbij de perceptie van spraak achteruitgaat wanneer er bijvoorbeeld ruis aanwezig is. D.w.z dat mensen geluid niet kunnen waarnemen wanneer het energieniveau daarvan lager is dan die van een ander geluid dat tegelijkertijd wordt afgespeeld. Dit fenomeen kan worden gebruikt om ruis te maskeren. Toegepast in

spraakherkenning, verkleint noise masking in het spectrale domein de bijdragen van de delen met weinig energie. Alleen de frequentiegebieden in het spectrum die een hoger energieniveau hebben dan de masking drempelwaarde, worden gebruikt voor herkenning. Op deze manier worden spectra zodanig getransformeerd dat ze minder gevoelig worden voor variaties in achtergrond ruis.

Voor meer informatie, zie [131] en [81]

D. LDA

LDA (Linear Discriminant Analyse) is een techniek die veelvuldig is toegepast in de patroonherkenning om het aantal dimensies / klassen in de feature space te reduceren. Hiertoe wordt een lineaire transformatie gedaan van de features door de within-classverschillen te minimaliseren en de between-classverschillen te maximaliseren. Behalve voor patroonherkenning, is ook voor robuuste spraakherkenning de toepasbaarheid van deze techniek bewezen, zie [14] en [118].

ad 2.

A. MLLR

MLLR staat voor Maximum Likelihood linear regression. MLLR schat een aantal lineaire transformaties voor de gemiddelden en variaties in een HMM op basis van een set adaptatiedata, om zo de likelihoods te maximaliseren. Hoewel MLLR oorspronkelijk bedoeld was voor sprekeradaptatie, kan deze techniek ook worden gebruikt voor omgevingsnormalisatie, omdat het ook de mismatch tussen train en test condities verkleint wanneer het additive noise of kanaal betreft. Deze methode blijkt vooral goed te werken wanneer slechts weinig adaptatiemateriaal beschikbaar is. Voor meer informatie, zie [62].

B. MAP adaptation

MAP (Maximum a Posteriori) adaptation is een vorm van directe adaptatie, waarbij de akoestische eenheden worden herschat waarvoor adaptatiedata beschikbaar is. Bij MAP adaptation wordt de adaptatiedata gecombineerd met a priori kennis over de model parameters gegeven hun a priori distributie. Hierbij worden niet alleen de gemiddelden en variaties aangepast op basis van het adaptatiemateriaal, maar ook de mixture weights, de initiële waarschijnlijkheden en de transitie waarschijnlijkheden. Voor een significante verbetering van de performance is echter een redelijk grote hoeveelheid data nodig. Voor meer informatie, zie [61]. Voor een overzicht van artikelen over MAP adaptation, zie `http://www.bell-labs.com/org/1133/Research/SpeechRecognition/sv-adaptation.html`

ad 3.

Missing Feature theory

Missing feature theorie wordt toegepast als een manier om het zoekalgoritme minder gevoelig te maken voor de aanwezigheid van ruis. Bij MFT worden alle feature componenten opgedeeld in delen die niet zijn aangetast door ruis, en delen die wel zijn aangetast door ruis. Tijdens de herkenning worden de delen die onbetrouwbaar zijn omdat ze ruis bevatten genegeerd of voorzichtig behandeld, en wordt alleen of voornamelijk gebruik gemaakt van

die delen die wel betrouwbaar zijn. Zie ook [88] en [46].

Een overzicht van literatuur over robuuste spraakherkenning is te vinden op `http://www.dcs.shef.ac.uk/~jeremy/litrev.htm`.

9.5.2 Evaluatiecriteria

Net als standaard spraakherkenners worden robuuste spraakherkenners veelal geëvalueerd in termen van Word Error Rate. In het geval van robuuste spraakherkenning wordt vaak gekeken naar het gedrag van WER in verschillende ruiscondities.

9.5.3 Inventaris beschikbare software

Er is geen software specifiek voor het Nederlands bekend.

9.6 Non-native spraakherkenning

Voor spraakherkenners in de meeste toepassingen geldt, dat ze in staat moeten zijn om spraak te herkennen van veel verschillende personen, dialecten, accenten, etc. De meeste spraakherkenners tegenwoordig zijn redelijk robuust tegen sprekervariabiliteit en andere oorzaken voor mismatches tussen de condities tijdens training, en de modellering in het lexicon en de test data, maar robuustheid tegen sprekers die geen moedertaalsprekers zijn, is nog niet zo hoog. Een bemoeilijkende factor is, dat de non-native sprekers geen homogene groep vormen: er zijn veel verschillende accenten. Een van de belangrijkste toepassingsgebieden voor non-native spraakherkenning is CALL (Computer Aided Language Learning), zie [31, 30, 53, 3].

Er zijn een aantal verschillende niveaus waarop non-native spraak kan worden gemodelleerd.

- *Specifieke modellen.*

  Ten eerste kan non-nativeness worden aangepakt tijdens de training van de akoestische en taalmodellen. Hierbij wordt een corpus van non-native spraak toegevoegd aan het corpus van native spraak, waarbij de twee verschillende soorten verder als gelijk worden beschouwd. De mismatch tussen train- en testdata is hiermee niet verdwenen, maar in elk geval wel gereduceerd doordat een deel van de traindata gelijk is aan de testdata. Een andere aanpak is om de twee soorten data niet tijdens de training bij elkaar op te tellen, maar om aparte modellen te trainen voor native en non-native spraak. Tijdens het herkenproces kunnen deze modellen dan ofwel sequentieel, ofwel parallel worden gebruikt. Wanneer ze sequentieel worden gebruikt, wordt eerst vastgesteld of het om native of non-native spraak gaat, en worden vervolgens de juiste modellen ingezet. Wanneer de modellen parallel worden gebruikt wordt niet specifiek vastgesteld of het om native of non-native spraak gaat, maar wordt dat model gekozen dat het beste matcht.

- *Modeladaptatie.*

  Omdat er meestal niet voldoende non-native spraak voorhanden is om volledige modellen mee te trainen, wordt vaak modeladaptatie toegepast. Hierbij worden de bestaande native modellen aangepast met een beperkte hoeveelheid non-native spraak.

- *Uitspraakvarianten in het lexicon.*

  Een ander niveau waarop non-native spraak aangepakt kan worden, is het niveau van het lexicon. Hierbij worden in het lexicon uitspraakvarianten opgenomen die beter passen bij de uitspraak van non-natives dan de canonieke transcripties.

9.6.1 State of the art internationaal

Het gebruik van specifieke akoestische modellen die getraind zijn op non-native spraak werkt niet optimaal, omdat de non-native sprekers een niet-homogene groep vormen. Wanneer echter gebruik wordt gemaakt van specifieke modellen voor specifieke accenten, kunnen wel significant betere resultaten worden geboekt. Hiervoor moet echter veel traindata beschikbaar zijn. Deze methode is met succes toegepast voor zowel non-native spraak als voor dialect accenten, zie bijvoorbeeld [126].

Omdat vaak niet genoeg data beschikbaar is om specifieke akoestische modellen te trainen, wordt meer gebruikt gemaakt van adaptatie. In een aantal studies is aangetoond dat de geadapteerde modellen veel beter scoren dan de native modellen en/of modellen die alleen op non-native spraak zijn getraind. Het is echter niet zo, dat het bijtrainen van modellen met uitingen met een bepaald accent ook direct een positieve invloed heeft op de herkenning van andere sprekers met hetzelfde accent [126].

Ook modellering van non-native spraak in het lexicon is met succes toegepast, bijvoorbeeld door Teixeira en door Bonaventura [16]. Echter, in veel onderzoeken is ook gezien dat het toevoegen van uitspraakvarianten aan het lexicon leidt tot een grotere verwarbaarheid, waardoor de stijging in performance door de betere modellering teniet wordt gedaan.

Voor meer informatie, zie [16, 31, 132, 53, 89, 2, 126].

9.6.2 Evaluatiecriteria

Er zijn geen specifieke evaluatiecriteria voor herkenning van non-native spraak. Voor evaluatie van spraakherkenners in het algemeen: zie sectie 3.3.

9.6.3 Inventaris beschikbare software

Non-native spraakherkenning is een (aangepaste) vorm van spraakherkenning, en daarom zal als uitgangspunt gebruik gemaakt kunnen worden van standaardspraakherkenners (zie sectie 9.3). Verder zullen technieken uit spreker- en lexiconadaptatie (zie sectie 9.9.1) waarschijnlijk ook bruikbaar zijn. Maar naar non-native spraakherkenning zelf is nog weinig onderzoek gedaan, en bijgevolg is er eigenlijk geen software beschikbaar specifiek voor non-native spraakherkenning. Wat in ieder geval nodig is voor het ontwikkelen van dergelijke software zijn corpora met spraak van non-natives. Op dit moment zijn er een aantal dergelijke corpora beschikbaar (zie ook hoofdstuk 10):

- ESFSLD (Migranten Databank) (1982-1987)
- Multi Tongue Dutch I & II (1997)

9.7 Sprekerherkenning: identificatie, verificatie en tracking

Met sprekerherkenning wordt bedoeld het automatisch herkennen van wie er aan het woord is op basis van specifieke individuele kenmerken in het spraakgeluid.

Sprekerherkenning kan worden opgedeeld in:

1. sprekeridentificatie,
2. sprekerverificatie en
3. spreker-tracking.

Het doel van *sprekerverificatie* is vast te stellen of de spreker daadwerkelijk degene is die hij claimt te zijn. Bij *sprekeridentificatie* is de taak de identiteit van de spreker uit een eindige set sprekers vast te stellen. *Spreker-tracking* betekent vaststellen welke spreker wanneer aan het woord is.

Onderstaand figuur toont de verschillen tussen sprekeridentificatie en sprekerverificatie: [84])

Figuur 2: Sprekerverificatie vs. sprekeridentificatie

Toepassingsgebieden van sprekerherkenning zijn het beveiligen van telefonische diensten, forensische toepassingen, informatietoegang via internet, etc.. Sprekeridentificatie wordt het meest gebruikt in forensische toepassingen, terwijl sprekerverificatie meestal wordt toegepast ter beveiliging van diensten. Tenslotte bestaat er nog zoiets als open-set-identificatie. Dit is een specifiek soort sprekeridentificatie, waarbij een referentiemodel voor een onbekende spreker niet bestaat, dit is meestal het geval in forensische toepassingen. In dit geval moet er een extra beslissingscriterium, namelijk 'matcht met geen van de modellen' worden vastgesteld.

Er zijn verschillende soorten sprekerherkenningssystemen:

- tekstafhankelijke systemen
- tekstonafhankelijke systemen

Voor tekstafhankelijke systemen geldt dat de spreker keywords of zinnen moet uitspreken die gelijk zijn aan de zinnen / woorden in de trainingsfase (vaak is dit een wachtwoord). Voor tekstonafhankelijke systemen geldt dit niet. Probleem met tekstafhankelijke systemen is echter, dat ze makkelijk kunnen worden misleid door spraak op te nemen van een toegestane spreker. Om deze reden wordt vaak gebruik gemaakt van methodes waarbij een kleine verzameling woorden (bijv. cijfers) wordt gebruikt als keywords en de gebruiker wordt gevraagd een bepaalde (random gekozen) sequentie van die woorden uit te spreken, hetzij van tekst voor te lezen of na te zeggen van een geluidsfragment.

Sprekeridentificatie, -verificatie en -tracking hebben gemeen, dat niet zozeer de software die wordt gebruikt taalafhankelijk is, als wel de data waarmee de referentiemodellen getraind worden. Het betreft hier fine-tuning, d.w.z. de identificatie, tracking of verificatie gaat beter wanneer de train en test data meer met elkaar overeenkomen, ook in taal.

Voor meer informatie, zie [59, 84].

Onderstaande website geeft een uitgebreid overzicht van technologie en toepassingen op het gebied van sprekerherkenning:
`http://www.ispeak.nl/start.html`

9.7.1 State of the art internationaal

Inmiddels wordt er op internationaal niveau veel onderzoek naar sprekerherkenning verricht, bijvoorbeeld door MIT, Dragon, Nuance, Idiap, OGI, etc.

De laatste jaren is veel vooruitgang geboekt in sprekerherkenning, maar er blijft een aantal problemen liggen. Net als bij spraakherkenning blijft bij sprekerherkenning variabiliteit in zowel sprekers als opnamecondities en kanaaleigenschappen de grootste bottleneck. Een ander probleem is het feit dat alleen de akoestische kenmerken van de spreker beschikbaar zijn om de identiteit vast te stellen of te verifiëren. Het gebruik van andere biometrische kenmerken zou het proces vergemakkelijken.

De *state of the art* in sprekerherkenning wordt voor een belangrijk deel bepaald tijdens de NIST Sprekerherkenningsevaluaties die al sinds 1996 ieder jaar worden uitgevoerd. Het doel van deze evaluaties is het voortdrijven van de technologie, het vaststellen van de *state of the art*, en het vinden van de meest veelbelovende algoritmen en benaderingen. In de loop der jaren hebben 16 onderzoeksinstituten meegedaan met verschillende evaluaties. Hierbij krijgen zij ieder jaar een andere taak (handset, spreker-tracking, meerdere sprekers tegelijk, etc.) waarvoor zijn hun eigen software dienen te ontwikkelen, welke vervolgens geëvalueerd worden op één algemene testset. De NIST evaluaties van 1997 – 1999 zijn uitgebreid beschreven in een aantal tijdschriftartikelen en papers in conferentie proceedings.

Meer informatie over NIST: `http://www.itl.nist.gov/iad/894.01/tests/spk/2000/index.htm`.

Overzichtsverhalen over sprekerherkenning: [94, 1].

9.7.2 Evaluatiecriteria

Sprekeridentificatie- en *sprekertracking* systemen worden doorgaans geëvalueerd in termen van het aantal misclassificaties. D.w.z. de spreker wordt aangezien voor een andere spreker.

Ook *sprekerverificatiesystemen* worden geëvalueerd in termen van misclassificatie, echter hier kunnen twee typen classificatiefouten optreden.

Bij *sprekerverificatie* kunnen twee typen fouten optreden:

1. false rejections: wanneer een toegestane spreker wordt afgewezen.
2. false acceptances: wanneer een inbreker wordt geaccepteerd als de spreker die hij claimde te zijn.

De performance van het systeem (het aantal false acceptances vs. het aantal false rejections) hangt sterk af van de drempelwaarde voor het al dan niet accepteren van een spreker. Om de performance weer te geven onafhankelijk van een specifieke drempelwaarde, kan een Receiver Operating Characteristic (ROC) curve worden gebruikt, zie Figuur 3.

Figuur 3: Voorbeeld van een ROC curve

Zowel voor ontwikkeling als voor evaluatie van sprekerherkenningssystemen is een geannoteerde spraakdatabase nodig. Voor het Nederlands zijn een aantal van dergelijke corpora beschikbaar:

- SESP I + II + III (Cave)
- SpeechDat
- JOHO
- PIG I & II (Picasso)

9.7.3 Inventaris beschikbare software

Voor het Nederlands zijn geen off-the-shelfproducten beschikbaar. Wel is in het kader van Europese projecten een aantal tools ontwikkeld.

- **HTK scripts (uit de *Picasso* en *Cave* projecten)**

  Omschrijving: In het kader van de Picasso en Cave projecten is aan de Afdeling Taal en Spraak van de KUN door Johan Koolwaaij sprekerverificatiesoftware gemaakt op basis van HTK. De software bestaat uit een verzameling scripts om alle stappen in de verificatie-experimenten uit te voeren en gebruikt HTK als de search-engine. Het is een generiek systeem waarmee verschillende aanpakken kunnen worden uitgeprobeerd (ergodische HMM-en, Gaussian mixture modelling, etc.) en dat gebruikt kan worden voor zowel tekstafhankelijke en tekstonafhankelijke verificatie. Deze software wordt getraind met spraak van de sprekers die gebruik willen maken van de dienst, en de wereld modellen die gebruikt worden voor de verificatie kunnen worden getraind op normale spraakdatabases die worden gebruikt voor sprekeronafhankelijke spraakherkenning (cf. sectie 9.3).

  Om gebruik te kunnen maken van deze software is het volgende nodig:

  - het bovengenoemde generieke systeem
  - een Unix omgeving
  - HTK (versie 2.1 of later)
  - een sprekerverificatie database (bijv. SESP)
  - een database om wereldmodellen te trainen (bijv. Polyphone).

  Meer informatie: `http://www.PTT-Telecom.nl/cave`

  Zie ook: [11, 74, 84].

9.8 Taal- en dialectidentificatie

Taal- en dialectidentificatie is het proces dat gebruik maakt van specifieke kenmerken in het geluidssignaal om vast te stellen welke taal, dialect of accent wordt gesproken (kortweg gesproken taalidentificatie). Er zijn twee soorten toepassingen denkbaar voor taalidentificatiesystemen: 1) als voorbewerking voor automatische systemen, en 2) als voorbewerking voor menselijke operators. Voor het eerste toepassingsgebied kun je bijvoorbeeld denken aan een meertalig vertaalsysteem. In het tweede toepassingsgebied wordt taalidentificatie bijvoorbeeld toegepast om telefoontjes direct door te schakelen naar een operator die de betreffende taal spreekt, bijvoorbeeld voor het afhandelen van oproepen bij de alarmcentrale (112).

Er zijn veel verschillende kenmerken die kunnen helpen bij het identificeren van de taal, veel van deze kenmerken worden ook door mensen gebruikt:

- fonologische kenmerken
- morfologische kenmerken
- syntactische kernmerken
- prosodische kenmerken.

Bijna alle systemen voor gesproken taalidentificatie bestaan uit een trainingsfase en een herkenningsfase. In de trainingsfase wordt het systeem gevoerd met spraak uit de doeltalen. Gebruikmakend van verschillende taalafhankelijke kenmerken uit deze spraak worden vervolgens de doeltalen gemodelleerd. Tijdens de herkenning worden deze features vergeleken met dezelfde features uit de te testen uiting, en op basis hiervan wordt besloten welke taal is gesproken.

Verschillende identificatiesystemen verschillen in de volgende dimensies:

- welke akoestische representatie wordt gebruikt? (bijv. prosodische features, fonemen, etc.)
- wat is de mate van akoestische modellering? (taalmodellen, meerdere akoestische modellen voor iedere taal, grammatica's)
- tekstafhankelijkheid of tekstonafhankelijkheid

Meer informatie: [72, 101, 147].

9.8.1 State of the art internationaal

Het onderzoek naar taalidentificatie wordt bemoeilijkt door het feit dat er geen meertalig spraakcorpus beschikbaar is (public-domain) waarmee verschillende aanpakken kunnen worden getest. Recentelijk is een dergelijk corpus beschikbaar gekomen: het OGI 'Multi-language Telephone Speech Corpus' (OGI_TS), dat spontane en fixed-vocabularyspraak bevat in elf talen. Dit corpus maakt het mogelijk om verschillende benaderingen objectief met elkaar te vergelijken. Gebruikmakend van dit corpus, organiseert NIST jaarlijks evaluaties van gesproken taalidentificatie systemen (acht Europese en Amerikaanse deelnemers). Zie voor resultaten van de NIST evaluatie in 1996: [2].

De grootste bottlenecks in gesproken taalidentificatie zijn ongeveer gelijk aan de bottlenecks bij spraakherkenning: veel problemen worden veroorzaakt door een mismatch tussen train- en testcondities. Daarnaast is (voor tekstonafhankelijke identificatie) veel data nodig in verschillende talen, en voor tekstafhankelijke identificatie is een multilinguale herkenner nodig.

Het gebruik van prosodie in taalidentificatie net als in spraakherkenning steeds meer aandacht, maar heeft tot nog toe niet veel succes opgeleverd.

Tot nog toe zijn de resultaten voor identificatie uit een groot aantal talen nog lang niet goed genoeg om in een echte dienst ingezet te worden, voor een kleine subset zijn ze acceptabel. De resultaten zijn sterk afhankelijk van de duur van het segment waarop de identificatie moet plaatsvinden.

Meer informatie: `http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/gac1/report.html#3`

9.8.2 Evaluatiecriteria

Identification error rates geven aan hoe vaak het systeem de verkeerde keuze heeft gemaakt.

9.8.3 Inventaris beschikbare software

Er is geen software specifiek voor het Nederlands bekend.

Er zijn een aantal multilinguale corpora beschikbaar, waarvan Nederlands een van de talen is: 'Eurom0' en 'Eurom1' (waarin Britsengels, Deens, Duits, Frans, Grieks, Italiaans, Nederlands, Noors, Portugees, Spaans, en Zweeds) en 'Polyglot' (zeven Europese talen). Zie ook hoofdstuk 10 over spraakcorpora.

9.9 Adaptatie

Het aantal fouten dat gemaakt wordt door een spraakherkenner, is sterk afhankelijk van de overeenkomst in train- en testcondities. Verschillen tussen beide condities kunnen verschillende oorzaken hebben: uitspraakvariatie, ruis, dialect, etc. Om het aantal fouten dat de spraakherkenner maakt te verminderen, wordt vaak adaptatie toegepast. Er zijn verschillende vormen van adaptatie: spreker adaptatie, lexicon/uitspraakadaptatie, taalmodel adaptatie, database adaptatie, ruisadaptatie, etc.

Adaptatie aan ruis en non-native sprekers is al aan de orde gekomen in eerdere secties (secties 9.5 en 9.6, resp.). In deze sectie zullen sprekeradaptatie en lexiconadaptatie worden besproken.

Sprekeradaptatie Door de enorme variabiliteit in sprekers is de performance van sprekeronafhankelijke systemen sterk afhankelijk van de betreffende spreker. Sprekeronafhankelijke systemen worden getraind op een grote hoeveelheid geannoteerde spraak van verschillende sprekers, en de resulterende modellen vormen daardoor min of meer een gemiddelde over alle sprekers. Voor iedere individuele spreker zijn deze modellen daardoor niet optimaal. Bovendien zijn er altijd sprekers die sterk afwijken van de sprekers die gebruikt zijn voor de training, en deze sprekers zijn daardoor niet goed gemodelleerd. Sprekerafhankelijke systemen zijn beter in staat specifieke kenmerken van één spreker te modelleren, en presteren daardoor beter, maar voor het trainen van dergelijke systemen is erg veel data nodig, wat niet altijd haalbaar is. Bovendien moet iedere gebruiker van een sprekerafhankelijk systeem een langdurige initialisatiefase doorlopen. Sprekeradaptatietechnieken proberen de sprekeronafhankelijke spraakherkenner te af te stemmen op de karakteristieken van de betreffende spreker, om op deze manier de performance van het systeem te verbeteren.

Er zijn verschillende methoden om sprekeradaptatie te doen. De keus voor een methode wordt bepaald door de volgende factoren: wat voor soort modellen worden gebruikt, welke applicatie wordt beoogd, hoeveel medewerking van de gebruiker is nodig, etc.

Er wordt onderscheid gemaakt in adaptatie op modelniveau en adaptatie op feature niveau. Bij *modeladaptatie* wordt een bestaande set HMM-en aangepast aan de huidige spreker. Bij *feature adaptatie* worden transformaties op de akoestische feature vectoren toegepast die de verschillen tussen sprekers verkleinen.

Een ander onderscheid dat gemaakt kan worden in adaptatietechnieken is het onderscheid in *supervised* en *unsupervised* adaptatie. Supervised wil zeggen dat bekend is wat er uitgesproken is, unsupervised betekent dat de woorden niet bekend zijn, in dit geval wordt meestal door de spraakherkenner een hypothese opgeleverd van wat er gezegd is. Een derde opdeling is die in *batch* en *incrementele*. Bij batch adaptatie wordt een hoeveelheid spraakmateriaal verzameld alvorens de adaptatie te doen (zoals bijvoorbeeld in de trainingsfase van een dicteersysteem). Bij incrementele adaptatie wordt de adaptatie direct gedaan zodra er data van de gebruiker binnenkomen, bijvoorbeeld na elke uiting. Incrementele adaptatie is doorgaans unsupervised.

Voor meer informatie, zie: [65, 73, 122, 123].

Lexiconadaptatie Bovenstaande technieken zijn er allemaal op gericht de akoestische modellen of de feature vectoren aan te passen. Hierbij wordt er vanuit gegaan dat ook variatie in uitspraak hierdoor verdisconteerd wordt. Echter, voor die gevallen waar de uitspraak zeer sterk afwijkt van de uitspraken in de train set, kan sprekermodellering door middel van het lexicon uitkomst bieden, dit noemen we lexiconadaptatie.

De term lexiconadaptatie wordt ook gebruikt voor het toevoegen van woorden aan het lexicon om het aantal OOV woorden te verkleinen (dit wordt soms gebruikt in bijv. dicteersystemen). In deze sectie wordt lexiconadaptatie alleen gebruikt in de zin van uitspraakadaptatie in het lexicon.

Lexiconadaptatie kan worden toegepast in verschillende situaties en kan worden gedaan op de volgende manieren:

- on line (adaptatie wordt gedaan tijdens de herkenning) of off line (adaptatie wordt gedaan in de ontwikkelfase),
- handmatig door een expert of automatisch,
- supervised of unsupervised (zie ook sprekeradaptatie).

Welke soort op welk moment het meest geschikt is, hangt af van de hoeveelheid en het soort materiaal dat beschikbaar is voor adaptatie. De hoeveelheid beschikbaar materiaal hangt af van de applicatie. Bijvoorbeeld bij dicteersystemen is er relatief veel materiaal beschikbaar per spreker, bij de meeste dialoogsystemen zijn per spreker slechts enkele uitingen beschikbaar.

Lexiconadaptatie kan worden gedaan op basis van het spraaksignaal, op basis van tekst of beide. In sommige applicaties behoort tekstinvoer (voor on line adaptatie) niet tot de mogelijkheden, in dat geval wordt adaptatie op het spraaksignaal gebaseerd.

Lexiconadaptatie kan worden gebruikt om het systeem aan te passen aan een specifieke spreker, een accent, een dialect of een taalachtergrond. Ook kan het worden gebruikt voor mensen met een handicap c.q. spraakstoornis. Of lexiconadaptatie al dan niet moet worden toegepast is afhankelijk van de frequentie, grootte en beschrijving van de verschillen tussen de uitspraken in het lexicon en de werkelijke uitspraak.

Voor meer informatie, zie [63, 121, 120].

Overzicht van papers over spraakherkenning en adaptatie:
`http://www.univ-st-etienne.fr/eurise/pdupont/bib/speech_basic.html`

9.9.1 State of the art internationaal

Sprekeradaptatie Op dit moment zijn er sprekeradaptatie technieken die maar weinig data vergen, en die vergelijkbare resultaten behalen als sprekerafhankelijke systemen [73]. De huidige adaptatiesystemen zijn echter off line. Ze kunnen dus niet gebruikt worden voor publiek toegankelijke systemen, zoals databanken die telefonisch kunnen geraadpleegd worden.

Experimenten op het Wall Street Journal corpus hebben aangetoond dat adaptatie de foutpercentages kan reduceren. Er wordt ook steeds meer aandacht besteed aan adaptatie in systemen die gebruik maken van neurale netwerken, in plaats van HMM-en.

Lexiconadaptatie Er zijn zeer veel applicaties waarbij met uitspraakadaptatie op het niveau van het lexicon waarschijnlijk een verbetering van de prestaties van een spraakherkenner bewerkstelligd zou kunnen worden [120]. Echter, voor zover ons bekend is, zijn er op

dit moment nog geen bestaande applicaties waarbij dit ook daadwerkelijk toegepast wordt. Verder is het onderzoek naar lexiconadaptatie tot nu toe ook zeer beperkt geweest, en heeft zich vnl. gericht op een bepaald soort adaptatie: Off line adaptatie, waarbij een lexicon met canonieke transcripties geoptimaliseerd wordt voor een specifiek spraakcorpus.

Entries in een lexicon zijn foontranscripties (zie ook sectie 9.3). Voor lexiconadaptatie moeten dergelijke transcripties eerst gegenereerd worden, en vervolgens moet bepaald worden welke transcripties toegevoegd worden aan het lexicon (selectie). Voor het genereren van transcripties is automatische transcriptie van spraak nodig (evt. in combinatie met grafeem-naar-foneem conversie, als er naast spraakinvoer ook tekstinvoer is). De kwaliteit van automatisch gegenereerde transcripties is echter op dit moment niet goed genoeg (zie sectie 9.4.3), en het tot nu toe uitgevoerde onderzoek heeft nog geen duidelijke richtlijnen opgeleverd m.b.t. selectie. Dit zou kunnen verklaren waarom lexiconadaptatie nog niet toegepast wordt in bestaande applicaties, en het maakt tevens duidelijk dat meer onderzoek op dit gebied nodig is.

9.9.2 Evaluatiecriteria

Omdat adaptatie wordt gebruikt voor het verbeteren van spraakherkenners wordt voor evaluatie van een adaptatiemethode gekeken naar het verschil in prestaties zonder en met adaptatie. Daarom worden voor adaptatie meestal dezelfde maten gebruikt als voor spraakherkenning (zie sectie 9.3.2): Word Error Rate en Word Accuracy.

9.9.3 Inventaris beschikbare software

Er is geen software specifiek voor het Nederlands bekend

- **HTK – HEadapt / HVite**

  Omschrijving: Deze module uit HTK (versie 2.2 of hoger) kan off line adaptation toepassen met gebruik van een kleine verzameling data. Zowel supervised adaptatie (HEadapt) als unsupervised adaptatie (HVite) worden ondersteund. Bij unsupervised adaptatie wordt de data herkend, en met het herkenresultaat worden direct de modellen aangepast.

  Beschikbaarheid: HTK is gratis te downloaden

  Meer informatie: `http://ciips.ee.uwa.edu.au/~roberto/research/speech/local/entropic/HTKBook/node184.html`

- **Departement elektrotechniek (E.S.A.T.) - Universiteit Leuven**

  “Algoritmes voor het on line aanpassen van spraakherkenningsystemen aan de sprekerkarakteristieken”.

  Omschrijving: De prestaties van de huidige sprekeronafhankelijke spraakherkenningsystemen zijn sterk afhankelijk van de spreker (geslacht, stemgeluid,...). Het aanpassen van de parameters in het herkenningssysteem aan de karakteristieken van de huidige spreker kan dit probleem grotendeels verhelpen. De meeste aandacht in ons onderzoek zal gaan naar adaptatietechnieken die geen supervisie behoeven en die on line werken, daar deze technieken algemeen toepasbaar zijn.

  Meer informatie:

  Patrick Wambacq (hoofdpromotor)
  `http://cwisdb.cc.kuleuven.ac.be/onderzoek/P/3E99/project3E990380.htm`

9.10 Betrouwbaarheidsmaten / uitingverificatie

De meest gangbare automatische spraakherkenners rekenen voor een binnenkomend spraaksignaal kansscores uit en leveren uiteindelijk de woordreeks op die het meest waarschijnlijk gesproken is. Deze kansscores worden uitgerekend met behulp van statistische modellen voor zinnen, woorden en klanken die van tevoren in een trainsituatie zijn gebouwd. Deze benadering heeft twee implicaties:

1. in het geval dat de trainsituatie een vertekening is van de operationele situatie, zullen de kansen inaccuraat worden geschat en neemt het risico dat herkenfouten optreden toe.

2. er wordt *altijd* een resultaat opgeleverd, zelfs als het systeem de invoer niet eens had kunnen herkennen.

Anders gezegd bevat de oplossing nogal eens herkenfouten en ontbreekt aan het systeem de mogelijkheid om te kunnen zeggen: "Sorry, ik geloof niet dat ik u helemaal goed verstaan heb". In plaats daarvan gokt het altijd maar een woordreeks die het best lijkt te passen.

In een groot aantal toepassingen, waaronder de befaamde 'sprekende computer' van Openbaar Vervoer Reisinformatie (OVR), is gebleken dat herkenfouten uitermate schadelijk kunnen zijn voor het algehele succes van het systeemgebruik. Gebruikers blijken het heel moeilijk te vinden om in de eerste plaats al te constateren dat een fout heeft plaatsgevonden en in de tweede plaats om daarvoor een correctie door te voeren. En als dat al lukt, is er vaak veel tijd mee verloren. Het vervelende is nu dat gemiddelde duur van een dialoog het absolute sleutelelement is voor bruikbaarheid, gebruikerstevredenheid en daarmee uiteindelijke acceptatie van het systeem. Niet altijd, maar wel vaak zou een spraakherkenner daarom liever *helemaal geen* dan een *onjuist* resultaat moeten opleveren.

Bij **uitingverificatie** wordt ernaar gestreefd om onjuist herkende (deel)uitingen vroegtijdig in de kraag te vatten en te verwerpen. De beslissing om een oplossing te verwerpen is veelal gebaseerd op de toetsing van een kansscore aan een vooraf gestelde drempelwaarde. De kansscores die een spraakherkenner gebruikt om de best passende oplossing te zoeken zijn hiervoor niet adequaat. Deze zijn namelijk geen echte wiskundige kansen, maar alleen maar geschikt om meerdere oplossingshypothesen onderling te kunnen vergelijken; de berekende waarde zegt op zich zelf heel weinig. Daarom hebben we voor uitingverificatie een maat nodig die onafhankelijk van het resultaat en het spraaksignaal betekenis draagt.

Betrouwbaarheidsmaten zijn daar een voorbeeld van. Dit zijn schattingen van de waarschijnlijkheid dat een oplossing correct is. Met succes zijn er overal ter wereld verschillende varianten geïmplementeerd. Sommigen zijn gebaseerd op de 'nabijheid' van de secundaire oplossing ("in welke mate had het net zo goed iets anders kunnen zijn?"). Anderen gaan terug naar het akoestische spraaksignaal en gaan foneem voor foneem na of er niet toch vreemde eenden in de bijt zitten ("in hoeverre is dit stukje spraak typisch of juist atypisch voor deze klank?").

De toepassing van betrouwbaarheidsmaten strekt zich verder uit dan alleen uitingverificatie. De maten vinden zo hun toepassing in 'unsupervised' sprekeradaptatie. Het systeem past zijn akoestische modellen vooral aan op die segmenten die met grote zekerheid zijn herkend. Ook worden ze toegepast in de dialoogmanager van OVR systeem. Als de spraakherkenner aangeeft dat hij onzeker is over een herkende stationsnaam, wordt een voorzichtige houding naar de gebruiker aangenomen. Zo kan correctie soepeler verlopen. Maar in geval van grote zekerheid, gaat het systeem als een speer door de dialoog.

9.10.1 State of the art internationaal

Tegenwoordig is bijna elke succesvolle spraakherkenner (in verschillende toepassingsgebieden, zoals automatic inquiry, dicteersystemen, voice control) zoals die van Lucent, Nuance, Philips, L& H, BBN, AT& T, etc., uitgerust met een component voor uitingverificatie en betrouwbaarheidsmaten. Hiernaar wordt nog steeds veel onderzoek verricht: bijv. France Telecom, RWTH Aachen & Philips, Lucent Technologies en IBM hebben grote bijdragen geleverd aan het onderzoek op dit gebied.

Zoals al eerder gezegd, geldt voor spraakherkenning in het algemeen dat de performance afneemt naarmate het aantal klassen en/of de variabiliteit van features toeneemt. De performance van betrouwbaarheidsmaten hangt hier sterk mee samen. De kwaliteit van een bepaald algoritme kan alleen worden geëvalueerd in de context van zijn concrete toepassingsdomein.

Specifiek voor het Nederlands geldt dat het belangrijk is dat Nederland aansluiting ziet te vinden bij het internationale niveau. Op dit moment wordt er in Nederland weinig aandacht geschonken aan onderzoek naar uitingverificatie en betrouwbaarheidsmaten. Het onderzoek naar spraakherkenning op zich hangt hier sterk mee samen. Op dit moment wordt in het onderzoek op dit gebied nog veel gewerkt met off-the-shelfspraakherkenners.

Meer informatie over dit onderwerp in [145].

9.10.2 Evaluatiecriteria

Bij het toepassen van betrouwbaarheidsmaten en uitingverificatie in combinatie met een drempelwaarde voor het al dan niet accepteren van uitingen/woorden, kunnen twee typen fouten optreden:

1. false rejections - een woord/uiting wordt als onbetrouwbaar aangemerkt en dus afgewezen, terwijl het goed herkend was.

2. false acceptances - een woord/uiting dat fout herkend was, werd ten onrechte als betrouwbaar bestempeld, en daarom geaccepteerd.

Het aantal false rejections en het aantal false acceptances hangen nauw met elkaar en met de drempelwaarde voor acceptatie samen. Om de relatie tussen deze drie dingen te tonen, wordt vaak gebruik gemaakt van een Receiver Operating Characteristic (ROC) curve. Deze beeldt voor verschillende drempelwaardes het aantal false rejections en het aantal false acceptances (zie Figuur 3 in sectie 9.7.2) [10].

9.10.3 Inventaris beschikbare software

Er zijn geen off-the-shelfproducten beschikbaar voor het Nederlands. De software die binnen verschillende onderzoeksprojecten is ontwikkeld, is vaak zeer taakspecifiek en daardoor niet breed inzetbaar.

9.11 Standaardisatie

Tot slot een korte noot over de meer en meer opkomende standaardisatie van software componenten. Voor spraaktechnologische modules zal steeds meer moeten worden voldaan aan standaard applicatie interfaces voor integratie in grotere applicaties.

- **Java Speech API**

  `http://java.sun.com/products/java-media/speech/`

- **Microsoft Speech API**

  `http://www.microsoft.com/speech/technical/SAPIOverview.asp`

- **VoiceXML** voor dialogen

  `http://www.w3.org/TR/voicexml/`

- **IBM Voicetimes** voor mobiele spraaktechnologie

  `http://www-4.ibm.com/software/speech/enterprise/ms_2.html`

- **W3C** (World Wide Web Consortium)

  De W3C Voice Browser Working groep definieert een ML (markup language) voor spraakherkenning grammatica's, gesproken dialogen, NLP semantiek, multimodale dialogen, spraaksynthese, maar ook een collectie van herbruikbare dialoogcomponenten.

  `http://www.w3.org/TR/2000/WD-voice-intro-20001204/#spif`

10 Spraaktechnologie data

ASH maakt veel gebruik van statistische methodes waar heel veel data voor nodig zijn. Het devies is dan ook nog steeds: “hoe meer, hoe beter”. Data zijn nodig voor het trainen van onder andere de akoestische modellen en de taalmodellen.

Voor verschillende klasse van applicaties zijn verschillende datasoorten nodig. Onderstaande beschreven corpora zijn verschillende soorten corpora, zoals op meerdere niveaus geannoteerde corpora, multilinguale corpora, telefoonspraak corpora, etc.

10.1 State of the art internationaal

Er zijn veel gesproken corpora voor de talen van de geïndustrialiseerde landen, zoals Amerikaansengels, Japans, Duits, Frans, Spaans. Veel wordt verspreid via LDC & ELRA (voor Duits ook BAS). Van deze corpora wordt een groot gedeelte gevalideerd door SPEX (Speech Processing EXpertise Centre, zie `http://www.spex.nl/`) conform internationale standaarden.

Veel corpora zijn, en worden nog opgenomen, binnen verschillende SpeechDat projecten, zowel ‘fixed, mobile & car’. Daarnaast wordt momenteel het Corpus Gesproken Nederlands opgenomen en geannoteerd.

Meer informatie:
ELRA: `http://www.icp.inpg.fr/ELRA/`
LDC: `http://www.ldc.upenn.edu/`
BAS: `http://www.phonetik.uni-muenchen.de/Bas/`
SpeechDat: `http://www.speechdat.org/`
CGN: `http://lands.let.kun.nl/cgn/ehome.htm`

10.2 Evaluatiecriteria

Evaluatie criteria die aan spraakcorpora gesteld kunnen worden, zijn ruwweg op de volgende wijze in te delen.

- Documentatie
  - over specificaties, bijvoorbeeld opnameconditie, hoeveelheid woorden
  - een eventueel bestaand validatie rapport
- Technische specificaties
  - bemonsteringsfrequentie
  - bitresolutie
  - opname kanaal (microfoon, telefoon, radio, televisie)
  - formaat (wave, a-law, raw)
- Kwantitatieve informatie
  - hoeveelheid spraak (aantal woorden, totale duur)
  - aantal sprekers

  - kwalitatieve informatie, zoals spreekstijl (voorgelezen, spontaan) en spreekmodus (monoloog, dialoog)

- Annotatie
  - transcriptie (orthografisch, fonemisch, fonetisch) van spraaksignaal, automatisch of handmatig
  - oplijning, segmentering (automatisch of handmatig) op welk niveau (zin, woord, foneem)
  - syntactische annotatie / analyse (POS, syntactisch)
  - prosodische annotatie
  - lexicon met fonetische transcripties

- Beschikbaarheid
  - vrij toegankelijk, te koop, niet beschikbaar voor derden

10.3 Evaluatie van spraakdatabases

In [134] wordt een duidelijk overzicht gegeven van de huidige stand van zaken met betrekking tot corpusevaluatie. Dit artikel is dan ook het uitgangspunt van de onderstaande tekst.

10.3.1 Introductie

In [134] wordt validatie gedefinieerd als de kwaliteitsevaluatie van een corpus met behulp van een controlelijst van relevante criteria. ELRA heeft voor validatie handleidingen gemaakt (ELRA Validation Manual for SLR, op `http://www.icp.inpg.fr/ELRA/validat.html`), en is bezig met het ontwikkelen van een procedure die, op de lange termijn, moet leiden tot een situatie dat er voor iedere corpusspecificatie en iedere kwaliteitscontrole documenten aanwezig zijn. In het algemeen zullen de specificaties geleverd worden door de 'bouwer' van het corpus, en zal de kwaliteitscontrole gebeuren door een onafhankelijke partij. Deze laatste taak heeft ELRA toevertrouwd aan SPEX.

10.3.2 Validatie en verbetering

Bij validatie kunnen twee dimensies onderscheiden worden (zie 5).

	Gedurende	Na productie
Validator		
Intern	1	2
Extern	3	4

Tabel 5: Vier types validatie strategieën.

De validatie kan gebeuren tijdens of na productie van het corpus; en de validatie kan uitgevoerd worden door de bouwer zelf (intern) of door een onafhankelijke partij (extern). Dit leidt tot vier types validatie strategieën (zie 5). In het ideale geval worden alle vier types

validatie uitgevoerd, wat bijv gedaan is in de SpeechDat-projecten (`http://www.phonetik.uni-muenchen.de/SpeechDat.html`).

Validatie en correctie moeten goed uit elkaar gehouden worden. Zij verschillen in een aantal opzichten:

- *Inhoud van de procedure:* Validatie is een kwaliteitsmeting, een diagnostische operatie die niets verandert aan het corpus zelf; bij correctie wordt het corpus juist wel veranderd, om onvolmaaktheden te corrigeren.
- *Chronologische volgorde:* In het algemeen zal de validatie eerst plaatsvinden, gevolgd door de correctie.
- *Verantwoordelijken:* De validatie gebeurt idealiter door een onafhankelijk instantie, terwijl de correctie veelal door de bouwer zelf gedaan zal worden.

10.3.3 Te valideren aspecten

De volgende acht aspecten van een corpus kunnen gevalideerd worden (zie ook [135]):

1. *Documentatie.* Zijn alle aspecten van het corpus duidelijk, compleet en correct beschreven?
2. *Formaat.* Zijn alle relevante bestanden aanwezig in de juiste mappen structuur en in het goede formaat?
3. *Design.* Is het corpus geschikt voor de voorziene applicaties?
4. *Spraakbestanden.* Akoestische metingen worden verricht (bijv. signaal-ruisverhouding, duur, oversturing, gemiddelde amplitude), en er vindt een auditieve inspectie plaats.
5. *Labelfiles.* Zijn ze aanwezig, in het juiste formaat, en kunnen ze automatisch geparseerd worden?
6. *Lexicon.* Is het compleet (dwz. bevat alle woorden), en correct (o.a. is er gebruik gemaakt van het juiste computer-fonetisch alfabet)?
7. *Distributie van sprekers en opnameomgevingen.*
8. *Orthografische transcripties.*

10.3.4 Controlelijst / specificaties

In de hierboven genoemde validatie wordt gecontroleerd of een opgenomen corpus voldoet aan de opgegeven specificaties. Die specificaties verschillen sterk per corpus. Bijv. syntactische en prosodische annotatie is wel onderdeel van de specificaties bij het CGN, maar niet bij de meeste andere corpora.

Daarnaast worden vaak naderhand dingen toegevoegd aan een corpus (bijv. analyse gegevens). Bijv. een syntactische en prosodische annotatie. Deze toegevoegde dingen kunnen ook voor anderen nuttig zijn, maar zullen meestal niet officieel deel uitmaken van het corpus.

Mede om genoemde redenen is het belangrijk dat er goede documentatie is. Hierin moeten allereerst de specificaties duidelijk beschreven zijn, maar daarnaast ook de validatie daarvan.

Als bovengenoemde validatie goed uitgevoerd is, zal er documentatie zijn met een controlelijst (per corpus). Daarnaast zal binnen het platform een 'meer algemene' controlelijst ontwikkeld worden, zie 10.2.
Meer informatie:
`http://www.speechdat.org/`
`http://www.icp.inpg.fr/ELRA/validat.html`
`http://www.spex.nl/`
Zie ook [134, 135]. ELRA Validation Manual for SLR (Spoken Language Resources), 36 p. SpeechDat `http://www.speechdat.org/SpeechDat.html`) Deliverable: SD1.3.3; Version: 1.9; Date: 21. Feb. 1997. Title: Validation criteria for Databases (zie de file sd133v19.doc).

10.4 Alfabetische opsomming van enkele Nederlandstalige corpora

Onderstaande is een lijst van bestaande Nederlandstalige corpora. Ongetwijfeld bestaan er talloos veel meer, maar onderstaande corpora worden hier vermeld omdat deze in zekere mate voldoen aan enkele criteria zoals beschreven in ELRA Validation Manual for SLR (`http://www.elda.fr/validat.html`). De belangrijkste criteria zijn:

- Spraaksignaal moet aanwezig zijn;
- Transcriptie, documentatie en eventueel lexicon moeten in een door computer leesbare vorm aanwezig zijn.

Naast deze criteria is herbruikbaarheid een ander belangrijk doel dat nagestreefd kan worden. Met herbruikbaarheid bedoelen we niet direct 'openbaar beschikbaar', maar bruikbaar ten behoeve van spraaktechnologische toepassingen en onderzoek.

Een ander niet te negeren aspect is het privacy aspect van de sprekers in de corpora en de copyrights. Voor veel bestaande Nederlandstalige corpora is dit niet goed geregeld of zijn de sprekers niet meer te achterhalen voor toestemming waardoor deze corpora niet herbruikbaar zijn.

- Bloemendal ASSP
- Casimir
- CGN
- COGEN
- DDAC
- Demsi
- Dutch Speech Styles Corpus
- ELIS-PBS
- ESFSLD
- EUROM-0 & EUROM-1
- Groningen corpus

- IDD
- IFA-DSLC
- MIVA
- Mobiel
- Multi Tongue Dutch I, II
- PBS corpus
- Philips Car
- PIG I, II
- Polyglot
- Polyphone
- Promocor
- Promotex
- Read Texts Corpus
- Sesp I, II, III
- SpeechDat Mobile (Dutch)
- TNO non-natives database
- Tootsie
- Van Der Wijst Corpus
- VIOS
- VNC
- VoiceDialling I, II, III
- Vowels In Concert
- (Noise-Rom-0)
- ...

Niet alle hierboven genoemde corpora zijn beschikbaar. Sommige corpora zijn nog in opbouw en enkele zijn 'in-housecorpora' die slechts bedoeld zijn voor 'in-housegebruik'. Compleetheid van de lijst is een streven maar mede door vele onbekende 'in-housecorpora' is dit niet mogelijk. Suggesties voor toevoegingen zijn welkom.

10.5 Inventarisatie van de verschillende Nederlandse spraakcorpora

Het Speech Processing Expertise Centre speelt ook voor Nederlandse spraakcorpora een grote rol. Naast bovengenoemde validatieactiviteiten van internationale corpora, is SPEX ook actief met het opnemen van spraakdatabases. Veel van onderstaande databases zijn ook door SPEX getranscribeerd.

- **Bloemendal ASSP (1987)**

  Omschrijving: Het corpus is opgenomen voor onder andere onderzoek naar effecten van spreeksnelheid op prosodie en uitspraak. Er bestaat echter geen orthografische transcriptie van het hele corpus, waarschijnlijk wel van delen ervan.

  Beschikbaarheid: nog onbekend

  Meer informatie: IFA, dr R. van Son; IPO, dr J. Terken

- **Casimir (1995)**

  Omschrijving: Database bevat cijferuitingen van 274 verschillende sprekers. De cijferuitingen zijn bestaan voornamelijk uit pincodes en veertiencijferige kaartnummers.

  Meer informatie: SPEX

- **CGN: Corpus Gesproken Nederlands (1999-2002)**

  Omschrijving: Het project Corpus Gesproken Nederlands 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 Nederland, en eenderde uit Vlaanderen. Het Corpus Gesproken Nederlands wordt gevormd door een selectie van een groot aantal fragmenten van (opnames van) gesproken tekst. In totaal gaat het hierbij om een duizendtal uren spraak. Al het materiaal wordt orthografisch getranscribeerd, terwijl er tevens een oplijning plaatsvindt waarbij de orthografische transcriptie gekoppeld wordt 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 (geverifieerde) brede fonetische transcriptie wordt vervaardigd, er een geverifieerde oplijning 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.

  Beschikbaarheid: De Nederlandse Taalunie zal te zijner tijd verantwoordelijk zijn voor de distributie van de corpora.

  Delen van het corpus worden al tijdens de looptijd van het project ongeveer om de zes maanden beschikbaar gesteld. Het volledige corpus zal medio 2003 beschikbaar zijn.

  Het corpus wordt beschikbaar gesteld 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.

  Meer informatie:

  `http://lands.let.kun.nl/cgn/`
  `http://lands.let.kun.nl/cgn/epublicat.htm`

- **COGEN (1996-1997)**

  Omschrijving: Het Corpus Gesproken Nederlands COGEN werd ontwikkeld in het kader van het Vlaams korte termijnprogramma Spraak- en Taaltechnologie voor het Nederlands (STTN). Het bevat vier subcorpora:

  1. WL-OFF (word list office), een corpus van gespelde woorden, commandowoorden, cijfers, fonetisch rijke woorden, gelezen door in totaal 174 sprekers, opgenomen in een kantooromgeving (i.e. een omgeving die niet speciaal voor opnames geprepareerd is en die dus achtergrondgeluiden bevat). Totaal 2.16 uur gespelde woorden, en 5.83 uur voorgelezen woorden.
  2. RS-OFF (read speech office, een corpus van voorgelezen tekstfragmenten (5 paragrafen), door 174 sprekers, in kantooromgeving. Totale duur: 7.02 uur.
  3. WL-TEL (word list telephone), een corpus van voorgelezen woordenlijsten, opgenomen via een telefoonverbinding, opgenomen voor 185 sprekers. Duur: 5.85 uur.
  4. SS-TEL (spontaneous speech telephone), een corpus van spontane uitingen, opgenomen via een telefoonverbinding, opgenomen voor 126 sprekers. Duur: 2 uur.

  Beschikbaarheid: Geen gegevens.

  Meer informatie:

  ELIS Speech Laboratory (Gent) and ESAT/PSI (Leuven)

  Zie ook [47].

- **DDAC (2000)**

  Omschrijving: DDAC is het Dutch Directory Assistance Corpus. Dit corpus is opgenomen omwille van de ontwikkeling en testen van een spraakgestuurde dienst. Het corpus bevat dialogen over telefoonnummerinformatie, de uitingen bestaan uit plaatsnamen, namen, bedrijfsnamen, straatnamen, etc. Het corpus bevat meer dan 65000 dialogen en is geheel orthografisch getranscribeerd door SPEX

  Beschikbaarheid: beschikbaar voor onderzoek

  Meer informatie: SPEX

  `http://lands.let.kun.nl/literature/bouwman.2001.2.pdf`

- **Demsi**

  Omschrijving:

  Beschikbaarheid:

  Meer informatie:

  `http://coral.lili.uni-bielefeld.de/EAGLES/`

- **Dutch Speech Styles Corpus (1994)**

  Omschrijving: Het Dutch Speech Styles Corpus is gecreëerd ten behoeve van stemkwaliteitsonderzoek van sprekers van het Standaard Nederlands. Het corpus bevat

drie verschillende spraakstijlen, namelijk spontane spraak (monologen), semi-spontane spraak (beschrijvingen van plaatjes) en voorgelezen spraak. Het corpus bevat spraak van 127 sprekers, waarvan 60 mannen en 67 vrouwen, verdeeld in drie leeftijdscategorieën. Alle spraak is getranscribeerd door één transcribent. De totale hoeveelheid spraak bedraagt meer dan 19 uur.

Beschikbaarheid: SPEX

Meer informatie: Zie ook [48].
`http://fonsg3.let.uva.nl/Proceedings/Proceedings18/Els_den_Os/ELSSPEX_Proc_18.html`
`http://www.icp.inpg.fr/Relator/dutch/dutchspstyle.html`
SPEX (`http://www.spex.nl`)

- **ELIS-PBS**

  Omschrijving: Dit is een Vlaams corpus van Phonetically Balanced Sentences (PBS) opgesteld aan de Universiteit van Gent. Het bevat 13 verschillende fonetisch gebalanceerde zinnen van 130 verschillende sprekers. De totale duur is ongeveer anderhalf uur. De uitingen zijn orthografisch en fonetisch getranscribeerd. Tevens zijn de uitingen fonetisch gesegmenteerd.

  Beschikbaarheid: niet beschikbaar

  Meer informatie:

  `http://elis.rug.ac.be`

- **ESFSLD (Migranten Databank) (1982-1987)**

  Omschrijving: Dit corpus, European Science Foundation Second Language Databank (ESFSLD), is een elektronisch archief van longitudinale studies naar de tweede taalverwerving van een aantal volwassen immigranten uit 6 landen. Voor elk van de 6 moedertalen (Punjabi, Italiaans, Turks, Arabisch, Spaans en Fins), werden twee groepen geselecteerd, die elk dezelfde tweede taal (Engels, Duits, Nederlands, Frans of Zweeds) moesten leren. De studie startte in 1982, en werd in 1987 voltooid.

  In totaal werden 40 allochtone werknemers geselecteerd, wier conversatie met native speakers van de doeltaal op de band werd vastgelegd en later getranscribeerd volgens een centraal vastgesteld protocol. Een grote variëteit aan activiteiten werd vastgelegd: socio-biografische conversatie (soort intake-gesprek), rollenspel, plaatjesbeschrijving, filmbeschrijving, routebeschrijving, zelfconfrontatie (commentaar op bekijken eigen handelen), etc.

  Alle migranten werden maandelijks geïnterviewd over een periode van 2,5 jaar. Daarnaast werd een controlegroep van in totaal 24 migranten aan het begin, in het midden en tegen het einde van de opnameperiode geïnterviewd ter vergelijking met de hoofdgroepen. Voor het Nederlands werden twee groepen participanten vastgesteld: één met moedertaal Turks, en één met moedertaal Arabisch. Dit deel van het onderzoek werd geleid door Guus Extra van de Katholieke Universiteit Brabant. De centrale coördinatie was in handen van het Max Planck Instituut voor Psycholinguïstiek in Nijmegen. Van 39 participanten is een orthografische transcriptie voorhanden en slechts van 8 van de controlegroep.

Beschikbaarheid: Max Planck Instituut voor Psycholinguïstiek, Nijmegen

Meer informatie:

`http://lands.let.kun.nl/cgn/publ/1999_01.pdf`

Feldweg, H. (1992). The European Science Foundation Second Language Databank. Ongepubliceerd document, MPI Nijmegen.

- **EUROM0 & EUROM1** (The multilingual European speech database).

  Omschrijving: Dit is een Europees initiatief om platformonafhankelijke, uniform gecodeerde en ontsloten gesproken taalcorpora (met slechts voorgelezen spraak) samen te stellen voor alle Europese talen. Het maakt gebruik van de in Europa erkende SAMPA transcriptiestandaard (ESPRIT SAM 2589). Het is vooral geschikt voor industriële toepassingen. Samenstelling (teksttypes): 100 voorgelezen getallen, 60-100 CVC-patronen, 10 woorden in isolatie, 50 zinnen en 40 alinea's van 5 zinnen. Sprekergegevens: 60 sprekers per taal. De vertegenwoordigde talen zijn Brits Engels, Deens, Duits, Frans, Grieks, Italiaans, Nederlands, Noors, Portugees, Spaans, en Zweeds. 30 mannelijke en 30 vrouwelijke sprekers per taal, alle tussen de 20 en 60 jaar.

  Beschikbaarheid: Distributie door ELRA. In de praktijk blijken er grote problemen vanwege de afstemming tussen alle Europese partners, het gedeelde auteursrecht, en copyright op het GERSONS-databasesysteem, dat berust bij het bedrijf ICP. Momenteel zijn alleen Italiaanse data beschikbaar bij ELRA.

  Meer informatie:

  `http://www.icp.inpg.fr/Relator/multiling/eurom1.html`

  `http://www.icp.inpg.fr/Relator/multiling/eurom1.html`

  `http://www.icp.grenet.fr/ELRA/cata/spee_det.html#eurom1`

- **GRONINGEN Corpus (1996)**

  Omschrijving: Dit is een corpus met Nederlandse voorgelezen spraak, verzameld door A.M. Sulter en H.K. Schutte. Samenstelling (teksttypes): voorgelezen tekst, getallen, eenlettergrepige woorden, 23 fonetisch rijke korte zinnen, twee stukken tekst met veel directe rede om 'emotionele spraak' op te wekken. Sprekergegevens: 238 sprekers. 94 sprekers lezen ook nog een uitgebreide woordenlijst voor. Gegevens over leeftijd, lengte, gewicht, rook- en drinkgedrag zijn opgenomen. Er zijn ook pathologische sprekers opgenomen. De stemkwaliteit is beschreven door de spreker zelf en een panel van luisteraars. De sprekers worden gekarakteriseerd als sprekers van het Standaardnederlands. De omvang van de vier CDROMS is meer dan 20 uur spraak. De data is verwerkt en getranscribeerd door SPEX en op CD-ROM gedistribueerd, met steun van ELSNET, en de pre-mastering is gedaan bij LIMSI_CNRS.

  Beschikbaarheid: Distributie door ELRA.

  Meer informatie:

  `http://www.icp.inpg.fr/ELRA/cata/tabspeech.html`
  `http://www.icp.inpg.fr/ELRA/cata/spee_det.html#gron`
  `http://www.elsnet.org/groningen.html`
  SPEX

- **IDD (2000)**

  Omschrijving: IDD bevat cijferreeksen opgenomen in de periode november 1997 tot september 1999 in twee internet experimenten:

  1. *The CAVE speaker recognition demo*

     Deze website demonstreert hoe veertiencijferige kaartnummers gebruikt kunnen worden als een 'sleutel' ter beveiliging van toegang tot websites. Om zich in te schrijven moest de spreker een naam opgeven, het geslacht, leeftijd en moest 8 maal zijn of haar kaartnummer inspreken. Om de sprekerherkenning van het systeem te testen hoefde de spreker maar 1 maal zijn of haar kaartnummer uit te spreken.

  2. *The speech recognition demo*

     Deze website demonstreert herkenning van verbonden cijferreeksen. Sprekers spraken hun banknummer (7-10 cijfers) in en het herkenresultaat werd gepresenteerd in de vorm van een H-best lijst.

     Opname condities: alle cijferreeksen zijn opgenomen met behulp van de opname mogelijkheden van de individuele spreker thuis. Deze files werden opgestuurd via het web. Door de niet te controleren opname conditie ontstond er een wijde variatie aan geluidskwaliteit. Alle geluiden zijn uiteindelijk opgeslagen als 8 kHZ, 16-bit lineair. Het corpus is getranscribeerd door SPEX.

  Beschikbaarheid: SPEX

  Meer informatie: Zie [84].

- **IFA-DSLC (2001)**

  Omschrijving: Dit corpus is ontworpen, opgenomen en getranscribeerd onder verantwoordelijkheid van R.J.J.H van Son aan het Institute of Phonetic Sciences, Universiteit van Amsterdam.

  Het corpus bevat spraak van zowel mannen als vrouwen die in verschillende spraakmodi zijn opgenomen. Het corpus is breed fonetisch getranscribeerd en op foneemniveau opgelijnd door SPEX waar ook de manuele verificatie plaatsvond.

  Beschikbaarheid: vrij beschikbaar via Nederlandse Taalunie en R.J.J.H van Son

  Meer informatie:

  `http://www.fon.hum.uva.nl/rob/`
  `http://www.spex.nl`

- **MIVA (1996)**

  Omschrijving: Multilingual Interactive Voice Activated (MIVA) services bevat spraak van 387 mannelijke sprekers en 316 vrouwelijke sprekers. De sprekers is gevraagd een woordenlijst (applicatiewoorden) in te spreken. De sprekers zijn verdeeld over leeftijd, geslacht en soort telefoon (vast, mobiel, telefooncel). De uitingen zijn orthografisch getranscribeerd. MIVA was een Europees project, waarin verschillende landen een soortgelijk corpus opnamen ten behoeve van een multilinguale informatiedienst. Voor het eind van het project zijn de meeste deelnemers eruit gestapt

  Meer informatie: KPN Research, Leidschendam; SPEX

- **Mobiel (1995)**

  Omschrijving: Corpus is orthografisch getranscribeerd door SPEX. De spraakdata is opgenomen ter ontwikkeling van een spraakgestuurde dienst over de mobiele telefoon. De sprekers zijn evenwichtig verdeeld over geslacht, leeftijd en regio. De spraakdata bestaat uit applicatiewoorden en cijferreeksen.

  Meer informatie: KPN Research, Leidschendam; SPEX

- **Multi Tongue Dutch I & II (1997)**

  Omschrijving: Corpus is ontstaan door een dataverzameling ten behoeve van onderzoek naar taalverwerving, in bijzonderheid uitspraakevaluatie en testen van non-native sprekers van het Nederlands. De sprekers is gevraagd de fonetisch rijke zinnen uit het Polyphone corpus via de telefoon in te spreken. Alle uitingen zijn uitgebreid orthografisch getranscribeerd door SPEX.

  Beschikbaarheid: KUN

  Meer informatie: Catia Cucchiarini, KUN; SPEX

- **Philips Car (2000)**

  Omschrijving: Het corpus is opgenomen door SPEX in opdracht van Philips. Zo'n tweehonderd sprekers is gevraagd ongeveer 300 items in te spreken terwijl zij in een auto zaten. De opnamesituaties varieerden van stilstaande auto met motor aan tot rijdend op de snelweg. Er is gebruik gemaakt van verschillende type auto's. De opnames zijn bedoeld ter ontwikkeling van een spraakgestuurd product. Alle uitingen zijn orthografisch getranscribeerd door SPEX.

  Beschikbaarheid: Niet beschikbaar

  Meer informatie: SPEX

- **PIG I & II : Picasso Investment Game (1999-2000)**

  Omschrijving: De opnames zijn gemaakt in de testperiode van een sprekerverificatie systeem in de vorm van een beleggingsspel, waarin via een spraakgestuurde dienst aandelen verhandeld konden worden. De dienst, ontwikkeld bij KPN Research, was beveiligd met een persoonlijk nummer voor de deelnemers en een verificatie van de uitspraak van een depotnummer. De database bestaat dan ook uit cijferreeks uitingen. Ongeveer 30 personen hebben hieraan meegewerkt. Alle uitingen zijn orthografisch getranscribeerd.

  Beschikbaarheid: niet beschikbaar.

  Meer informatie: KPN Research, Leidschendam; SPEX

- **Polyglot (1989 – 1992)**

  Omschrijving: Dit multilinguale corpus kwam voort uit het ESPRIT (European Strategic Programme for Research and development in Information Technology) project, genaamd POLYGLOT. In het project is onderzoek gedaan naar automatische spraakherkenning alsmede spraaksynthese van zeven talen van de EG.

  Meer informatie:

  Mr. Francesco Lovecchio, SYNTAX SISTEMI SOFTWARE SpA
  `http://www.newcastle.research.ec.org/esp-syn/text/2104.html`

http://lands.let.kun.nl/literature/strik.1992.2.html

- **Polyphone (1995)**

  Omschrijving: POLYPHONE is een internationaal corpus van telefoonspraak, gecoördineerd door het Linguistic Data Consortium in de VS. Het Nederlandse deel werd verzameld in samenwerking tussen PTT-Telecom en het Expertisecentrum SPEX. Het is te verkrijgen bij het Europese consortium ELRA. Er zijn inmiddels ook Amerikaansengelse, Amerikaansspaanse, Franse, Duitse, Japanse, Mandarijns-chinese, Zwitsersfranse en Deense versies beschikbaar. Het Amerikaanse deel van het POLYPHONE-project staat bekend onder de naam MACROPHONE.

  Samenstelling (teksttypes): geëliciteerde spontane spraak (beantwoording 14 voorgedrukte vragen, zoals "Is Nederlands uw moedertaal?", "Heeft U ooit in een ander land dan Nederland gewoond?","In welke plaatsen bent u opgegroeid?", "Bent u een vrouw of een man?", en 4 niet-voorgedrukte vragen ("Spel uw naam alstublieft", "Hoe laat is het nu?"), 32 stukken voorgelezen tekst (getallen, woorden, gespelde woorden, datum, bedrag, tijdsaanduiding, hoeveelheid, zinnen met een applicatiewoord, fonetisch rijke zinnen). In totaal 50 items per spreker.

  Sprekergegevens: 5050 sprekers, zo mogelijk gelijkelijk verdeeld over geslacht, leeftijd (16-20, 21-40, 41-60, 61-), regio en sociaal-economische klasse. De sociaal-economische klasse werd gedefinieerd in termen van opleiding: alleen lagere school, middelbare school en hbo/universiteit.

  Beschikbaarheid: Distributie door ELRA.

  Meer informatie:

  http://www.icp.inpg.fr/ELRA/cata/tabspeech.html
  http://www.icp.inpg.fr/ELRA/cata/spee_det.html#dutpoly
  Zie ook [49].

- **Promocor (1995 – 1998)**

  Omschrijving: In samenwerking met Lernout & Hauspie Speech Products heeft ELIS (Gent) gewerkt aan het project PROMOCOR. Hierin is onderzoek gedaan naar duurmodellering en voorspelling van prosodische markers (woord prominentie en prosodische grenzen). Tijdens dit project zijn EUROM0 en COGEN prosodisch geannoteerd door het ontwikkelde systeem.

  Beschikbaarheid: niet beschikbaar

  Meer informatie:

  http://chardonnay.elis.rug.ac.be/en/research/tts.html
  RUG-ELIS The Speech Processing Lab, Gent

- **Promotex (1999 – 2001)**

  Omschrijving: Het PROMOTEX project (samenwerking Lernout & Hauspie en ELIS) had tot doel het modelleren van intonatie. Het onderzoek heeft zich gericht op intonatiemodellering van geïsoleerde zinnen van zes verschillende talen (Engels, Nederlands, Frans, Duits, Italiaans en Spaans). Voor deze modellering is gebruik gemaakt van datagedreven technieken, zoals (recurrente) neurale netwerken.

Voor elk van de bovengenoemde talen zijn ongeveer 1200 zinnen met verschillende spreekstijlen, syntactische patronen en van verschillende lengte opgenomen. Elke taal in ingesproken door een professionele moedertaalsprekende vrouwelijke spreker. Het spraakmateriaal is prosodisch geannoteerd.

Beschikbaarheid: niet beschikbaar.

Meer informatie:

`http://chardonnay.elis.rug.ac.be/en/research/tts.html`
RUG-ELIS The Speech Processing Lab, Gent

- **Read Texts Corpus (1994)**

  Omschrijving: Het corpus bevat voorgelezen spraak van 1 spreker. Deze spreker heeft 45 korte Nederlandse teksten voorgelezen in een echovrije ruimte. Een doorsnede van deze teksten zijn op twee spreeksnelheden ingesproken, normaal en snel. Ongeveer 6 minuten (20 teksten) is fonemisch getranscribeerd en gesegmenteerd. De overige teksten zijn morfologisch geannoteerd, gesyllabificeerd en voorzien van woordklasse.

  Beschikbaarheid: Beschikbaar via ELRA

  Meer informatie:

  `http://www.icp.grenet.fr/Relator/dutch/dutchreadtext.html`
  `http://www.icp.grenet.fr/ELRA/home.html`

- **SESP I, II, III (1996 – 1999)**

  Omschrijving: Deze corpora zijn opgenomen tijdens het CAVE project. SESP I en SESP II verschillen van SESP III in opnameomstandigheden. SESP I en II zijn database collecties terwijl SESP III is opgenomen tijden een test van een scopecard-applicatie waarbij sprekerverificatie is gebruikt. Alle SESP's bestaan uit cijferreeks uitingen van los uitgesproken cijfers, namelijk scopecard-nummers, telefoonnummers en pincodes, van 'nul' tot en met 'negen'.

  Meer informatie: KPN Research, Hans Jongebloed; SPEX

- **SpeechDat II Mobile**

  Omschrijving: Vijf corpora in SpeechDat zijn opgenomen met een mobiel netwerk. Alle SpeechDat corpora bevatten applicatiewoorden, cijferreeksen, getallen, geldbedragen, datum en tijdsexpressies, gespelde woorden, 'directory-assistancewoorden', ja / nee uitingen en fonetisch rijke zinnen. Het Nederlandse gedeelte is opgenomen door SPEX in opdracht van Philips en bevat spraakmateriaal van 250 verschillende sprekers die allen vier opnamesessies hebben doorlopen in vier verschillende omgevingen. De SpeechDat corpora zijn gevalideerd door SPEX.

  Beschikbaarheid: ELRA

  Meer informatie:

  `http://www.speechdat.org/`
  SPEX

- **SpeechDat II Vlaams**

  Omschrijving: Vlaams, 1000 sprekers, partner: Lernout & Hauspie

Beschikbaarheid: niet beschikbaar

Meer informatie:

`http://www.speechdat.org/speechdt/db_info.html`

- **TNO non-natives database (1996)**

  Omschrijving: In 1996 hebben 82 Nederlandstalige sprekers, 58 mannen en 24 vrouwen, meegedaan aan opnames ten behoeve van een multifunctioneel continue spraak corpus. De meeste sprekers waren medewerkers van TNO TM. Elke spreker is gevaagd 10 zinnen in het Nederlands, Engels, Frans en Duits voor te lezen. De zinnen kwamen respectievelijk uit de volgende kranten: NRC/Handelsblad, Wall Street Journal, Le Monde en Frankfurter Rundschau. Van de laatste drie kranten zijn de zinnen overgenomen uit het SQALE project (1993-1995).

  Alleen de Nederlandse zinnen zijn orthografisch getranscribeerd. Voor de andere talen is de prompttekst beschikbaar.

  Beschikbaarheid: TNO heeft het corpus beschikbaar gesteld aan deelnemers van de ESCA/NATO workshop MIST (Multi-lingual Interoperability in Speech Technology) 1998. Nu is het niet vrij beschikbaar meer.

  Meer informatie: `http://lands.let.kun.nl/mist/database.html`
  `http://www.nist.gov/speech/publications/darpa99/html/rel240/rel240.htm#15`
  `http://www.limsi.fr/tlp/sqale.html`

- **Tootsie (1995)**

  Omschrijving: Corpus bevat applicatie woorden, is opgenomen ter ontwikkeling van een spraakgestuurde dienst. Alle uitingen in het corpus zijn orthografisch getranscribeerd.

  Meer informatie:

- **Van Der Wijst Corpus (1991-1992)**

  Omschrijving: Corpus opgenomen tijdens een onderzoek naar beleefdheid in onderhandelingen, waarbij de 56 verschillende, mannelijke sprekers via een fictief telefoongesprek een onderhandeling moesten voeren. De opnames zijn niet opgenomen via de telefoon. Een deel van de uitingen wordt gebruikt in het CGN en orthografisch getranscribeerd.

  Beschikbaarheid: niet algemeen beschikbaar, maar informatie bij Per van der Wijst, Université de Liège.

  Meer informatie: Zie [137].

- **VIOS (1999)**

  Omschrijving: Dit corpus is ook wel bekend onder de naam OVIS. Er zijn verschillende versies in omloop van verschillende grootte ten behoeve van verschillende soorten onderzoek en bewerking. Het betreft hier opnames van een openbaarvervoerreisinformatiesysteem. De opnames zijn gemaakt door KPN t.b.v. het uittesten van dit mens-machine-dialoogsysteem. De uitingen zijn te karakteriseren als quasi spontaan en

bevatten allerhande vragen over reistijden en bestemmingen en antwoorden op ja / nee vragen. De uitingen zijn orthografisch getranscribeerd door SPEX.

Beschikbaarheid: Momenteel wordt er over onderhandeld door NWO.

Meer informatie:

- **VNC (vanaf 1994)**

  Omschrijving: De vakgroep Algemene Taalwetenschap en Dialectologie van de Katholieke Universiteit Nijmegen door het Vlaams Nederlands Comitë (VNC) gesubsidieerde project, getiteld "De uitspraak van het Standaardnederlands. Variatie en varianten in Nederland en Vlaanderen". Het onderzoek beoogt een brede registratie en gedetailleerde inventarisatie van de hedendaagse variatiepatronen in de standaarduitspraak. Naast deze descriptieve doelstelling is er een normatief evaluatieve doelstelling die zich richt op de oordelen van standaardtaalsprekers over varianten in het hedendaagse Standaardnederlands. Het onderzoek wordt parallel uitgevoerd in Nederland en Vlaanderen.

  De verzameling van een corpus spraak (bij leerkrachten Nederlands en bij radioomroepen), beschrijving van de uitspraak (met behulp van auditieve en akoestische analyses).

  Een deel van het VNC corpus is opgenomen in het CGN en is volgens CGN protocollen orthografisch getranscribeerd en geannoteerd.

  Meer informatie:

  dr. H. Van de Velde, KUN, Vakgroep Algemene Taalwetenschap en Dialectologie

- **Voice Dialling I, II, III (1996, 1997, 1997)**

  Omschrijving: Opnames zijn gemaakt gedurende drie veldtesten van een voice-diallingsysteem van KPN Research. De opnames zijn opgenomen via een mobiel netwerk en bevatten applicatie woorden, ja/nee antwoorden en cijferreeksen. Alle uitingen zijn orthografisch getranscribeerd

  Beschikbaarheid: niet vrij beschikbaar

  Meer informatie: KPN Research, Leidschendam; SPEX

- **Vowels in Concert**

  Omschrijving: VOWELS IN CONCERT is een corpus van gekalibreerde opnames van klinkers en frases die zijn gezongen door 14 professionele Nederlandse zangers, zeven mannen en zeven vrouwen.

  In totaal zijn 4099 klinkers opgenomen in verschillende contexten en in verschillende modes. De 168 frases zijn gezongen met variabele emoties, neutraal, angstig, blij. Alle opnames zijn gekalibreerd op geluidsdruk.

  Beschikbaarheid: SPEX

  Meer informatie:

  Zie [12].
  SPEX (`http://www.spex.nl`).

Referenties

[1] *Proceedings van COST250 Workshop "Application of Speaker Recognition Techniques in Telephony"*, Spain, 1996.

[2] *Proceedings van Workshop "Multi-lingual Interoperability in Speech Technology" (MIST)*, Nederland, 1999.

[3] C. Cucchiarini A. Neri and H. Strik. Effective feedback on l2 pronunciation in asr-based call. In *Proceedings of the workshop on Computer Assisted Language Learning, Artificial Intelligence in Education Conference, AIED, San Antonio, Texas USA*, pages 40–48, 2001.

[4] S. Abney. Parsing by chunks. In Robert Berwick, Steven Abney, and Carol Tenny, editors, *Principle-based parsing*. Kluwer Academic Publishers, Dordrecht, 1991.

[5] J. Allen, M.S. Hunnicutt, and D.H. Klatt. *From Text to Speech: the MITalk System*. Cambridge, England: Cambridge University Press, 1987.

[6] E.L. Antworth. *PC-KIMMO: a two-level processor for morphological analysis*. Number 16 in Occasional Publications in Academic Computing. Summer Institute of Linguistics, Dallas,TX, 1990.

[7] B. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *JASA*, 55:1304–1312, 1974.

[8] J. T. Berghmans. Wotan: een automatische grammaticale tagger voor het Nederlands. Master's thesis, Katholieke Universiteit Nijmegen, 1994.

[9] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proceedings of ICASSP79*, 1979.

[10] F. Bimbot. Assessment of speaker verification systems. In *EAGLES handbook of standards and resources for spoken language systems*. 1997.

[11] F. Bimbot. Speaker verification in the telephone network. research activities in the cave project. In *Proceedings of Eurospeech97*, 1997.

[12] G. Bloothooft. *Spectrum and timbre of sung vowels*. PhD thesis, University of Amsterdam, 1985.

[13] G. Bloothooft. Automatische spraakherkenning. *Informatie*, (31), 1989.

[14] G. Bocchieri and J. Wilpon. Discriminative analysis for feature reduction in automatic speech recognition. In *Proceedings of ICASSP92*, 1992.

[15] S.F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, 1979.

[16] P. Bonaventura. Speech recognition methods for non-native pronunciation variations. In *Proceedings of Eurospeech97*, 1997.

[17] L. Ten Bosch. Algorithmic classification of pitch movements. In *Proceedings of the ESCA workshop on Prosody*, 1993.

[18] L. Ten Bosch. On the automatic classification of pitch movements. *Journal of the Acoustic Society of America 94*, 1994.

[19] L. Ten Bosch. The potential role of prosody in automatic speech recognition. In *Proceedings of the Twente Workshop on Language Technology 8*, 1995.

[20] G. Bouma. A finite-state and data-oriented method for grapheme to phoneme conversion. In *Proceedings of the first conference of the North-American Chapter of the Association for Computational Linguistics*, pages 303–310, Somerset, NJ, 2000.

[21] G. Bouma. Extracting dependency frames from existing lexical resources. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations. Somerset, NJ. Association for Computational Linguistics.*, 2001.

[22] G. Bouma, G. Van Noord, and R. Malouf. Alpino: Wide-coverage computational analysis of dutch. In *Proceedings of CLIN 2000*, 2001.

[23] G. Bouma and I. Schuurman. De positie van het Nederlands in taal- en spraaktechnologie. Technical report, Nederlandse Taalunie, Den Haag, 1998.

[24] H. Bourlard. Links between markov models and multilayer perceptrons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (12), 1990.

[25] H. Bourlard. *Continuous speech recognition: A hybrid approach.* Kluwer academic publishers, 1993.

[26] T. Brants. Tnt - a statistical part-of-speech tagger. Technical report, Saarland University, Computational Linguistics, 1996. Saarbrücken, Germany.

[27] E. Brill. A simple rule-based part of speech tagger. In *Proceedings of ANLP'92*, 1992.

[28] P.F. Brown, S.A Della Pietra, V.J. Della Pietra, and R.L. Mercer. The mathematics of statisitical machine translation. *Computational Linguistics*, 19(2):263–313, 1993.

[29] S. Buchholz and A. van den Bosch. Integrating seed names and ngrams for a named entity list and classifier. In *Proceedings of LREC 2000 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May - 2 June 2000.

[30] H. Strik C. Cucchiarini and L. Boves. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. In *J. of the Acoustical Society of America*, volume 107 (2), pages 989–999. 2000.

[31] R. Carlson. Speech technology in language learning. In *Proceedings of STiLL, Marhölmen, Zweden*, 1998.

[32] P. Clarckson. Statistical language modelling using the cmu-cambridge toolkit.

[33] R.A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, V. Zue, G. B. Varile, and A. Zampolli, editors. *Survey of the state of the art in human language technology.* Cambridge University Press, 1996.

[34] P.A. Coppen. Morane, een systeem voor morfologische analyse. In *Verslagen computerlinguistiek*, volume 3, pages 1–40. KU Nijmegen, 1983.

[35] N. Cremelie. *Heuristische zoekstrategie en automatische aanmaak van lexica voor continue spraakherkenning.* PhD thesis, University of Gent, 2000.

[36] C. Cremers. On parsing coordination categorially. Master's thesis, HIL Leiden, 1993.

[37] C. Cucchiarini. *Phonetic transcription: a methodological and empirical study.* PhD thesis, University of Nijmegen, 1993.

[38] C. Cucchiarini. Assessing transcription agreement: Methodological aspects. *Clinical Linguistics and Phonetics*, 10:131–155, 1996.

[39] S. Cucerzan and D. Yarowsky. Language independent named entity recognition combining morphological and contextual evidence. In *Proceedings, 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, pages 90–99, 1999.

[40] W. Daelemans. Grafon: A grapheme-to-phoneme conversion system for Dutch. In *Proceedings twelfth international conference on cmputational linguistics (COLING-88)*, Budapest, 1988.

[41] W. Daelemans and A. van den Bosch. Language-independent data-oriented grapheme-to-phoneme conversion. In J. Van Santen, R. Sproat, J. Olive, and J. Hirschberg, editors, *Progress in speech synthesis*, pages 77–90. Springer Verlag, 1996.

[42] W. Daelemans, J. Zavrel, P. Berck, and S. Gillis. MBT: A memory-based part of speech tagger-generator. In E. Ejerhed and I. Dagan, editors, *Proceedings of the Fourth Workshop on Very Large Corpora*, pages 14–27, 1996. Copenhagen, Denmark.

[43] R. I. Damper, Y. Marchand, M. J. Adamson, and K. Gustafson. Evaluating the pronunciation component of text-to-speech systems for english: A performance comparison of different approaches. *Computer Speech and Language*, 13(2):155–176, 1999.

[44] E.D. de Jong. *Spreektaal: Woordfrequenties in Gesproken Nederlands.* Utrecht: Bohn, Scheltema and Holkema, 1979.

[45] F. de Jong, J.L. Gauvain, D. Hiemstra, and K. Netter. Language-based multimedia information retrieval. In *Proceedings RIAO 2000, Paris*, pages 713–725, 2000.

[46] J. de Veth. *On speech sound model accuracy.* PhD thesis, University of Nijmegen, 2001.

[47] K. Demuynck, F. Schoeters, K. Verschaeren, and D. van Compernolle. Cogen: a large database of spoken Dutch. Technical report, ESAT/PSI, Universiteit Leuven, 1997.

[48] E. A. den Os. Transliteration of the Dutch speech styles corpus. In *Proceedings of Institute of Phonetic Sciences, University of Amsterdam*, 1994.

[49] E. A. den Os, T. I. Boogaart, L.Boves, and E. Klabbers. The Dutch Polyphone Corpus. In *Proceedings of Eurospeech95*, 1995.

[50] E. Dermatas and G. Kokkinakis. A language-independent probabilistic model for automatic conversion between graphemic and phonemic transcription of words. In *6th European Conference on Speech Communication and Technology, September 5-9, 1999 Budapest, Hungary*, 1999.

[51] T. Dutoit. High-quality text-to-speech synthesis: an overview. *Journal of Electrical and Electronics Engineering*, 1997.

[52] W. Daelemans en H. Strik. Actieplan voor het nederlands in de taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen (1e versie). september 2001.

[53] M. Eskenazi. Using automatic speech processing for foreign language pronunciation tutoring: some issues and a prototype. *Language Learning and Technology*, 1999.

[54] W. Haesereyn et al. *Algemene Nederlandse Spraakkunst*. Martinus Nijhoff Uitgevers Groningen / Wolters Plantyn Deurne, 1997. Tweede, geheel herziene druk.

[55] F. Van Eynde. Part of speech tagging en lemmatisering. Technical report, CCL, K.U.Leuven, 2001.

[56] F. Van Eynde, J. Zavrel, and W. Daelemans. Part of speech tagging and lemmatisation for the Spoken Dutch Corpus. In *Proceedings of the second LREC*, pages 1427–1433, Athens, Greece, 2000.

[57] W.N. Francis. A tagged corpus - problems and prospects. In S. Greenbaum et al, editor, *Studies in English linguistics for Randolph Quirck*, pages 192–209. Longman, 1979.

[58] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29:254–272, 1981.

[59] S. Furui. An overview of speaker recognition technology. In *Automatic speech and speaker recognition*. 1996.

[60] W. A. Gale and K.W. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–103, 1993.

[61] J.L. Gauvain and C.-H. Lee. Maximum-a-posteriori estimation for multi-variate gaussian observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994.

[62] J.L. Gauvain and C.-H. Lee. Maximum likelihood linear regression for speaker adaptation of continuous density hmms. *Computer, Speech and Language*, 9:171–186, 1995.

[63] A. Geutner. Selection criterion for hypothesis driven lexical adaptation. In *Proceedings of ICASSP '99*, 1999.

[64] D. Gibbon, I. Mertins, and R. Moore. Consumer off-the-shelf (cots) product and service evaluation. In *Handbook of Multimodal and Spoken Dialogue Systems*. 2000.

[65] D. Giuliani. Speaker adaptation. In *Spoken Dialogues with computers*. 1999.

[66] G. Grefenstette. Tokenization. In H. van Halteren, editor, *Syntactic Wordclass Tagging*. Kluwer Academic Publishers, Dordrecht, 1999.

[67] R. Heemels. Morfo-analyzer - een experimenteel morfologisch systeem. Master's thesis, Computerlinguïstiek KU Nijmegen, 1985.

[68] W.J. Hess. Speech synthesis – a solved problem? In *Proceedings of EUSIPCO-92*, 1992.

[69] D. Hiemstra. Multilingual domain modelling in twenty-one: automatic creation of a bi-directional translation lexicon from a parallel corpus. In *Proceedings of the eighth CLIN meeting*, pages 41–58, 1997.

[70] D. Hiemstra, W. Kraaij, R. Pohlmann, and T. Westerveld. Translation resources, merging strategies and relevance feedback for cross-language information retrieval. In *Lecture Notes in Computer Science 2069: Cross-langauge Information Retrieval and Evaluation*, pages 102–115. 2001.

[71] V. Hoste, W. Daelemans, and S. Gillis. A rule induction approach to modeling regional pronunciation variation. In *Proceedings of COLING 2000, Saarbrcken, Germany*, pages 327–333. San Francisco: Morgan Kaufman Publishers, 2000.

[72] E. Hovy. Multilingual information management: Current levels and future abilities. Technical report, US National Science Foundation, 1999.

[73] X. Huang. On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. In *IEEE Transactions on Speech and Audio Processing*, 1993.

[74] Jaboulet. The cave-wp4 generic speaker verification system. In *Proceedings of RLA2C*, 1998.

[75] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, 2000.

[76] F. Karlsson, A. Voutilainen, and J. Heikkil, editors. *Constraint Grammar; A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin, New York, 1995.

[77] L. Karttunen. Constructing lexical transducers. In *The proceedings of the 15th international conference on computational linguistics. Coling 94*, pages 406–411, Kyoto, Japan, August 5-9 1994.

[78] L. Karttunen. Directed replacement. In *The Proceedings of the 34rd Annual Meeting of the Association for Computational Linguistics. ACL-96*, Santa Cruz, California, 1996.

[79] J. M. Kessens and H. Strik. Lower wers do not guarantee better transcritions. In *Proceedings of Eurospeech 2001, Aalborg, Denmark*, volume 3, pages 1721–1724, 2001.

[80] E. Klabbers, K. Stober, R. Veldhuis, P. Wagner, and S. Breuer. Speech synthesis development made easy: The bonn open systhesis system. In *Proceedings of Eurospeech2001*, 2001.

[81] D. Klatt. A digital filter bank for spectral matching. In *Proceedings of ICASSP76*, 1976.

[82] D.H. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustic Society of America*, (67), 1980.

[83] E. M. Konst. Automatic grapheme-to-phoneme conversion of Dutch names. In *Proceedings of ICSLP94*, pages 735–738, 1994.

[84] J. Koolwaaij. *Automatic speaker verification in telephony: a probabilistic approach*. PhD thesis, University of Nijmegen, 2001.

[85] K. Koskenniemi. A general computational model for word-form recognition and production. In *Proceedings of COLING-84*, pages 178–181, California, 2-4 July 1984. Stanford University, USA.

[86] W. Kraaij and R. Pohlmann. Porter's stemming algorithm for Dutch. In L.G.M. Noordman and W.A.M. de Vroomen, editors, *Informatiewetenschap 1994: Wetenschappelijke bijdragen aan de derde STINFON Conferentie*, pages 167–180, Tilburg, 1994.

[87] F. Kubala, R. Schwartz, R. Stone, and R. Weischedel. Named entity extraction from speech. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne Conference Resort Lansdowne, Virginia, February 8-11 1998.

[88] R. Lippmann and B. Carlson. Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise. In *Proceedings of Eurospeech97*, 1997.

[89] K. Livescu. *Analysis and Modeling of Non-Native Speech for Automatic Speech Recognition*. PhD thesis, MIT Department of Electrical Engineering and Computer Science, 1999.

[90] A. Ljolje. Automatic segmentation and labelling of speech. *Computer, Speech and Language*, 1994.

[91] C. Cucchiarini M. Wester, J. M. Kessens and H. Strik. Obtaining phonetic transcriptions: A comparison between expert listeners and a continuous speech recognizer. In *Language & Speech 44(3)*, pages 377–403. 2001.

[92] B. MacWhinney. *The CHILDES Project: Tools for Analyzing Talk, Second Edition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1995.

[93] M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[94] J. Markowitz. Voice biometrics: Speaker recognition applications and markets 1999. In *Voice Europe1999: European symposium on voice technologies*, 1999.

[95] E. Marsi. A reusable syntactic generator for Dutch. In P.A. Coppen, H. van Halteren, and L. Teunissen, editors, *Computational Linguistics in the Netherlands 1997*, pages 205–221. Rodopi, Amsterdam, 1999.

[96] E. Marsi. *Intonation in Spoken Language Generation*. PhD thesis, University of Nijmegen, 2001.

[97] A. Marzal. Computation of normalized edit distance and applications. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1993.

[98] A. Marzal and E. Vidal. A review and new approaches for automatic segmentation of speech signals. In *Proceedings of EUSIPCO '90*, 1990.

[99] A. Mikheev, C. Grover, and M. Moens. Xml tools and architecture for named entity recognition. *Markup Languages: Theory and Practice*, 1:89–113, 1999.

[100] N. Morgan. Continuous speech recognition using multilayer perceptrons with hidden markov models. In *Proceedings of ICASSP1990*, 1990.

[101] Y. Muthusamy. Reviewing automatic language identification. In *IEEE Signal Processing Magazine*, 1994.

[102] J. Nerbonne. Edit distance and dialect proximity. In *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. 1999.

[103] J.Y. Nie, M. Simard, P. Isabelle, and M. Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts in the web. In *Proceedings of ACM-SIGIR'99*, pages 74–81, 1999.

[104] G. Van Noord. Robust parsing of word graphs. In Jean-Claude Junqua and Gertjan van Noord, editors, *Robustness in Language and Speech Technology*. Kluwer Academic Publishers, Dordrecht, 2001.

[105] G. Van Noord, G. Bouma, R. Koeling, and M. Nederhof. Robust grammatical analysis for spoken dialogue systems. *Journal of Natural Language Engineering*, 5(1):45–93, 1999.

[106] A.M. Nunn and V.J. van Heuven. Morphon, lexicon-based text-to-phoneme conversion and phonological rules. In V.J. van Heuven and L.C.W. Pols, editors, *Analysis and synthesis of speech, strategic research towards high-quality text-to-speech generation*, pages 87–99. Mouton de Gruyter, Berlin, 1993.

[107] T. Kruyt en P. Van der Kamp P. Van Sterkenburg. Blauwdruk voor onderhoud, beheer en distributie van digitale materialen. december 2001.

[108] D. Palmer and M.A. Hearst. Adaptive sentence boundary disambiguation. In *Proceedings of the Applied Natural Language Processing Conference*, Stuttgart, October 1994.

[109] C. Peters. Results of the clef 2001 cross-language system evaluation campaign: Working notes for the clef 2001 workshop, 2001.

[110] R. Prins and G. Van Noord. Unsupervised pos-tagging improves parsing accuracy and parsing efficiency. In *IWPT 2001: International Workshop on Parsing Technologies, Beijing China*, volume 3, 2001.

[111] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, number 77, 1989.

[112] J. C. Reynar and A. Ratnaparkhi. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, D.C., March 31-April 3 1997.

[113] M. D. Riley. Some applications of tree-based modelling to speech and language. In *Proceedings of the DARPA Speech and Language Technology Workshop*, pages 339–352. Morgan Kaufman, 1989.

[114] G. D. Ritchie, G. J. Russell, A. W. Black, and S. G. Pulman. *Computational Morphology*. The MIT Press, Cambridge, Massachusetts, 1992.

[115] T. Robinson. The use of recurrent neural networks in continuous speech recognition. In *Automatic Speech and Speaker Recognition - Advanced Topics*. 1996.

[116] F. Schiel. Full automatic annotation of read, spontaneous and dialog speech corpora with maus. In *Proceedings of Cocosda97*, 1997.

[117] I. Schuurman. Anno: a multi-functional Flemish text corpus. In J. Landsbergen et al, editor, *CLIN VII. Papers from the Seventh CLIN meeting*, pages 161–176. IPO, Technische Universiteit Eindhoven, 1997.

[118] O. Siohan. On the robustness of linear discriminant analysis as preprocessing step for noisy speech recognition. In *Proceedings of ICASSP95*, 1995.

[119] R. Sproat. *Morphology and computation*. MIT Press, Cambridge, Massachusetts, 1992.

[120] H. Strik. Pronunciation adaptation at the lexical level. In J-C. Juncqua and C. Wllekens, editors, *Proceedings of the ISCA Tutorial & Research Workshop (ITRW) 'Adaptation Methods For Speech Recognition', Sophia-Antipolis, France, August 29-30*, pages 123–131, 2001.

[121] H. Strik and C. Cucchiarini. Modeling pronunciation variation for asr: a survey of the literature. In *Speech Communication 29 (2-4)*, pages 225–246. 1999.

[122] N. Ström. Speaker adaptation by modeling the speaker variation in a continuous speech recognition system. In *Proceedings of ICSLP '96*, 1996.

[123] N. Ström. Speaker modeling for speaker adaptation in automatic speech recognition. In *Talker Variability in Speech Processing*. 1997.

[124] J. 't Hart. *A perceptual study of intonation*. Cambridge University Press, 1990.

[125] M. Tatham. Spruce speech synthesis for dialogue systems. In *Proceedings of Multimodal Dialogue Workshop*, 1993.

[126] A. Teixera. Recognition of non-native accents. In *Proceedings of Eurospeech97*, 1997.

[127] H. ter Doest. Towards probabilistic unification-based parsing. Master's thesis, Universiteit Twente, 1999.

[128] D. Torre. Automatic alternative transcription generation and vocabulary selection for flexible word recognizers. In *Proceedings of ICASSP97*, 1997.

[129] P.C. uit den Boogaart (red.). *Woordfrequenties in geschreven en gesproken Nederlands*. Utrecht: Oosthoek, Scheltema en Holkema, 1975.

[130] R. van Bezooijen. Assessment of synthesis systems. In *Handbook of standards and resources for spoken language technology*. 1997.

[131] D. van Compernolle. Spectral estimation using a log-distance error criterion applied to speech recognition. In *Proceedings of ICASSP89*, 1989.

[132] D. van Compernolle. Recognizing speech of goats, wolves, sheep and ... non-natives. In *Speech Communication*, 2001.

[133] A. van den Bosch and W. Daelemans. Memory-based morphological analysis. In *Proceedings of the 37th annual meeting of the association for computational Linguistics*, pages 285–292, University of Maryland, USA, June 20-26 1999.

[134] H. van den Heuvel. Slr validation: evaluation of the speechdat approach. In *Proceedings LREC 2000 Satellite workshop XLDB - Very large Telephone Speech Databases*, 2000.

[135] H. van den Heuvel. Slr validation: Present state of affairs and prospects. In *Proceedings LREC 2000*, 2000.

[136] I. van der Sluis and F. de Jong. Enriching textual documents with time-codes from video fragments. In *Proceedings RIAO 2000, Paris*, pages 431–440, 2000.

[137] P. van der Wijst. Politeness in requests and negotiations. Master's thesis, KU Brabant, 1996.

[138] H. van Halteren, editor. *Syntactic Wordclass Tagging*. Dordrecht: Kluwer, 1999.

[139] H. van Halteren, J. Zavrel, and W. Daelemans. Improving accuracy in nlp through combination of machine learning systems. *Computational Linguistics*, 5:2071–2074, to appear.

[140] G. Veldhuijzen van Zanten, G. Bouma, K. Sima'an, G. van Noord, and R. Bonnema. Evaluation of the nlp components of the OVIS2 spoken dialogue system. In F. van Eynde, I. Schuurman, and N. Schelkens, editors, *Computational Linguistics in the Netherlands 1998*, pages 213–229. Rodopi, Amsterdam, 1999.

[141] A. Vorstermans. Automatic segmentation and labelling of multi-lingual speech data. *Speech Communication*, 1996.

[142] E.A. Wan and A.T. Nelson. Networks for speech enhancement. In S. Katagiri, editor, *Handbook of neural networks for speech processing*. Artech House, Boston, USA, 1999.

[143] A. Weijters and G. Hoppenbrouwers. Netspraak: een neuraal netwerk voor grafeem-foneem-omzetting. *TABU 20*, 1:1–25, 1990.

[144] T. Westerveld. Image retrieval: Content versus context. In *Proceedings RIAO 2000, Paris*, pages 276–284, 2000.

[145] G. Williams. A study of the use and evaluation of confidence measures in automatic speech recognition. Technical report, Department of Computer Science, University of Sheffield, 1998.

[146] J. Zavrel and W. Daelemans. Evaluatie van part-of-speech taggers voor het Corpus Gesproken Nederlands. Technical report, CGN, 1999.

[147] M. Zissman. Automatic language identification. In *Proceedings van MIST: Multi-lingual Interoperability in Speech Technology*, 2000.