De positie van het Nederlands in Taal- en Spraaktechnologie

Gosse Bouma
en
Ineke Schuurman

De positie van het Nederlands in Taal- en Spraaktechnologie

Gosse Bouma
Rijksuniversiteit Groningen
Alfa-informatica
gosse@let.rug.nl
en
Ineke Schuurman
Katholieke Universiteit Leuven
Centrum voor Computerlinguïstiek
ineke.schuurman@ccl.kuleuven.ac.be

Augustus 1998

Voorwoord

Dit rapport is het verslag van een terreinverkennend onderzoek naar de positie van het Nederlands in de taal- en spraaktechnologie, dat werd uitgevoerd in opdracht van de Nederlandse Taalunie, in de periode van oktober 1997 tot en met juni 1998.

Het onderzoek werd begeleid door Jeannine Beeken (Nederlandse Taalunie), Walter Daelemans (Katholieke Universiteit Brabant en Universiteit Antwerpen, UIA), Elisabeth D'Halleweyn (Nederlandse Taalunie), en Bruno Krekels (Vlaams Instituut voor de Bevordering van het Wetenschappelijk-technologisch onderzoek in de Industrie - IWT). Alle begeleiders hebben actief meegewerkt aan het tot stand komen van dit rapport, met name door het aandragen van informatie, het nader preciseren van de uitgangspunten van het onderzoek, het leveren van commentaar op voorlopige versies, en het mede vormgeven van de aanbevelingen. Daarnaast heeft Walter Daelemans een deel van de interviews voor zijn rekening genomen, en leverde Elisabeth D'Halleweyn de tekst van sectie 1.8 (juridische aspecten). Wij danken hen voor hun bijdrage aan dit rapport.

We willen hierbij ook de leden van de begeleidingscommissie bedanken voor hun opbouwende commentaar tijdens twee bijeenkomsten waarin de voorlopige resultaten van dit rapport werden besproken.

In het kader van dit onderzoek hebben we met enige tientallen personen gesproken die actief betrokken zijn bij de taal- en spraaktechnologie in Nederland en Vlaanderen (zie hoofdstuk 4). We willen onze gesprekspartners bedanken voor deze, soms zeer uitvoerige, gesprekken die een grote hoeveelheid nuttige informatie opleverden. Een aantal van hen heeft ook gereageerd op een eerdere versie van dit rapport, waardoor we nog een aantal correcties en aanvullingen konden verwerken.

Het *world wide web* is inmiddels de meest actuele en uitgebreide bron voor informatie over onderzoeksprojecten, software-producten, elektronische dataverzamelingen, *etc.* We hebben voor dit onderzoek dan ook intensief gebruik gemaakt van informatie die te vinden is op het internet. In dit rapport zijn daarom een groot aantal verwijzingen naar web-pagina's, zoals ze golden ten tijde van het schrijven van dit rapport, opgenomen. Een web-versie van dit rapport is beschikbaar als `http://www.let.rug.nl/~gosse/taalunie/webrapport`.

De belangrijkste resultaten van dit onderzoek zijn gepresenteerd op de door ELRA georganiseerde *First International Conference on Language Resources and Evaluation*, Granada, 28-30 Mei, 1998 (Bouma en Schuurman 1998).

Augustus 1998,
Gosse Bouma en Ineke Schuurman

Begeleidingscommissie

De begeleidingscommissie bestond uit de volgende leden:

Industrie:

dr. A. Zaenen (RANK XEROX)
prof. dr. D. Van Compernolle (Lernout & Hauspie/ESAT)
dr. E. den Os (KPN)

Europese Unie:

dhr. J. Roukens (EG, DG XIII)

Nederlandse overheid:

drs. R. Spoor (OC&W/OWB)
drs. H. de Brabander (Economische Zaken)

Vlaamse overheid:

ir. B. Krekels (IWT)
dr. ir. P. Dengis (Ministerie van de Vlaamse Gemeenschap/AWI)
dr. L. Vanfleteren (Kabinet van minister-president Van den Brande)
dr. F. Buekens (Ministerie van de Vlaamse Gemeenschap/dept. Onderwijs)

Wetenschapsfondsen:

drs. A.Dijkstra (NWO)
dhr. J. Traest (FWO)

Wetenschap:

dr. W. Daelemans (KU Brabant, Universiteit Antwerpen)
prof. dr. F. Van Eynde (KU Leuven)
prof. dr. G. Kempen (RU Leiden)

Nederlandse Taalunie:

drs. E. D'Halleweyn
dr. J. Beeken

Inhoudsopgave

Samenvatting

Hoofdstuk 1: Inleiding en Uitgangspunten

De **taal- en spraaktechnologie** (TST) houdt zich bezig met onderzoek naar de mogelijkheden om taal en spraak automatisch te herkennen, te analyseren, en te produceren, en met toepassingen van deze techniek binnen de informatie- en communicatietechnologie. Hierbij kan men denken aan spraakherkenning, het omzetten van tekst in spraak, spellingcorrectie, automatisch vertalen, samenvatten, en het classificeren van documenten.

Wanneer men in de toekomst in het Nederlands gebruik wil kunnen maken van alle mogelijkheden die TST biedt, is het noodzakelijk dat er TST-producten voor het Nederlands zijn. Een goede positie van het Nederlands binnen de TST is in het belang van de overheid (omdat hierdoor de positie van het Nederlands gewaarborgd wordt), van het bedrijfsleven (omdat het de communicatie met een belangrijke markt optimaliseert) en onderwijs en wetenschap (omdat hierdoor aansluiting bij internationaal onderzoek gewaarborgd wordt).

Voor het maken van Nederlandstalige TST-producten zijn evenwel een aantal hulpmiddelen nodig. De investeringen die nodig zijn om hulpmiddelen als elektronische woordenboeken en geannoteerde corpora te ontwikkelen, gaan echter de mogelijkheden van de meeste individuele onderzoeksgroepen en bedrijven te boven. Voor bedrijven die zich een dergelijke investering wel kunnen veroorloven speelt de vraag of een middelgrote taal als het Nederlands deze investering wel waard is. Om deze impasse te doorbreken is een gezamenlijke investering nodig in dataverzamelingen en hulpmiddelen. Het resultaat van deze investeringen is een deugdelijke *infrastructuur* voor TST, dat wil zeggen een situatie waarin de meest noodzakelijke, algemene, hulpmiddelen beschikbaar zijn. Een infrastructuur voor TST levert daarmee een belangrijke bijdrage aan de ontwikkeling van Nederlandtalige TST-producten.

Hoofdstuk 2: De huidige situatie

De materiële infrastructuur bestaat uit een groot aantal verschillende componenten, zoals spraakcorpora en tekstcorpora, elektronische woordenboeken, en programmatuur (voor spraakherkenning, spraaksynthese, voor het ontleden en vertalen van woorden, zinnen, en teksten, etc.). In hoofdstuk 2 geven we een overzicht van de belangrijkste hulpmiddelen voor het Nederlands, en een (schetsmatig) overzicht van belangrijke hulpmiddelen voor andere talen.

De immateriële infrastructuur bestaat uit het geheel van instellingen (voor onderzoek, onderwijs, en beleid) en bedrijven die actief zijn op het gebied van TST. In hoofdstuk 2 wordt ook een overzicht gegeven van instellingen en bedrijven in Nederland en Vlaanderen, en van belangrijke instellingen op dit terrein in het buitenland.

Hoofdstuk 3: Evaluatie

De bestaande materiële infrastructuur voor TST voor het Nederlands kent een aantal hiaten en zwakke plekken. Dit blijkt zowel uit het overzicht dat wordt gegeven in hoofdstuk 2 als uit de interviews met experts.

Aan corpora voor het Nederlands is er, naast een Europese CD-ROM, met daarop vijf miljoen woorden ruwe Nederlandstalige tekst, vrijwel niets dat gemakkelijk verkrijgbaar of toegankelijk is. De hoeveelheid beschikbare tekst is gering, maar belangrijker nog is het feit dat geannoteerde corpora vrijwel niet bestaan. De corpora die wellicht in de nabije toekomst beschikbaar komen of die in voorbereiding zijn, zullen in deze situatie maar weinig verandering kunnen brengen.

Het Nederlands-Vlaamse project voor een Corpus Gesproken Nederlands is een stimulans voor onderzoek naar gesproken taal. Het meest succesvolle hulpmiddel voor TST-onderzoek is de lexicale database van CELEX. De toekomst van CELEX is momenteel onzeker. In concreto zijn er momenteel geen lexica beschikbaar die voorzien in gedetailleerde syntactische en semantische informatie. Dit betekent dat een aantal toepassingen niet goed mogelijk zijn.

De situatie op het gebied van programmatuur (spraakherkenning en systemen voor tekst-naar-spraak, morfologische en syntactische analyse, automatisch vertalen) is tamelijk zorgelijk. Een verbetering van deze situatie kan alleen bereikt worden wanneer de noodzakelijke hulpmiddelen, in de vorm van corpora, woordenboeken, en testmateriaal, beschikbaar komen.

Hoofdstuk 4: Interviews

In het kader van dit onderzoek is met een dertigtal personen gesproken die op enigerlei wijze betrokken zijn bij TST.

Uit de interviews kwam naar voren dat men verwacht dat met name toepassingen van spraaktechnologie, en toepassingen op het gebied van *information* en *document retrieval* en *extraction* in de toekomst belangrijk zullen zijn. Men is vrij algemeen bekend met de lexicale database van CELEX, maar kent daarnaast slechts enkele hulpmiddelen voor het Nederlands. Voor sommige andere talen (met name het Engels) is er aanzienlijk meer beschikbaar.

Er blijkt een vrij algemene behoefte te bestaan aan grotere corpora, die rijk geannoteerd zijn. Daarnaast is er behoefte aan verschillende vormen van lexicale informatie. Een aanzienlijk aantal respondenten heeft behoefte aan meer formele en computationele beschijvingen en implementaties van de Nederlandse grammatica. Men wijst ook op het feit dat er niets beschikbaar is dat evaluatie van bestaande hulpmiddelen mogelijk zou maken.

Men staat positief tegenover het ontwikkelen van een basiscollectie met hulpmiddelen. Randvoorwaarden zijn dat het materiaal tegen een redelijke vergoeding beschikbaar moet zijn (ook voor het bedrijfsleven), en dat er garanties zijn voor continuïteit. Met name het bedrijfsleven verwacht dat met de beschikbaarheid van basisvoorzieningen de mogelijkheden om commerciële toepassingen te ontwikkelen zullen toenemen.

Universitaire instellingen hebben vooral behoefte aan materiaal dat voor onderwijs en (fundamenteel) onderzoek gebruikt kan worden.

Men is somber over de mogelijkheden om gekwalificeerd personeel te vinden. Dit geldt vooral voor meer technisch onderlegde medewerkers. Er heerst een vrijwel algemene onvrede over het huidige beleid met betrekking tot TST. Men verwacht echter dat het mogelijk moet zijn om binnen de bestaande structuren te komen tot een beter beleid en betere samenwerking. Anderzijds wordt ook de situatie in Griekenland of Denemarken als voorbeeld genoemd, waar een nationale instelling verantwoordelijk is voor o.a. het beheer van TST-hulpmiddelen. Degenen die voor een nieuwe instelling zijn denken dan ook vooral aan een instituut dat het technisch beheer van hulpmiddelen op zich neemt, en niet direct aan een beleidsinstantie. De meningen over de mogelijke rol van de Taalunie op het gebied van TST zijn verdeeld.

Hoofdstuk 5: Aanbevelingen

Aanbeveling 1: Het instellen van een Nederlands-Vlaams platform met als primaire taak het coördineren van activiteiten op het gebied van taal- en spraaktechnologie voor het Nederlands. De Taalunie zou hierbij als initiator en coördinator kunnen optreden.

Aanbeveling 2: Het stimuleren van zowel fundamenteel als toegepast onderzoek op het gebied van TST.

Aanbeveling 3: Het opzetten van een speciale (interuniversitaire) opleiding voor taal- en spraaktechnologie in Vlaanderen en het versterken van de opleidingen op dit gebied in Nederland.

Hoofdstuk 1

Inleiding en Uitgangspunten

1.1 Het belang van taal- en spraaktechnologie voor het Nederlands

Dit rapport is het verslag van een onderzoek naar de positie van het Nederlands in de taal- en spraaktechnologie. De **taal- en spraaktechnologie** (TST) houdt zich bezig met onderzoek naar de mogelijkheden om taal en spraak automatisch te herkennen, te analyseren, en te produceren, en met toepassingen van deze techniek binnen de informatie- en communicatietechnologie. Hierbij kan men denken aan programma's voor spraakherkenning en spraaksynthese (het automatisch produceren van gesproken taal), spellingcorrectie, automatisch vertalen, samenvatten, en het classificeren van documenten.

De mogelijkheden van TST worden steeds groter, en spelen in toenemende mate een rol in het dagelijks leven. We zijn reeds lang gewend aan tekstverwerkers die spellingcorrectie uitvoeren, die een synoniemenfunctie bevatten, en die woorden op het eind van een regel kunnen afbreken. De huidige generatie tekstverwerkers bevat modules die grammaticale fouten (zoals de beruchte *d/t*-fouten voor het Nederlands) opspeuren en die stilistische adviezen geven. De Nederlandse organisatie voor informatie over het openbaar vervoer, OVR, heeft onlangs, in navolging van landen als Duitsland en Zwitserland, een sprekende computer in gebruik genomen, die klanten informatie geeft over treinverbindingen. Daarbij moet worden opgemerkt dat deze computer in staat is een uitvoerige dialoog met de klant aan te gaan, waarbij de klant zich niet hoeft te beperken tot het inspreken van losse stationsnamen of tijdstippen. Wanneer men bedenkt dat tot voor enige jaren spraakherkenning zich beperkte tot het herkennen van een zeer beperkt aantal woorden, dat dergelijke systemen sprekerafhankelijk waren, en dat spraakherkenning via een telefoonlijn al helemaal uitgesloten was, moge duidelijk zijn dat er enorme vooruitgang is geboekt op dit terrein. Automatisch vertalen is misschien de oudste toepassing van de taal- en spraaktechnologie, en een toepassing waar lange tijd weinig vooruitgang werd geboekt. Ondertussen zijn er echter een aantal commerciële systemen beschikbaar die heel behoorlijk presteren. Met name pakketten die op maat gemaakt zijn voor toepassingen binnen het bedrijfsleven hebben voldoende kwaliteit om commercieel interessant te zijn. Daarnaast lijkt ook de consumentenmarkt langzamerhand binnen het bereik van dergelijke systemen. Een voorbeeld is de vertaalmogelijkheid (gemaakt door het bedrijf Systran) die sinds kort is toegevoegd aan AltaVista, een zoekmachine voor het internet. Overigens is opvallend dat de vertaalmodule vertaalt van of naar het Engels, Duits, Frans, Italiaans, Spaans, en Portugees, maar dat het Nederlands ontbreekt. Tenslotte speelt taal- en spraaktechnologie in toenemende mate een rol in het talenonderwijs (onder andere in de vorm van elektronische woordenboeken, oefenprogramma's voor grammatica, en interactieve CD-ROM's die de uitspraak van de student kunnen controleren en verbeteren) en als hulpmiddel voor gehandicapten (bijvoorbeeld door mensen met een stemprobleem te voorzien van een programma dat tekst in spraak om kan zetten, of door (elektronische) documenten voor te lezen aan mensen met zichtproblemen).

Nu er langzamerhand sprake is van een daadwerkelijke markt voor TST (die recentelijk nog uitvoerig in kaart is gebracht door het Europese EUROMAP project, (Dewallef 1998, van Staden 1998)), wordt duidelijk dat er binnenkort nog meer toepassingen mogelijk zullen zijn, en dat de mogelijkheden van bestaande systemen nog verbeterd kunnen worden. Over de kwaliteit van spelling- en grammaticacorrectie voor het Nederlands is lang niet iedereen tevreden. Toch valt te voorzien dat hier met relatief kleine investeringen kwalitatief goede systemen mogelijk moeten zijn. Wanneer de kwaliteit van spraakherkenning verder toeneemt zullen bijvoorbeeld dicteersystemen niet alleen meer zijn voorbehouden aan beroepsgroepen als juristen en medici, en zal er een heel scala aan telefonische informatiediensten ontstaan die geheel of grotendeels gebaseerd zijn op het gebruik van sprekende computers. In Duitsland wordt gewerkt aan een systeem (VERBMOBIL) dat vertalen op basis van gesproken taal mogelijk moet maken, zodat men bijvoorbeeld in de toekomst zonder tussenkomst van een tolk een telefoongesprek kan voeren met iemand uit een ander taalgebied. *Voice dialing* (het kiezen van een telefoonnummer met behulp van spraakherkenning) heeft inmiddels ook in Nederland en Vlaanderen zijn intrede gedaan. De bediening van apparatuur die deel uitmaakt van de hedendaagse auto, zoals een telefoon, een stereo-installatie, en een navigatiesysteem, zal in de zeer nabije toekomst spraakgestuurd zijn, en daarna zal deze techniek waarschijnlijk toegepast worden in meer situaties.

Wanneer men in de toekomst in het Nederlands gebruik wil kunnen maken van alle mogelijkheden die TST biedt, is het noodzakelijk dat de voorzieningen die essentieel zijn voor het ontwikkelen van dergelijke toepassingen aanwezig zijn en van voldoende kwaliteit zijn.

In dit rapport inventariseren we de stand van zaken met betrekking tot deze 'infrastructuur' voor taal- en spraaktechnologie voor het Nederlands. In dit hoofdstuk gaan we nader in op de vraag waarom de positie van het Nederlands in TST een zaak van algemeen belang is, wat er onder een infrastructuur voor TST moet worden verstaan, en wat overheid, bedrijfsleven, en universiteit van een goede infrastructuur mogen verwachten. In hoofdstuk 2 bespreken we welke hulpmiddelen en dataverzamelingen voor TST voor het Nederlands beschikbaar zijn, en welke organisaties en instellingen (op nationaal, Vlaams/Nederlands, en Europees niveau) zich bezig houden met TST-beleid. In hoofdstuk 3 evalueren we de materiële infrastructuur (het geheel van hulpmiddelen en dataverzamelingen). In hoofdstuk 4 doen we verslag van de interviews die we hebben gevoerd met een dertigtal deskundigen in Nederland en Vlaanderen. Op basis van het overzicht van de infrastructuur, onze evaluatie van deze infrastructuur, en de mening van de deskundigen komen we in hoofdstuk 5 tot een aantal aanbevelingen, die zijn gericht op het verbeteren en versterken van de infrastructuur voor TST, zowel in materieel als in organisatorisch/immaterieel opzicht. We besteden

hierbij speciale aandacht aan de rol die de Nederlandse Taalunie zou kunnen spelen.

1.2 Aandacht voor de digitalisering van het Nederlands

> Als ik denk aan de technologie, is het niet omdat ik het spijtig vind dat die wereld louter Engelstalig is, niet omdat ik het aardig zou vinden dat onze taal daar vertegenwoordigd zou zijn, maar wel omdat ik ervoor wil zorgen dat de Nederlandstalige erin kan participeren zonder dat dit afhangt van zijn kennis van vreemde talen.
>
> Koen Jaspaert, De Standaard (11-04-'98)

Natuurlijke talen vervullen een aantal functies: praktisch communicatiemiddel voor het dagelijks leven, instrument voor het verschaffen van informatie, voor het overdragen van kennis, voor het bedrijven van wetenschap, voor het beoefenen van rechtspraak, politiek en bestuur, een uitdrukkingsmiddel voor cultuur.

Van den Bergh (1996) stelt dat een taal die te veel van deze functies verliest vroeger of later in de gevarenzone terechtkomt. Dit doordat mensen het bijvoorbeeld te lastig vinden om van de ene naar de andere taal over te schakelen, of doordat de ene taal meer status heeft dan de andere. Dit heeft tot gevolg dat de meest "volledige" taal, of die met de hoogste status, aan invloed wint. Het verdient de voorkeur een dergelijk scenario te vermijden. Immers, zoals een van de door ons geïnterviewde personen het uitdrukte: "je taal is net zo van jezelf als de kleur van je ogen, en dat moet zo blijven". Er zijn derhalve verschillende motieven om de (technologische) positie van het Nederlands te versterken: culturele, economische, politieke, en sociale (zie ook Cherribi en Sannen (1998))

Het niet participeren in nieuwe technische ontwikkelingen, in dit geval het niet meegaan in de digitalisering van de taal (zie Van Eynde (1996); Van den Bergh (1996)), kan een belangrijke oorzaak van functieverlies zijn. Er zijn tot dusver drie taaltechnologische "revoluties" geweest: 1) de uitvinding van het schrift, 2) de uitvinding van de boekdrukkunst en 3) de uitvinding van de computer. De talen die niet hebben geparticipeerd in de eerste twee revoluties zijn uiteindelijk in een marginale positie beland. Te vrezen valt dat dit ook voor de derde revolutie zal gelden. Het participeren is hier minstens zo complex als bij voorgaande revoluties: hiervoor moet speciaal, namelijk *digitaal*, Nederlandstalig basismateriaal voorhanden zijn, respectievelijk beschikbaar komen. Hierbij gaat het om woordenlijsten, grammatica's, (meertalige) woordenboeken, enz. Soms kan worden volstaan met het digitaliseren van hetgeen er op papier voorhanden was: een niet-geannoteerd corpus kan bijvoorbeeld ingescand worden. In verreweg de meeste gevallen volstaat dit echter niet: een elektronische versie van een woordenboek moet aan heel andere eisen voldoen dat de papieren versie, wil ze geschikt zijn voor TST-toepassingen. Andere hulpbronnen, zoals grammatica- en stijlcorrectors, moeten (bijna) helemaal opnieuw worden ontwikkeld.

Digitalisatie kost derhalve veel geld. Er mag bovendien niet veel tijd verloren gaan. Als er eenmaal een overvloed aan Engelstalig materiaal aanwezig is valt te vrezen dat er bij de industrie (vanuit commerciële overwegingen) weinig of geen belangstelling meer zal zijn om ook nog iets voor het Nederlands te gaan ontwikkelen.

Voor de niet-grote taalgebieden is de financiële last van het ontwikkelen van taaltechnologische basisproducten erg zwaar, onder meer doordat het economisch draagvlak kleiner is. Het bedrijfsleven en de overheden (zowel nationale als Europese, bijvoorbeeld in het kader van het MLIS[1]-programma), zullen die inspanningen samen moeten doen.

1.3 De huidige stand van zaken

Wie onderzoek wil gaan doen op het gebied van taal- en spraaktechnologie, of een product wil gaan ontwikkelen waarin deze technologie een rol speelt, zal worden geconfronteerd met het feit dat veel van het benodigde materiaal ontbreekt. Dit geldt in het bijzonder voor de 'kleinere' en 'middelgrote' talen, zoals het Nederlands. Het ontbreken van elektronische woordenboeken, corpora, en hulpmiddelen voor het ontwikkelen van TST-producten heeft een negatief effect op de efficiëntie waarmee wetenschappelijk onderzoek en de ontwikkeling van commerciële producten kan worden uitgevoerd. Ook de kwaliteit van het uiteindelijke resultaat wordt nadelig beïnvloed. Tenslotte werkt het gebrek aan makkelijk toegankelijke materialen en informatiebronnen onnodig drempelverhogend voor onderzoeksteams die zich met het vakgebied willen gaan bezighouden. Dit betekent dat bepaalde producten wellicht niet beschikbaar komen.

We illustreren de gevolgen van de huidige situatie op het gebied van TST voor het Nederlands hieronder met een drietal voorbeelden, ontleend aan onze eigen ervaringen en die van onze interviewpartners (zie hoofdstuk 4), en geven aan waarom deze voorbeelden wijzen op een onderontwikkelde infrastructuur.

Voorbeeld 1: Het automatisch toekennen van woordsoorten

Een bedrijf dat zich bezig houdt met het ontwikkelen van intelligente grammatica-controle voor het Nederlands (zie Vosse, 1994) wil graag gebruik maken van een *part-of-speech tagger*. Een POS *tagger* is een programma dat aan de woorden in een tekst de juiste woordsoort toekent (en dat dus het gebruik van het woord *bedrijven* als zelfstandig naamwoord kan onderscheiden van het gebruik als werkwoord). Het bepalen van woordsoorten is een belangrijke eerste stap voor grammatica-controle. Bij de ontwikkeling van het programma ontmoet men de volgende problemen:

- Er zijn nauwelijks corpora beschikbaar die zijn voorzien van woordsoort en die kunnen worden gebruikt om het systeem te trainen en te testen. De corpora die er zijn, kunnen alleen langs informele kanalen ter beschikking worden gesteld, waarbij onduidelijk blijft of er auteursrechtelijke belemmeringen zijn voor het gebruik in (de ontwikkeling van) een commercieel product. Ook de kwaliteit van de corpora is niet altijd duidelijk (is het corpus gecorrigeerd op fouten of niet?).

- Er is geen duidelijke standaard voor het toekennen van woordsoorten voor het Nederlands. Binnen een Europees project als EAGLES (gewijd aan de ontwikkeling van standaards en evaluatiecriteria) is wel een aanzet geleverd voor de ontwikkeling van dergelijke standaard, maar dit heeft vooralsnog niet geresulteerd in een eenduidige EAGLES-*tagset* voor het Nederlands. De documentatie van bestaande programma's die automatisch woordsoorten toekennen is uiterst summier waar het de keuze van de gebruikte woordsoorten betreft (vaak niet meer dan een opsomming van de gebruikte categorieën).

- Er zijn geen mogelijkheden om de prestaties van het programma te evalueren. Er zijn geen corpora beschikbaar die kunnen worden gebruikt om de foutenmarge van het eigen programma te bepalen en om de prestaties van het eigen programma te vergelijken met die van anderen.

Voorbeeld 2: Automatisch zinsontleden

Een tweede belangrijke component voor een programma dat grammatica-controle uitvoert is een computationele grammatica voor het Nederlands. Nederland kent een rijke taalkundige traditie, en de zinsbouw van het Nederlands is dan ook vrij grondig beschreven (bijvoorbeeld in naslagwerken als de *Algemene Nederlandse Spraakkunst* (ANS) (Haesereyn, Romijn, Geerts, De Rooy, en Van den Toorn 1997) en in benaderingen vanuit een bepaald theoretisch kader, zoals in Model (1991)). De kloof tussen een taalkundige beschrijving en dat wat nodig is voor een computationele grammatica is evenwel aanzienlijk:

- Een belangrijk hulpmiddel voor het ontwikkelen van computationele grammatica's is een corpus, waarin iedere zin is voorzien van een constituentstructuur (en wellicht ook informatie over de syntactische functie van de verschillende zinsdelen), gebaseerd op algemeen aanvaarde (en gedocumenteerde) grammaticale regels. Een grammaticaal geannoteerd corpus kan zowel bij de ontwikkeling als bij de evaluatie van een grammatica een rol spelen. Een dergelijk corpus ontbreekt.

- Er zijn geen elektronische woordenboeken beschikbaar waarin voor ieder woord is vastgelegd wat de grammaticale (met name combinatorische) eigenschappen van het woord zijn. Voor grammaticale analyse is het niet voldoende om te weten dat een bepaald woord een werkwoord is, het moet ook bekend zijn aan welke eisen het onderwerp moet voldoen, of het werkwoord overgankelijk is of niet, of het werkwoord voorkomt met een voorzetselvoorwerp of niet, enz. Het ontwikkelen van een omvangrijk woordenboek waarin dergelijke informatie is vastgelegd, is arbeidsintensief en kostbaar.

Voorbeeld 3: Spraaksynthese

Bij Europese projecten voor spraaktechnologie wordt in toenemende mate verlangd dat voldoende hulpmiddelen en corpora beschikbaar zijn voor de talen waarop men zich richt. Aangezien deze hulpmiddelen voor het Nederlands niet altijd beschikbaar zijn, kost het moeite om de aansluiting te behouden:

- Om te kunnen participeren in een project op het gebied van tekst-naar-spraaksynthese is de beschikbaarheid van goede spraaksynthese essentieel. Een

[1] *multilingual information society*

dergelijk product komt er alleen als er uitgebreide en gedetailleerd geannoteerde corpora beschikbaar zijn, waarin voor ieder deel van de geluidsopname nauwkeurig is aangegeven met welke zin, met welk woord, en zelfs met welke klank het overeenkomt. Het ontbreken van dergelijke corpora voor het Nederlands, en de producten die ervan afgeleid kunnen worden, vormt een obstakel voor samenwerking met onderzoeksgroepen die werken aan talen waarvoor dergelijke hulpmiddelen wel beschikbaar zijn.

- Een vergelijkbare situatie doet zich voor bij onderzoek naar spraak-naar-spraak vertalen. De ontwikkeling van een spraakgebaseerde automatische vertaler bestaat voor een belangrijk deel uit het op de juiste wijze combineren van systemen voor spraakherkenning, automatisch vertalen, en spraaksynthese. Omdat de meeste vertaalsystemen geen Nederlandse module bevatten, is het voor Nederlandse onderzoeksgroepen onmogelijk om te participeren in projecten op dit gebied.

Deze voorbeelden zijn illustratief voor de huidige situatie. Er zijn momenteel een aantal bedrijven serieus bezig met de *ontwikkeling* van TST. Deze bedrijven zien in principe ook mogelijkheden om TST-producten voor het Nederlands op de markt te brengen. Daarnaast zijn verschillende bedrijven met name geïnteresseerd in het *toepassen* van TST binnen op maat geleverde software. Tenslotte is er binnen de academische wereld een behoorlijke belangstelling voor toegepast onderzoek op TST-gebied. Voor de ontwikkeling van TST-producten zijn echter hulpmiddelen nodig, met name in de vorm van corpora en woordenboeken, en deze hulpmiddelen zijn niet of onvoldoende beschikbaar.

Bedrijven die zich vooral richten op het toepassen van TST in andere producten hebben vooral behoefte aan software-modules voor TST, zoals modules voor spraakherkenning, morfologische of syntactische analyse, etc. Het ontbreken van hulpmiddelen en modules voor TST is een obstakel voor het efficiënt ontwikkelen van TST-producten. De investeringen die nodig zijn om bijvoorbeeld een elektronisch valentiewoordenboek samen te stellen of om een groot, geannoteerd, corpus hedendaags Nederlands aan te leggen, gaan de mogelijkheden van de meeste individuele onderzoeksgroepen en bedrijven te boven.[2] Voor bedrijven die zich een dergelijke investering wel kunnen veroorloven speelt de vraag of een middelgrote taal als het Nederlands deze investering wel waard is. Om deze impasse te doorbreken is een investering nodig in, algemeen beschikbare, dataverzamelingen en hulpmiddelen. Het subsidiariteitsbeginsel[3] indachtig is dit het moment waarop de overheid zou moeten bijspringen: immers, industrie en wetenschappelijke instituten kunnen de benodigde middelen niet alleen opbrengen.

[2] De benodigde investeringen zullen des te hoger zijn wanneer we er ons rekenschap van geven dat er, vooral bij aan spraak gerelateerde projecten, ook aandacht zal moeten worden geschonken aan taalvariatie (regio, sexe, leeftijd, . . .).

[3] Zaken die door een lager orgaan kunnen worden verricht worden niet door een hoger ter hand genomen.

1.4 Naar een infrastructuur voor TST

Er zijn in principe drie categorieën actoren actief op het terrein van de taal- en spraaktechnologie: overheidsinstellingen, bedrijfsleven, en wetenschappelijke instituten. In deze sectie zetten we uiteen welk belang de verschillende actoren hebben bij een goede positie van het Nederlands binnen de taal- en spraaktechnologie en bepalen we welke doelstellingen op het gebied van TST bereikt kunnen worden door een versterking van de infrastructuur.

1.4.1 De belangen van overheid, bedrijfsleven, en wetenschap

De overheid is verantwoordelijk voor het in stand houden van de Nederlandse taal in al haar functies (als cultuurtaal, als taal van de overheid, de rechtspraak, en het onderwijs, en als taal in het economisch verkeer). Dit betekent in toenemende mate aandacht voor de rol van het Nederlands in de informatietechnologie. De overheid kan bijvoorbeeld stimulerend optreden wanneer de ontwikkeling van een bepaalde productcategorie (spraakherkenning, grammaticacorrectie) niet of te traag voor het Nederlands tot stand komt, of wanneer in multilinguale (vertaal-)programma's het Nederlands niet aan bod komt. De overheid is daarnaast betrokken bij de taal- en spraaktechnologie doordat ze de belangrijkste beleidsmaker en fondsenverstrekker is op het gebied van onderwijs en onderzoek, ze een belangrijke rol speelt in het industriebeleid, en daarnaast in een aantal Europese programma's (zoals bijvoorbeeld MLIS), het initiatief heeft bij het verwerven van Europese fondsen.[4]

Het bedrijfsleven heeft er baat bij wanneer ze een (thuis-)markt van circa 21 miljoen personen op de juiste wijze kan bedienen. Binnen de (zakelijke) dienstverlening en binnen de informatie- en communicatietechnologie speelt TST in toenemende mate een rol. Voor zover het bedrijfsleven zich richt op de Nederandstalige markt is het van belang dat hierbij gebruikt kan worden gemaakt van innovatieve en concurrerende ICT-producten die, waar nodig, gebruik maken van Nederlandstalige TST. Om dergelijke producten te kunnen ontwikkelen zal het bedrijfsleven vaak afhankelijk zijn van samenwerking met universitaire partners en van financiële steun van de nationale overheid of de EU.

Wetenschappelijke instellingen die actief zijn op het gebied van TST dienen onderzoeks- en onderwijsprogramma's uit te voeren die innovatief en van internationaal niveau zijn. Wanneer men zich hierbij niet alleen wil beperken tot onderzoeksvragen gericht op de Engelse (of Duitse of Franse) taal, dienen de hulpmiddelen die noodzakelijk zijn voor onderzoek en onderwijs op het gebied van TST voor het Nederlands beschikbaar te zijn.

1.4.2 Doelstellingen voor een TST-infrastructuur

Een infrastructuur voor TST dient een bijdrage te leveren aan de versterking van de positie van het Nederlands binnen TST. Het realiseren van deze doelstelling is een

[4] De overheid zou, cf. Roukens (1998), ook op dergelijke Europese programma's een beroep kunnen doen, daarnaast zou ze, in geval van het ontwikkelen van bi-, respectievelijk multilinguale hulpbronnen, op bilaterale basis, de andere nationale overheden om cofinanciering kunnen verzoeken.

gezamenlijk belang van bovengenoemde actoren: voor de overheid wordt op deze manier de positie van het Nederlands gewaarborgd; voor het bedrijfsleven wordt op deze manier de communicatie met een belangrijke (zakelijke) markt verbeterd; en voor de wetenschap wordt op deze manier aansluiting bij internationaal onderzoek gewaarborgd.

Het verbeteren van de positie van het Nederlands op het gebied van TST is het meest gebaat bij Nederlandstalige TST-producten en bij TST-onderzoek gericht op het Nederlands. Voor het maken van dergelijke producten en voor het doen van TST-onderzoek zijn evenwel een aantal hulpmiddelen nodig. Een *infrastructuur* voor TST heeft in de eerste plaats als doel de beschikbaarheid van deze hulpmiddelen te verbeteren, en daarmee een bijdrage te leveren aan de primaire doelstelling (TST-producten en TST-onderzoek gericht op het Nederlands).

Een infrastructuur voor TST ontstaat niet alleen door de ontwikkeling van hulpmiddelen te stimuleren. Naast ontwikkeling moet er aandacht zijn voor onderhoud, ondersteuning, en de wijze en voorwaarden waarop materiaal beschikbaar wordt gesteld. Om te garanderen dat hulpmiddelen nuttig zijn voor productontwikkeling, toegepast en fundamenteel onderzoek, is bovendien overleg nodig tussen researchafdelingen, wetenschappelijke instituten, en de overheid.

Een kwalitatief goede infrastructuur voor TST voor het Nederlands betekent dat alle corpora, woordenboeken, software-modules, en andere zaken die als hulpmiddelen kunnen worden ingezet bij het ontwikkelen van TST-producten en bij wetenschappelijk onderzoek in dit gebied, beschikbaar zijn en worden onderhouden. Een dergelijke infrastructuur zal bijdragen tot een verbetering van de positie van het Nederlands binnen TST. Meer in het bijzonder mag van een goede infrastructuur worden verwacht dat ze zal leiden tot:

- **kwaliteitsverbetering** van Nederlandstalige TST-producten, doordat gebruik kan worden gemaakt van de juiste hulpmiddelen en er duidelijke evaluatiecriteria beschikbaar zijn,
- **kostenbesparing** in onderzoek, research, en productontwikkeling, doordat hulpmiddelen beschikbaar zijn of voor gezamenlijk gebruik ontwikkeld kunnen worden, en niet door iedere instelling apart hoeven worden aangemaakt,
- **stimulering** van wetenschappelijk onderzoek, doordat de aansluiting bij internationale programma's gewaarborgd is,
- **drempelverlaging**, doordat instituten die zich bezig willen gaan houden met (aspecten van) TST gebruik kunnen maken van het werk dat reeds gedaan is,
- **synergie**, doordat het samenvoegen en op elkaar afstemmen van hulpmiddelen nieuwe toepassingen mogelijk maakt,
- **aandacht voor onderhoud**, doordat het ontwikkelen van hulpmiddelen niet langer de verantwoordelijkheid is van individuele instellingen,

Een infrastructuur voor TST is per definitie een zaak die alle actoren die actief zijn op het gebied van TST aangaat. Dit betekent dat een infrastructuur alleen zal onstaan wanneer de verschillende actoren van elkaars wensen en mogelijkheden op de hoogte

zijn, en er voldoende mogelijkheden zijn voor samenwerking. Dit betekent onder andere dat informatieuitwisseling en regelmatig overleg vereist is. Zo'n *overlegstructuur* voor TST kan bijdragen aan de volgende doelstellingen:

- **ondersteuning van fondsenwerving**. Doordat de behoeften van de verschillende partijen duidelijk in kaart gebracht worden, en kunnen worden afgezet tegen het bestaande aanbod aan hulpmiddelen, kan men effectiever pleiten voor projecten die gericht zijn op het wegwerken van hiaten.
- **bewustmaking (awareness).** Een samenwerkingsverband voor TST waarin alle actoren vertegenwoordigd zijn kan het belang van deze technologie onder de aandacht brengen van derden, en kan potentiële afnemers van deze technologie informeren over de mogelijkheden die TST biedt.
- **een sterkere positie in Europa.** De EU speelt een beslissende rol in vele projecten op het gebied van ICT, en stimuleert de ontwikkeling van multilinguale ICT-hulpmiddelen. Een infrastructuur voor TST maakt het mogelijk deze programma's effectief om te zetten in nationale acties.
- **aandacht voor opleidingen.** Voor onderzoek en productontwikkeling zijn gekwalificeerde medewerkers nodig. Het opleiden van onderzoekers op het gebied van TST is in de eerste plaats de verantwoordelijkheid van de wetenschappelijke instituten. Een infrastructuur voor TST kan echter een bijdrage leveren aan het opzetten van nieuwe opleidingen, aan het onderling afstemmen van de curricula en aan het afstemmen van opleiding op de wensen van het afnemend veld.

1.5 Een minimale infrastructuur

Een minimale infrastructuur voor TST dient er in ieder geval zorg voor te dragen dat die hulpmiddelen voor TST die van belang zijn voor ieder van de betrokkenen, beschikbaar zijn. Het gaat hierbij in de eerste plaats om hulpmiddelen die niet, of slechts met grote inspanning, door de individuele instituten kunnen worden geproduceerd, en om hulpmiddelen die alleen na overleg tot stand kunnen komen. Hierbij kan men denken aan:

- grote **elektronische woordenboeken** met lexicale, fonologische, combinatorische, morfosyntactische, syntactische en semantische informatie,
- omvangrijke **corpora**, zowel tekst als spraak, geannoteerd als ruw,
- **modules** voor TST, zoals bijvoorbeeld een algemene, corpus-gebaseerde, computationele grammatica die kan dienen als basis voor de ontwikkeling van applicatie-specifieke computationele grammatica's,[5]
- **standaards**, zoals annotatierichtlijnen voor spraak- en tekstcorpora,

[5] In het verslag van de interviews (hoofdstuk 4) wordt een dergelijke grammatica omschreven als de 'Groene Grammatica', naar analogie met het Groene Boekje, om te benadrukken dat het hier om een algemeen aanvaard hulpmiddel gaat dat bij voorkeur in het publieke domein beschikbaar is.

- **evaluatiehulpmiddelen** (*benchmarks, testsuites*) waarmee de prestaties van spraakherkenners, computationele grammatica's, etc. gemeten kunnen worden.

Om te garanderen dat bestaande hulpmiddelen daadwerkelijk aangeschaft en gebruikt kunnen worden is een minimale vorm van informatievoorziening en onderhoud nodig:

- Een **internet-pagina** waar bestaande hulpmiddelen worden beschreven, en waar men deze hulpmiddelen kan verkrijgen, dan wel waar men wordt doorverwezen naar de instantie die verantwoordelijk is voor het beheer.

- Aandacht voor het **onderhoud van hulpmiddelen**. Daar waar mogelijk wordt er op toegezien dat hulpmiddelen die van algemeen nut zijn beschikbaar worden gemaakt en beschikbaar blijven (ook na afloop van een project) op een manier die zoveel mogelijk partijen in staat stelt hiervan gebruik te maken.

Het ontwikkelen van elektronische woordenboeken, corpora, en hulpmiddelen als (algemene) computationele grammatica's gaat de (financiële) mogelijkheden van individuele instellingen te boven of is vanuit bedrijfseconomisch oogpunt gezien niet lonend voor een taal als het Nederlands. Daarnaast is het nut van ieder van deze hulpmiddelen beperkt zolang men over deze middelen in isolatie beschikt: de ontwikkeling van woordenboeken veronderstelt de beschikbaarheid van corpora, de ontwikkeling van een algemene computationele grammatica veronderstelt de beschikbaarheid van corpora en woordenboeken, en het toepassen van deze hulpmiddelen in een praktische applicatie veronderstelt weer dat de gegevens uit verschillende (algemene en applicatiespecifieke) corpora gecombineerd worden en dat verschillende modules (woordenboek of morfologische analyse en grammatica, grammatica en spraakherkenner, etc.) gekoppeld kunnen worden.

Het ontwikkelen van standaards en evaluatiehulpmiddelen is bijna per definitie een kwestie die niet door individuele instellingen ondernomen kan worden. Zowel standaards als evaluatiemethoden dienen te worden gevalideerd door een onderzoeksgemeenschap als geheel, en dienen bij voorkeur door zoveel mogelijk groepen toegepast te worden.

Naast aandacht voor hulpmiddelen dient er daarom een *minimale overlegstructuur* te zijn. Hierbij kan men bijvoorbeeld denken aan:

- een **overzicht** (in de vorm van een website of nieuwsbrief) van lopende of nieuw op te starten programma's, waarin een systematisch overzicht wordt gegeven van subsidiemogelijkheden van de nationale overheid en de EU,

- een **overlegorgaan** (waarin bijvoorbeeld o.a. NWO, IWT, en de Taalunie vertegenwoordigd zijn), dat zich richt op het bevorderen van samenwerking op TST-gebied, ook tussen Nederland en Vlaanderen, en dat erop toeziet dat beschikbare fondsen zo effectief mogelijk worden ingezet (waarbij criteria als het vermijden van doublures en het opvullen van hiaten, en de beschikbaarheid na afloop van een project een rol kunnen spelen),

- overeenstemming over een **curriculum** voor taal- en spraaktechnologie.

1.6 Een ideale infrastructuur

Een ideale infrastructuur voor TST is een uitbreiding van de minimale infrastructuur. In het ideale geval zijn alle hulpmiddelen die van belang zijn voor TST aanwezig en voor alle partijen beschikbaar. De lijst van hulpmiddelen die men tot de infrastructuur kan rekenen is omvangrijk. Zo kunnen bijvoorbeeld zeer verschillende soorten corpora worden ingezet (spraak of tekst, algemeen of domeinspecifiek, ongeannoteerd of geannoteerd met fonetische en fonologische informatie, woordsoort, of constituentstructuur, gekoppeld aan een woordenboek of niet, spontane spraak of voorgelezen tekst, multilinguaal, parallel, met spreek-, schrijf- of spelfouten, enz.), en kan een woordenboek zeer diverse vormen van informatie bevatten (spelling, uitspraak, syntactische categorie, betekenis, afbreekstreepjes, informatie over vaste verbindingen en idiomatische uitdrukkingen, enz.). Een uitgebreid overzicht van de componenten die deel uitmaken van een ideale infrastructuur voor TST, en die tevens dient als uitgangspunt voor de inventarisatie in hoofdstuk 2, wordt gegeven in de volgende sectie.

In een ideale infrastructuur is er ook aandacht voor organisatorische aspecten die het onderhoud van de infrastructuur betreffen, voor fundamenteel, lange-termijn, wetenschappelijk onderzoek, en voor onderwijs. Meer in het bijzonder mag men verwachten dat er in het ideale geval sprake is van:

1. **structurele aandacht voor hulpmiddelen**:
    - Producenten en afnemers van hulpmiddelen voor TST zijn van elkaars activiteiten op de hoogte en plegen regelmatig overleg,
    - Bij het ontwikkelen van nieuwe hulpmiddelen is er aandacht voor de wensen van het TST-veld als geheel, voor standaards, en voor de juridische aspecten,
    - Alle hulpmiddelen zijn onder duidelijke en redelijke voorwaarden beschikbaar (waarbij gedacht moet worden aan financiële vergoedingen voor het gebruik en mogelijkheden tot hergebruik in TST-producten),
    - er is een structurele voorziening voor het onderhoud van de bestaande hulpmiddelen,
    - de productie van hulpmiddelen wordt zonodig gestimuleerd door subsidies.
2. een **hoogwaardig educatief netwerk**:
    - Er bestaan voldoende en kwalitatief hoogstaande opleidingen die onderzoekers afleveren die actief kunnen worden op het gebied van TST. Te denken valt aan opleidingen op het gebied van de algemene taalwetenschap (met name formele taalkunde), computationele taalkunde, fonologie en fonetiek, informatica en bepaalde opleidingen op het gebied van de toegepaste informatica.[6]

[6] In Vlaanderen bestaan er in de basisopleiding slechts algemene talenopleidingen, en geen specialismen zoals in Nederland. Er zijn wel korte post-doc opleidingen. In Nederland is de situatie beter, maar zeker niet ideaal. De variatie in het onderwijsaanbod tussen unviersiteiten onderling is bijvoorbeeld erg groot. Dit komt de herkenbaarheid van de opleidingen niet ten goede (zie bv. het visitatierapport *Experimentele Letteren*).

3. een **hoogwaardig en efficiënt wetenschappelijk netwerk**:

    - Wetenschappelijk onderzoek richt zich op het uitvoeren van innovatief onderzoek met als doel de hiaten in het kennisdomein op te vullen en zodoende nieuwe toepassingsmogelijkheden te ontsluiten. Het onderzoek maakt gebruik van onderzoeksresultaten van derden en geschiedt in samenwerking met andere onderzoekers en onderzoekscentra.
    - Er bestaan mogelijkheden voor de uitwisseling van onderzoeksresultaten op TST-gebied. De uitwisseling vindt plaats in de vorm van conferenties, workshops, en tijdschriften.
    - Universitaire onderzoekscentra en de onderzoekers in het bedrijfsleven werken samen en wisselen ideeën uit. Samenwerking kan bijvoorbeeld worden gerealiseerd in expertisecentra en in (door derden gefinancierde) projecten.
    - Onderzoekers op het gebied van de taaltechnologie en op het gebied van de spraaktechnologie onderhouden regelmatige contacten. Hierbij is het van belang te streven naar een daadwerkelijke onderzoeksgemeenschap voor taal- *en* spraaktechnologie. In de huidige situatie is dit nauwelijks het geval. Taaltechnologie is vaak een onderdeel van de computationele taalkunde (de linguïstische benadering), terwijl spraaktechnologie deel uitmaakt van de fonetiek (de ingenieursbenadering).

1.7 Componenten van een infrastructuur voor TST

In deze sectie geven we een overzicht van de hulpmiddelen die in het ideale geval deel uitmaken van de infrastructuur voor TST. Twee duidelijk te onderscheiden onderdelen van de materiële infrastructuur zijn corpora en lexica. Daarnaast kunnen verschillende andere producten, in de sfeer van software-modules, *tools*, halffabrikaten, evaluatiehulpmiddelen en standaards, tot de materiële infrastructuur gerekend worden. In een ideale infrastructuur staat dit alles ter beschikking van onderzoekers en ontwikkelaars.

Corpora

Een corpus is een verzameling tekst of gesproken taal. Corpora worden typisch gebruikt voor het trainen, testen, en evalueren van TST-programma's (zie hoofdstuk 3). Corpora kunnen verschillen in de aard van de data (tekst of spraak), tekst-soort (alleen literaire of journalistieke teksten, of een mix van tekst-soorten, spraak in formele en informele contexten), omvang (een historisch corpus kan de omvang van enkele duizenden woorden hebben, andere corpora hebben een omvang van vele miljoenen woorden), en aard van de annotaties (geen annotaties of uitgebreide syntactische en semantische annotatie, al dan niet gecontroleerd, en alles daartussen). Hieronder noemen we enkele soorten corpora die gangbaar zijn in TST-onderzoek.

- **Ruwe tekst.** Corpora waaraan geen enkele vorm van taalkundige informatie is toegevoegd kunnen informatie leveren over de frequentie van woorden, woordcombinaties (*bigram- en trigramstatistieken*), letters en letter-combinaties, gemiddelde woord- en zinslengte, etc. Ruwe corpora zijn (met de huidige overvloed aan elektronische documenten) relatief eenvoudig samen te stellen. Het samenstellen van grote corpora (met een omvang van tientallen miljoenen woorden), van corpora die gebalanceerd en representatief zijn, en het samenstellen van corpora die ter beschikking kunnen worden gesteld aan derden, vereist veel zorg een aandacht.

- **Woorden met woordsoort.** Een corpus waarin van ieder woord de woordsoort is aangegeven kan worden gebruikt om informatie over de frequentie van woordsoorten, van woordsoort per woord, en van combinaties van woordsoorten te verzamelen. Daarnaast kan het gebruikt worden om programma's die ontleden op woordsoort (*taggers*) te trainen en te evalueren. Ook bij de constructie van robuuste ontleders (programma's die willekeurige tekst syntactisch ontleden) wordt soms gebruik gemaakt van een corpus met woordsoorten. Het construeren van een dergelijk corpus is arbeidsintensief, omdat het handmatig of semiautomatisch (automatische *tagging* met handmatige correctie) uitgevoerd moet worden.

- **Woorden met concepten.** Woorden zijn vaak ambigu, en desambiguatie is een belangrijk probleem voor automatisch tekstbegrip en voor *information retrieval*. Een corpus waarin voor ieder woord is aangegeven in welke betekenis het gebruikt is, kan worden gebruikt om het probleem van (woord-)desambiguatie op basis van data te onderzoeken.

- **Zinnen met constituent-structuur.** Van iedere zin wordt aangegeven wat de zinsdelen zijn, en wat de syntactische categorie van die zinsdelen is. Dergelijke corpora worden gebruikt om de prestaties van automatische ontleders te meten en om probabilistische grammatica's af te leiden.

- **Zinnen met semantische annotaties.** Het is mogelijk zinnen te annoteren met woordbetekenissen (of concepten), de zinsbetekenis, en de betekenis die relevant is voor een bepaalde toepassing. Daarnaast kan men aandacht besteden aan aspecten van dialoog-voering (taaldaden) en contextuele interpretatie (zinsgebonden en niet-zinsgebonden anaforische en deictische relaties). Het semantisch annoteren van teksten is een nog zeer pril onderdeel van de corpustaalkunde. Standaards ontbreken vooralsnog.

- **Parallelle corpora.** In het onderzoek naar automatisch vertalen maakt men gebruik van parallelle corpora: corpora waarin dezelfde tekst in meerdere talen voorkomt. Dergelijke corpora worden gebruikt bij de constructie van multilinguale lexica en in corpus-gebaseerd onderzoek naar automatisch vertalen.

- **Parallel variantencorpus.** Een parallel corpus Vlaams-Nederlands kan inzicht verschaffen in de verschillen tussen het Nederlands en het Vlaams. Standaards ontbreken ook hier vooralsnog.

- **Gesproken tekst.** Geluidsfragmenten van gesproken taal voorzien van annotatie. De annotatie maakt het mogelijk specifieke delen van de opname te verbinden met klanken of woorden. De mate van detail waarin dit gebeurt kan verschillen. Dergelijke corpora spelen een belangrijke rol bij het ontwikkelen van uitspraakmodellen (voor spraaksynthese) en bij het trainen van spraakherkenners.

Een **suite** is een verzameling data die zowel overeenkomsten als verschillen vertoont met een corpus. Zo'n suite is samengesteld door een taalkundige, met als doel relevante (positieve en negatieve) voorbeelden van een bepaald fenomeen bijeen te brengen. Suites worden bijvoorbeeld gebruikt om de reikwijdte (*coverage*) van een grammatica te bepalen, of om de kwaliteit van bestaande grammatica's te bewaken: met behulp van een suite kan eenvoudig worden vastgesteld of veranderingen in de grammatica geen onvoorziene fouten introduceren. In tegenstelling tot een corpus is een suite meestal niet een verzameling ruwe data, maar een verzameling data die zorgvuldig is samengesteld met een bepaalde doelstelling voor ogen. Een suite is daarom in de eerste plaats bedoeld om informatie te krijgen over een taalkundig programma en niet om (statistische) informatie over de taal zelf te verkrijgen.

Lexica

Een *lexicon* of *woordenboek* is een woordenlijst met additionele informatie. De informatie die voor ieder woord wordt gegeven hangt sterk af van het doel waarvoor het woordenboek is ontworpen. Wanneer erg veel verschillende informatie wordt opgenomen, en het woordenboek in functie duidelijk verschilt van hetgeen normaal gesproken in een (gedrukt) woordenboek wordt aangetroffen, spreekt men ook wel van een lexicale database (bijvoorbeeld de CELEX-database (Baayen, Piepenbrock, en van Rijn 1993)).

Een lexicale database kan onder andere de volgende informatie bevatten:

- **De woordenlijst zelf.** Levert informatie over de grootte van de woordenschat en is nuttig voor spellingcorrectie.
- **Combinatorische informatie.** Aanduidingen van idiomen, vaste uitdrukkingen, collocaties, enz. zijn waardevol voor spellingcorrectie, automatisch vertalen, etc.
- **Afbreekstreepjes.** Nuttig voor de ontwikkeling van afbreekroutines.
- **Morfologische informatie.** Van ieder verbogen woord is de stam bekend en de morfologische structuur. Bijvoorbeeld van belang voor programma's die morfologische analyse uitvoeren en voor *lemmatizers* (programma's die een woord herleiden tot een stamwoord en die o.a. worden gebruikt in *information retrieval*).
- **Syntactische informatie.** Bijvoorbeeld woordsoort, geslacht (van zelfstandige naamwoorden), valentie (m.n. voor werkwoorden). Deze informatie is noodzakelijk voor het ontwikkelen van niet-triviale grammatica's.
- **Semantische en conceptuele informatie.** Woordbetekenissen, synoniemen, conceptuele verbanden. Van belang voor desambiguatie en toepassingen waarbij lexicale ambiguïteit een struikelblok is, zoals automatisch vertalen.

- **Fonologische en prosodische informatie.** Voor spraakherkenning, systemen die tekst naar spraak omzetten, en andere spraaktechnologische producten.
- **Frequentie-informatie.** Voor ontleden op woordsoort (*part-of-speech tagging*), optische lezers (*optical character recognition*), handschriftherkenning, etc.
- **Vertalingen.** Een tweetalig woordenboek is een bouwsteen voor automatische vertaalsystemen.

Een **concordantie** is een product dat eigenschappen van een lexicon en een corpus combineert. Het is een woordenlijst van een bepaald corpus, met verwijzingen naar de plaatsen waar het woord voorkomt. Tegenwoordig worden concordanties meestal niet als aparte producten beschouwd, maar als producten die automatisch afgeleid kunnen worden uit een bepaald corpus.

Overige hulpmiddelen

In deze sectie beschrijven we verschillende producten die, naast corpora en lexica, deel uit moeten maken van de TST-infrastructuur.

- **Corpus-hulpmiddelen.** Om corpora te kunnen exploreren kunnen verschillende computationele hulpmiddelen worden gebruikt. De meeste onderzoekers gebruiken de mogelijkheden die het operating system UNIX biedt, in combinatie met *script*-talen als AWK en PERL. Het gebruik van deze omgevingen vereist kennis van computers die niet bij iedere taalkundige voorhanden is. Om die reden, en omdat het handig is, worden er ook wel verzamelingen scripts aangelegd die de meest gangbare taken verrichten (zoeken, het maken van een woordenlijst, van een frequentielijst, van bigram- en trigramstatistieken, van concordanties of *keyword-in-context* overzichten, etc.). Veel gebruikte softwarepakketten zijn bijvoorbeeld Tact en Wordsmith. Tegenwoordig wordt soms ook het *world wide web* gebruikt om een corpus toegankelijk te maken.
- **Spraakherkenners.** Programma's voor het omzetten van geluidssignalen naar hun geschreven equivalent. Het kan hierbij gaan om het herkennen van losse woorden, maar steeds vaker probeert men ook vloeiend spraakgebruik (zogenaamde *connected speech*) juist te herkennen. Sommige systemen werken met een microfoon, andere systemen werken ook telefonisch (waarbij bedacht moet worden dat de geluidskwaliteit in het laatste geval veel minder is). Sommige systemen (met name dicteersystemen) zijn sprekerafhankelijk (dit betekent dat een gebruiker eerst een aantal oefenteksten moet inspreken zodat het systeem getraind kan worden), andere systemen zijn sprekeronafhankelijk.
- **Grafeem-naar-foneem-omzetters.** Programma's voor het omzetten van tekst in een reeks fonetische symbolen. Dit is bijvoorbeeld van belang voor het omzetten van willekeurige tekst in spraak.
- **Part-of-speech taggers.** Een POS *tagger* kent automatisch woordsoorten toe aan de woorden in een tekst.

- **Standaards.** Wanneer een corpus wordt voorzien van annotatie, of wanneer in een woordenboek syntactische categorieën of uitspraakgegevens worden vastgelegd, maakt men bij voorkeur gebruik van een notatie die niet specifiek voor dit corpus of voor dit woordenboek is, maar van een algemeen gangbare notatie. Het volgen van standaards maakt de kans op hergebruik groter, en maakt het ook eenvoudiger verschillende corpora of woordenboeken te combineren, en het maakt het ontwikkelen van *tools* die gebruik maken van deze data eenvoudiger en waardevoller.

- **Morfologische ontleders.** In IR wordt veel gebruik gemaakt van *lemmatizers*, programma's die een verbogen woord terug brengen tot de stam. Voor taalkundige toepassingen is het vaak ook van belang informatie te krijgen over de morfologische structuur en kenmerken van een verbogen woord.

- **Syntactische ontleders.** Programma's die grote hoeveelheden tekst efficiënt van een syntactische structuur kunnen voorzien.

1.8 Juridische aspecten

Zowel voor de makers als voor de afnemers van TST-materialen is het belangrijk dat de auteursrechtelijke kwesties helder geregeld zijn. Het is het belangrijk het juiste evenwicht te creëren tussen de rechten van de makers enerzijds en optimale openbaarheid en toegankelijkheid van de informatie voor het publiek anderzijds. Hieronder zetten we eerst de belangrijkste begrippen met betrekking tot het auteursrecht op een rij.

Aan elk werk van letterkunde, wetenschap of kunst is van rechtswege vanaf zijn ontstaan een exclusief beschikkingsrecht voor de maker verbonden. Dit exclusieve beschikkingsrecht omvat onder andere de rechten om te beslissen over openbaarmaking en verveelvoudiging (exploitatierechten), om als de maker te worden aangemerkt en om te beslissen over wijzigingen in het werk (persoonlijkheidsrechten). Voor auteursrechtelijke bescherming moet het betreffende werk wel een voldoende oorspronkelijk karakter hebben. De exploitatierechten zijn overdraagbaar. Persoonlijkheidsrechten zijn in principe niet overdraagbaar, de rechthebbende kan wel te kennen geven dat hij zich niet op bedoeld recht zal beroepen. De rechthebbende kan de aan hem voorbehouden handelingen met betrekking tot het exploiteren van auteursrechten zelf verrichten of hij kan het exploiteren van zijn werk overlaten aan een of meer anderen door zijn rechten aan die anderen over te dragen ('verkopen') of door hen een licentie te geven. Door overdracht van het auteursrecht komt de zeggenschap van het werk in handen van degene aan wie het wordt overgedragen. Met het verlenen van een licentie wordt aan een niet-rechthebbende de toestemming gegeven om het auteursrechtelijk beschermde werk op een of andere wijze openbaar te maken of te verveelvoudigen. Het auteursrecht zelf blijft in handen van de licentiegever. In geval van niet-exclusieve licentie blijft de auteursrechthebbende ook zelf gerechtigd tot het zelfstandig exploiteren van het werk en kan hij ook aan anderen een soortgelijke toestemming geven. Bij een exclusieve licentie verplicht hij zich tegenover de licentienemer om aan anderen geen soortgelijke toestemming te verlenen. De weg tussen aanbieder (maker/producent) en gebruiker is bij digitale, met name *on-line* media, veel korter dan bij gedrukte publicaties. De toegang tot de digitale producten en het gebruik van de data wordt meestal

geregeld in licentieovereenkomsten tussen afnemer en aanbieder. De aanbieder kan de voorwaarden voor toegang en gebruik in principe naar believen per afnemer aanpassen. Zo kan men aan onderzoeksinstellingen een niet-overdraagbare, niet-exclusieve licentie geven voor gebruik van de data in de eigen onderzoeksgroep; een commerciële gebruiker kan een niet-exclusieve licentie krijgen om het materiaal te gebruiken (en aan te passen) voor duidelijk afgesproken commerciële toepassingen, bijvoorbeeld om het materiaal te distribueren als deel van een eigen, nieuw product. De rechten en voorwaarden moeten duidelijk in de licentie-overeenkomst worden omschreven. Voor de verschillende gebruiksrechten kunnen verschillende tarieven worden gehanteerd. Zo hanteert ELRA (de *European Language Resources Association*) verschillende conceptovereenkomsten tussen ELRA en resp. aanbieder, eindgebruiker en VAR (= *value added resaler*)[7].

Elektronische raadpleging van gedigitaliseerde informatie gaat per definitie gepaard met kopiëren (verveelvoudigen). De moderne technologie maakt het mogelijk snel en zonder kwaliteitsverlies op voordelige wijze identieke kopieën te maken. Het is daarom van het grootste belang dat men niet alleen nationaal maar ook op internationaal vlak tot een akkoord komt over de beschermingsomvang van het auteursrecht in de digitale omgeving. De internationale wetgeving op dit gebied is in volle beweging. De Softwarerichtlijn en de Databankenrichtlijn werden in resp. mei 1991 en maart 1996 goedgekeurd door de Raad van Ministers van de Europese Unie. In december 1996 werden twee nieuwe WIPO-verdragen (verdragen van de *World Intellectual Property Organisation*) goedgekeurd en in december 1997 verscheen het voorstel van de Europese Commissie voor een Richtlijn Auteursrecht en naburige rechten in de Informatiemaatschappij. In de komende jaren moet de betrokken nationale staten hun eigen, nationale wetgeving aanpassen aan de wetgeving in deze internationale verdragen/richtlijnen.

De databankrichtlijn omschrijft een databank als 'een verzameling van werken, gegevens of andere zelfstandige elementen, die systematisch of methodisch zijn geordend en afzonderlijk toegankelijk zijn'. Computerprogramma's die worden gebruikt voor de totstandbrenging of de werking van de databank vallen niet onder deze term. De auteursrechtelijke bescherming van 'oorspronkelijke' computerprogramma's is vastgelegd in de Softwarerichtlijn. De databankrichtlijn voorziet in een vijftien jaar durende wettelijke bescherming van databanken. Naast de inhoud van de databanken wordt ook het copyright op de structuur ervan beschermd. Het extractierecht, een nieuw recht, verleent de maker een 'recht op verhindering van onrechtmatige opvraging', waardoor hij kan verhinderen dat gegevens uit de databank wordt opgevraagd en (voor commerciële doeleinden) hergebruikt.

De meeste landen in de wereld, zijn aangesloten bij de Berner Conventie (BC), het belangrijkste multilaterale verdrag op het gebied van het internationale auteursrecht. De BC garandeert de auteursrechtelijke bescherming over de grenzen heen. Sinds 1967 wordt de Berner Conventie behartigd door de World Intellectual Property Organisation (WIPO), een organisatie van de VN gevestigd in Genève. WIPO-verdragen vormen de basis voor de auteurswetgeving in de aangesloten landen. In de nieuwe WIPO-verdragen (1996) wordt de toepasselijkheid van het (bestaande) auteursrecht voor de nieuwe technologie principieel voor alle landen bevestigd. De transmissie van een

[7] http://www.icp.grenet.fr/ELRA/legals.html

werk van punt naar punt wordt als een openbaarmaking erkend. Het opslaan van een beschermd werk in digitale vorm wordt beschouwd als een verveelvoudiging. Het echte downloaden van een werk vormt een traditionele verveelvoudiging waarop de auteur zijn verbodsrecht kan uitoefenen.

Het doel van de Richtlijn Auteursrecht en naburige rechten in de Informatiemaatschappij is het harmoniseren van de nationale wetgevingen binnen de Europese Unie. Het voorstel zou vóór 1 juli 2000 door de Europese landen in hun wetten moeten zijn geïmplementeerd. De richtlijn kent auteurs het exclusieve recht toe op openbaarmaken van hun werk, onafhankelijk van het aantal keren dat het werk *on-line* wordt geraadpleegd. Onder reproductie wordt elke relevante handeling van directe of indirecte verveelvoudiging, tijdelijk of permanent, *on-line* of *off-line* verstaan. Iedere, ook de zeer tijdelijke vastlegging, valt onder het auteursrechtelijk verbodsrecht. Er worden een aantal uitzonderingen gemaakt zoals voor bepaalde reproductiehandelingen gedicteerd door de technologie maar zonder eigen economische waarde (bv. bepaalde vormen van browsing en *cache-copies* die optreden bij transmissie over internet). Voor openbaarmaken en verveelvoudigen gelden een aantal uitzonderingen waaronder gebruik voor onderwijs en wetenschappelijk onderzoek. Bibliotheken worden uitgezonderd van het reproductierecht maar niet van het recht op openbaarmaking. Met betrekking tot het distributierecht stelt de richtlijn dat eens een auteur de toestemming heeft gegeven tot verkoop van zijn werk in een lidstaat, deze toestemming voor alle lidstaten van de Europese Unie geldt. Parallelimport is verboden.

Voor een succesvolle exploitatie van TST-materialen moet duidelijk zijn waar de rechten op die materialen liggen. Bij bestaande materialen vereist dit vaak een hele zoektocht, de rechten blijken in veel gevallen niet eenduidig vastgelegd te zijn. Dit geldt met name bij projecten die in samenwerking tussen wetenschap en industrie tot stand kwamen. Een van de oorzaken is dat bij het creëren van TST-materialen meestal vele 'makers' betrokken zijn. In een aantal gevallen is het gezien de werkrelaties onduidelijk aan wie de auteursrechten toekomen. De feitelijke maker is bijvoorbeeld niet de juridische maker als het werk volgens een nauw omschreven opdracht wordt gemaakt of als het werk gemaakt wordt door een werknemer die uitdrukkelijk hiervoor in dienst is genomen. De rechten komen dan niet toe aan de werkelijke maker van een werk (de werknemer-auteur), maar aan degene die door de wet als fictieve maker (de werkgever) wordt beschouwd.[8] Een groot aantal potentiële auteursrechthebbenden kan de verdere exploitatie bemoeilijken: het voeren van onderhandelingen met en het contracteren van de vele rechthebbenden schept vaak grote logistieke problemen. Bij het ontwikkelen van nieuwe materialen is het daarom belangrijk al bij de start van een project bindende afspraken te maken met alle eventuele rechthebbenden zodat de materialen zonder ongewenste auteursrechtelijke beperkingen kunnen worden gedistribueerd. Het is in het belang van alle bij de exploitatie van het project betrokken partijen raadzaam de volledige auteursrechten op alle mogelijke exploitaties zoveel mogelijk in één rechtspersoon te concentreren.

Het auteursrecht schept overigens de bevoegdheid, niet de plicht het auteursrecht uit te oefenen. De maker kan besluiten zich niet op bepaalde rechten (exploitatierechten of persoonlijkheidsrechten) te beroepen. Bij projecten die door de overheid worden gefinancierd staat over het algemeen zowel de subsidiënten als de eigenlijke 'makers' een strategisch doel voor ogen waarvan de vrije toegankelijkheid van het materiaal een van de belangrijke elementen is.

[8]Deze restrictie geldt niet voor medewerkers aan de universiteit: hun functie-omschrijvingen zijn zo globaal dat de universiteit niet kan beschikken over het auteursrecht op de producten die zij vervaardigen. Dit geldt vaak ook voor medewerkers die voor een specifiek project zijn aangesteld.

Hoofdstuk 2

De huidige situatie

In dit hoofdstuk schetsen we de huidige infrastructuur voor taal- en spraaktechnologie. We geven achtereenvolgens een overzicht van bestaande dataverzamelingen en hulpmiddelen voor het Nederlands, een kort overzicht van vergelijkbare materialen voor het Engels, van onderzoeks- en onderwijsinstellingen, van beleidsinstellingen, en van buitenlandse organisaties voor taal- en spraaktechnologie.

2.1 De materiële infrastructuur

In deze sectie geven we een zo volledig mogelijk overzicht van beschikbare hulpmiddelen voor het Nederlands. We volgen de indeling uit sectie 1.7.

De informatie die hieronder volgt is ten dele ontleend aan:

- ANNO[1]: een geannoteerde publieke gegevensbank voor het geschreven Nederlands.
- ELRA[2]: *European Language Resources Association*.
- EUROMAP: de Vlaamse en Nederlandse rapporten (Dewallef 1998, van Staden 1998).
- *Language Engineering Directory*[3].

2.1.1 Corpora

Corpora met ruwe tekst

Het samenstellen van een corpus waaraan geen annotatie is toegevoegd is tegenwoordig vrij eenvoudig. Elektronische teksten zijn in overvloed aanwezig op het WWW. Op sommige van deze teksten rust geen auteursrecht (bijvoorbeeld de handelingen van de eerste en tweede kamer in Nederland, of het lijvige rapport van de Nederlandse parlementaire onderzoekscommissie Van Traa). Het verzamelen van een corpus over een bepaald onderwerp, of met een evenwichtige vertegenwoordiging van verschillende

[1] http://www.ccl.kuleuven.ac.be/about/ANNO.html
[2] http://www.icp.grenet.fr/ELRA/home.html
[3] http://www2.echo.lu/mlis/en/atlas/atlas-intr.html

tekstsoorten, kan toch nog een tijdrovende bezigheid zijn. Literaire teksten zijn beschikbaar via het Coster-project.[4] Ook de web-pagina's van de Stichting Tekstcorpora en Databestanden in de Humaniora (STDH)[5] geven verwijzingen naar literaire corpora. Daarnaast zijn enkele corpora speciaal voor taalkundig onderzoek beschikbaar gemaakt:

- **ECI/MCI CD-Rom.**
  Omschrijving: De CD-ROM *European Corpus Initiative Multilingual Corpus I* bevat een aantal Nederlandse corpora:
  - **dut01.** "Newspaper, Dutch, 600K tokens, Articles from the student newspaper Universiteitskrant of the University of Groningen from the academic years 1990/1991 and 1991/1992."
  - **dut02.** "Mixed, Dutch, 5203K tokens, A large Dutch corpus from INL including transcripts of radio programs, newspaper and magazine issues and some technical texts."
  - **dut03.** "Mixed, Dutch, 128K, A continuation of dut02."

  Beschikbaarheid: Distributie door Elsnet[6] en het Linguistic Data Consortium.[7]
- **ELRA-W0006.**[8]
  Omschrijving: "The Polylingual Document Collection (ELRA-W0006), a collection of newspaper articles from financial newspapers in 6 languages (Dutch, English, French, German, Italian and Spanish). It consists of the following subcorpora: *Dutch - Het Financiële Dagblad - 1992-1993.* The corpus contains articles from the Dutch financial newspaper Het Financiële Dagblad editions of 2nd January 1992 through to 24th December 1993. It contains around 8.5 million words of text."
  Beschikbaarheid: Distributie door ELRA.

Corpora voorzien van woordsoort

- **De INL Corpora.**
  Omschrijving: Het Instituut voor Nederlandse Lexicografie bezit verschillende corpora (Kruyt 1995):
  - **50 Miljoen Woorden Corpus 1994.** "Algemeen Nederlands, 1970-1990, 17 boeken over gevarieerde onderwerpen (30% fictie), niet taalkundig verrijkt."
  - **15 Miljoen Woorden Corpus.** "Automatisch taalkundig verrijkt (woorsoort en lemma); deel uit 50 Miljoen Woorden Corpus. "

[4] http://www.dds.nl/~ljcoster/
[5] http://CandL.let.ruu.nl/stdh/index.htm
[6] http://www.elsnet.org/resources/
[7] http://www.ldc.upenn.edu/
[8] http://www.icp.grenet.fr/ELRA/cata/text_det.html

– **5 Miljoen Woorden Corpus 1994.** "Algemeen Nederlands, 1989-1994, 17 tekstbronnen, geclassificeerd naar publicatiemedium en onderwerp taalkundig verrijkt met woordsoort en lemma. *Er zijn nauwelijks extra correctieslagen uitgevoerd.* Dit geldt zowel voor de teksten zelf, als voor de linguïstische gegevens woordsoort en lemma. Woordsoortcodes en lemmavormen zijn automatisch toegekend. *Dit corpus heeft een andere samenstelling dan dat op de multilinguale ECI/MCI CD-ROM*."

– **27 Miljoen Woorden Corpus 1995.** "Taalkundig verrijkt met woordsoort en lemma."

– **38 Miljoen Woorden Corpus 1996.** "Gevarieerde samenstelling met 3 hoofdcomponenten: krantenteksten (1992-1995), juridische component (1814-1989), gevarieerd samengestelde component (1970-1995). Teksten geclassificeerd volgens onderwerp en publicatiemedium. Taalkundig verrijkt met lemma en twee woordsoortcategorieënstelsels: een globale en een verfijnde met subcategorisatie. 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[9] standaard. *Er zijn nauwelijks correctieslagen uitgevoerd.*"

Beschikbaarheid: De 50 miljoen en 15 miljoen corpora zijn uitsluitend voor onderzoeksdoeleinden raadpleegbaar op het INL. De 5, 27, en 38 miljoen corpora zijn voor onderzoeksdoeleinden ook raadpleegbaar via internet (telnet) door middel van een retrievalprogramma.

- **Het Eindhoven (Uit den Boogaart) corpus (Uit den Boogaart 1975, de Jong 1979).**
  Omschrijving: "*Herkomst:* Werkgroep Frequentie-Onderzoek van het Nederlands, gesubsidieerd door Z.W.O. (het Nederlandse Fonds voor Zuiver Wetenschappelijk Onderzoek, nu het N.W.O.) en de Technische Hogeschool Eindhoven (geschreven taal); Instituut voor Dialectologie, Volks- en Naamkunde van de Koninklijke Nederlandse Academie voor Wetenschappen te Amsterdam (gesproken taal). *Inhoud:* Geschreven en (getranscribeerd) gesproken Nederlands, respectievelijk uit de periodes 1964-1971 en 1960-1973. *Omvang:* Geschreven taal: plm. 600.000 woorden; gesproken taal: plm. 120.000 woorden. *Codering:* voornamelijk morfo-syntactisch (woordsoort en flexievorm)."
  Beschikbaarheid: "Op verschillende instituten 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."

- **ANNO**[10]: **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.

[9] http://www.ids-mannheim.de/ldv/mecolb.html
[10] http://www.ccl.kuleuven.ac.be/about/ANNO.html

Het corpus bestaat uit de tekst van BRTN-radio nieuwsuitzendingen en Actueel uitzendingen. De transcripties van interviews binnen die uitzendingen betreft spontane uitingen. Het corpus heeft een omvang van in totaal ruim 640.000 woorden. Het gehele corpus is voorzien van morfosyntactische en fonetische annotaties. Deze annotatie is automatisch aangebracht (m.b.v. de WOTAN-tagger en TREETALK, een door W. Daelemans beschikbaar gesteld programma voor grafeem-naar-foneem conversie), en *deels* gecorrigeerd.
  Beschikbaarheid: Het corpus zal beschikbaar gesteld worden zodra de auteursrechtelijke kwesties zijn geregeld.

Rijker geannoteerde corpora

Er zijn geen corpora voor het Nederlands voorzien van constituentstructuur of semantische annotatie.

Parallelle corpora

- **ELRA-W0007.**[11]
  Omschrijving: "A Multilingual Parallel Corpus consisting of translated data in nine European languages: Danish, Dutch, English, French, German, Greek, Italian, Portuguese and Spanish. The parallel data, provided by the European Commission, comprises two sub-corpora from the Official Journal of the European Communities:

  – *Official Journal of the European Commission, C Series: Written Questions 1993.* This corpus contains written questions asked by members of the European Parliament and corresponding answers from the European Commission in 9 parallel versions. The total size of the corpus is approximately 10.2 million words (ca. 1.1 million words per language).

  – *Official Journal of the European Commission, Annex: Debates of the European Parliament 1992-1994.* The Parliamentary Debates are a record of what was said by members of the meeting as well as written input provided to the meeting. The original data from which the translations are produced consist of a transcript of the sittings, each member speaking in the language of his choice. The final version consists of nine parallel versions of the material. This sub-corpus contains some 5 to 8 million words per language."

  Beschikbaarheid: Distributie door ELRA.

Gesproken tekst

Binnen de spraaktechnologie spelen corpora reeds langere tijd een belangrijke rol. Er zijn dan ook verschillende corpora voor spraak in omloop, en aan een omvangrijk nieuw corpus wordt gewerkt. We noemen hieronder de voor spraaktechnologie belangrijkste corpora. Bij de hierboven genoemde bronnen zijn verwijzingen naar nog een aantal andere corpora te vinden.

[11] http://www.icp.grenet.fr/ELRA/cata/text_det.html

- **Eindhoven (Uit den Boogaart) Corpus.**
  Omschrijving: Zie boven.

- **Polyphone-NL (SPEX)[12].**
  Omschrijving: Dit is een geheel van gedirigeerde spraak van 5000 telefoonsprekers. Iedereen kon een vrij antwoord formuleren op 17 vaste vragen:

  "The Dutch Polyphone corpus contains telephone speech from 5050 speakers. The corpus comprises 222,075 speech files, which all have been orthographically transcribed. The data were collected directly off an ISDN telephone line interface. The corpus contains both read and extemporaneous items. Items to be read consist of isolated digits, numbers a postal code, guilder amounts, time, date, amounts, application words, sentences with application words, phonetically rich sentences, spelled words, city names. Several questions were asked to get the spontaneous part of the speech."
  Beschikbaarheid: Distributie door ELRA.

- **CHILDES (CHIld Language Data Exchange System)**.
  Omschrijving: Het Nederlands deel van dit corpus bevat uitsluitend kindertaal.
  Beschikbaarheid: Het corpus is verkrijgbaar op de CHILDES-site in Antwerpen[13] en bij de initiatiefnemers in Pittsburgh, Carnegie Mellon. Voor onderzoeksdoeleinden is het corpus ook toegankelijk op het Nijmeegse Max Planck Instituut[14].

- **GRONINGEN Corpus.**
  Omschrijving: Korte gelezen teksten, woorden, zinnen en klanken, meer dan 20 uur:

  "The Groningen Corpus[15] consists of 4 CD-ROMs containing over 20 hours of speech. It is a corpus of read speech material in Dutch, recorded on PCM tape under fairly good conditions. These 4 CD-ROMs contain speech from 238 speakers who read: 2 short texts, 23 short sentences (containing all possible vowels and all possible consonants and consonant clusters in Dutch), 20 numbers, 16 monosyllabic words (containing all possible vowels in Dutch), and 3 long vowels. The production on CD-ROM was partially supported by ELSNET and the pre-mastering was done at LIMSI-CNRS."
  Beschikbaarheid: Distributie door ELRA.

- **EUROM1 (The multilingual European speech database).**
  Omschrijving: "The first really multilingual speech database produced in Europe. Equivalent corpora for each of the European languages: same number of speakers selected in the same way, and recorded in the same conditions with common file formats. The content consists of Numbers, Passages, Sentences and CVC. More than sixty speakers per language."
  Beschikbaarheid: Distributie door ELRA.

[12] `http://iris1.let.kun.nl/spex/`
[13] `http://atila-www.uia.ac.be/childes/`
[14] `http://www.mpi.nl/world/index.html`
[15] `http://www.elsnet.org/resources/`

- **COGEN.**
  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:

  – WL-OFF (*word list office*), een corpus van gespelde woorden, commandowoorden, cijfers, en fonetisch rijke woorden, gelezen door in total 174 sprekers, opgenomen in een knatooromgeving (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.
  – RS-OFF (*read speech office*, een corpus van voorgelezen tekstfragmenten (5 paragrafen), door 174 sprekers, in kantooromgeving. Totale duur: 7.02 uur.
  – WL-TEL (*word list telephone*), een corpus van voorgelezen woordenlijsten, opgenomen via een telefoonverbinding, opgenomen voor 185 sprekers. Duur: 5.85 uur.
  – SS-TEL (*spontaneous speech telephone*), een corpus van spontane uitingen, opgenomen via een telefoonverbinding, opgenomen voor 126 sprekers. Duur: 2 uur.

  Beschikbaarheid: Geen gegevens.

- **Corpus Gesproken Nederlands.**
  Omschrijving: NWO, IWT, en een aantal andere partners bereiden momenteel een project voor (begroot op 5 miljoen ECU, looptijd 5 jaar) dat tot doel heeft een corpus van plusminus 10 miljoen woorden samen te stellen. Het gehele corpus zal worden voorzien van orthografische transcriptie, woordsoorten, morfologische analyse, en lexicologische koppeling (lemmatisering). Een kerncorpus van plusminus één miljoen woorden zal bovendien worden voorzien van syntactische analyse, en fonetische en fonologische transcripties, gekoppeld aan het akoestische signaal.
  Beschikbaarheid: De Nederlandse Taalunie zal ter zijner tijd verantwoordelijk zijn voor de distributie van de corpora.

2.1.2 Lexicale informatie

Algemene woordenboeken

- **van Dale.**
  Omschrijving: Van Dale heeft het Van Dale Groot woordenboek hedendaags Nederlands en het Van Dale Groot Synoniemenwoordenboek (niet haar bekendste product, de grote Van Dale (Geerts en Heestermans 1995)) op CD-ROM beschikbaar gemaakt (in totaal ongeveer 90.000 trefwoorden en 45.000 betekenisverwante woorden). De CD-ROM is bedoeld als elektronisch woordenboek en thesaurus, en als hulpmiddel dat kan worden gebruikt in combinatie met bekende tekstverwerkers als Word en WordPerfect. Merk op dat dit niet betekent dat de

woordenboeken bij automatische spellingcorrectie of bij afbreken gebruikt kunnen worden.
Beschikbaarheid: Distributie door Van Dale.[16]

- **WNT.**
  Omschrijving: Het Woordenboek der Nederlandsche Taal, dat wordt samengesteld op het INL, is sinds 1995 op CD-ROM beschikbaar. Behalve een zeer uitgebreid woordenboek van het oudere Nederlands, bevat het woordenboek ook een grote hoeveelheid bewijsplaatsen in de vorm van (met name literaire) citaten.
  Beschikbaarheid: Distributie door AND Publishers.[17]

Woordenlijsten voor correctie en afbreken

- **SDU/Elektronisch Groene Boekje.**
  Omschrijving: Dit is de elektronische versie van het nieuwe Woordenlijst Nederlandse Taal (Woordenlijst 1996). De functionaliteit is te vergelijken met de Van Dale CD-ROM: het is vooral bedoeld om de juiste spelling van een woord op te zoeken.
  Beschikbaarheid: Distributie door de SDU[18].

- **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 Woordelijst Nederlandse taal worden toegevoegd.
  Beschikbaarheid: Distributie door de SDU[19].

- **Words-L.**
  Omschrijving: Op de web-pagina van WORDS-L worden een aantal woordenlijsten beschikbaar gesteld die kunnen worden gebruikt voor spellingcorrectie en woorden afbreken 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.
  Beschikbaarheid: Distributie via de web-pagina van WORDS-L[20].

[16] http://www.vandale.nl/current/
[17] http://www.and.nl/publishers/wnt.html
[18] http://www.sdu.nl/uitg/tgp/taal/elekgb.html
[19] http://www.sdu.nl/uitg/tgp/taal/spellings.html
[20] http://www.iaf.nl/Users/Meridian/words.htm

Woordenlijsten met taalkundige informatie

- **CELEX. Omschrijving:** CELEX (Centre for Lexical Information)[21] heeft elektronische databases ontwikkeld die verschillende types van lexicale informatie over het hedendaagse Nederlands, Engels en Duits bevatten. De Nederlandse database bevat ongeveer 400.000 woordvormen uit het hedendaagse Nederlands. Het Nederlandse deel is voornamelijk afgeleid uit het INL 50 miljoen woorden corpus. Er is nu gedetailleerde informatie beschikbaar over de orthografie (spelling), fonologie (uitspraak), morfologie (woordstructuur: flexie en derivatie), syntaxis (grammatica) en woordfrequentie. Het belangrijke aan de CELEX databases is dat alle informatie gerepresenteerd is om tegemoet te komen aan de formele en strikte voorwaarden voor computationele toepassingen.

  Beschikbaarheid: "The CELEX database is open to all academic researchers and people associated with other not-for-profit research institutes free of charge (at least until 1998). For prospective customers from abroad, we recommend the stand-alone CD-ROM version distributed by the Linguistic Data Consortium[22]."

- **EuroWordNet.**[23]
  Omschrijving: Dit project heeft als doel de ontwikkeling van een multilinguale lexicale database in de stijl van WordNet (Miller, Beckwith, Fellbaum, Gross, en Miller 1990). De voorziene omvang van de database is 50.000 trefwoorden per taal. "The project aims at developing a multilingual database with basic semantic relations between words for several European languages (Dutch, Italian and Spanish). The wordnets will be linked to the American wordnet for English and a shared top-ontology will be derived, while language specific properties are maintained in the individual wordnets. The database can be used for multilingual information retrieval which will be demonstrated by Novell Linguistic Development."
  Beschikbaarheid: De resultaten van het project zullen ter zijner tijd onder licentie beschikbaar worden gesteld aan derden.

- **RBN.**
  Omschrijving: Het Referentiebestand Nederlands, een multifunctionele lexicale databank met informatie met betrekking tot morfologie, syntaxis, combinatoriek, semantiek en pragmatiek, op een expliciete, formele wijze weergegeven in de vorm van feature–value paren (45.000 lemmata).
  Beschikbaarheid: Het RBN zal ter zijner tijd beschikbaar worden gemaakt voor wetenschappelijk onderzoek.

Meertalige woordenlijsten

- **Van Dale.**
  Omschrijving: De vertaalwoordenboeken Engels-Nederlands/Nederlands-Engels, Frans-Nederlands/ Nederlands-Frans, en Duits-Nederlands/

[21] http://www.kun.nl/celex/
[22] http://www.ldc.upenn.edu/
[23] http://www.let.uva.nl/~ewn/

Nederlands-Duits zijn op CD-ROM beschikbaar.
Beschikbaarheid: Distributie door Van Dale[24].

- **ECHO Eurodicautom.**
  Omschrijving: "The EURODICAUTOM database is an multilingual source for all aspects of European institutions terminology, contextual phrases and abbreviations in all official languages of the European Union. Whilst the database is produced by the Translation service of the European Commission (EC), the content is an accumulation of terminology collected from outside sources such as international organisations, dictionaries and glossaries. Eurodicautom currently contains definitions for more than 4.500.000 terms and 180.000 abbreviations. The number of translated terms and abbreviations are distributed throughout the languages. EURODICAUTOM is a multilingual data bank. The languages available are : Danish, Dutch, English, French, German, Greek, Italian, Portuguese, Spanish, Finnish and Swedish. The database is updated monthly and the data are collected from 1976 onwards."
  Beschikbaarheid: De database kan on-line geraadplegd worden via de Eurodicautom web-pagina[25] .

- **Euterpe (Trados).**
  Omschrijving: "This dictionary was created by the European Parliament. It now contains over 150,000 entries in the 12 official languages of the European Community."
  Beschikbaarheid: Distributie via Trados[26].

Gesproken Taal

- **Multi-language pronunciation dictionary (Onomastica).**
  Omschrijving: "This resource is from the ONOMASTICA LRE-61004 project related to the Multi-language pronunciation dictionaries of proper names. The project ONOMASTICA, funded by the LRE programme, has built pronunciation dictionaries for the names of the European Union. These are city and town names, street names, family names, first names, product names, for 11 languages - Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish and Swedish. An extension to Eastern European languages is ongoing. ELRA and the Dutch PTT are negotiating the terms, conditions and modalities of distribution of the complete Dutch Onomastica dictionary."
  Beschikbaarheid: Distributie door ELRA.

- **Speri-Data AG Basic dictionaries (colloquial language).**
  Omschrijving: "These dictionaries contain a daily-life vocabulary. They include phonetic transcription with related phoneme lists. The following languages are available: Dutch 12,000 entries."
  Beschikbaarheid: Distributie door ELRA.

[24] `http://www.vandale.nl/current/`
[25] `http://www2.echo.lu/edic/`
[26] `http://germany.trados.com:4712/MTW_LOGON`

- **FONILEX.**
  Omschrijving: Fonilex[27] is een uitspraaklexicon voor het Nederlands in Vlaanderen. Het werd ontwikkeld in het kader van het Vlaams korte termijnprogramma Spraak- en taaltechnologie voor het Nederlands. Fonilex bevat de uitspraak van ruim 200.000 woorden, ontleend aan de CELEX database. Voor elke woordvorm vermeldt de databank de spelling, de fonologische vorm, en minstens één en ten hoogste drie (regelmatige) fonetische uitspraakvormen. Elke ingang bevat tevens een identificatienummer dat het verband legt met de CELEX database, zodat daar vermelde gegevens eveneens toegankelijk zijn.
  Beschikbaarheid: Sinds kort is het programma voor wetenschappelijk onderzoek beschikbaar gesteld voor derden.

2.1.3 Overige Hulpmiddelen

In deze sectie beschrijven we een aantal hulpmiddelen voor TST die een algoritmisch aspect bevatten en die speciaal gericht zijn op het Nederlands. Daarnaast noemen we een aantal standaards. We gaan in deze sectie niet in op algemene software voor het exploreren van corpora en woordenboeken, het maken van morfologische software, of het ontwikkelen van grammatica's.

Morfologische analyse

- **Xerox.**
  Omschrijving: Xerox Research Centre Europe[28] heeft programma's ontwikkeld voor de morfologische analyse van diverse Europese talen, gebaseerd op het gebruik van *finite state* technologie. De on-line demo geeft bijvoorbeeld het resultaat in figuur 2.1. De technologie wordt ingezet bij toepassingen waarin verschillende talen een rol spelen (automatisch vertalen, multilinguale IR).
  Beschikbaarheid: Er is een on-line demo versie. Voor commerciële licenties en toepassingen is de Xerox-dochter Inxight[29] verantwoordelijk.

- **Uplift.**
  Omschrijving: Het Uplift project[30] onderzoekt de mogelijkheden van taaltechnologie voor IR. Als onderdeel van een project is een *stemmer* gemaakt (een programma dat de stam van verbogen woorden bepaalt). Merk op dat een stemmer geen informatie geeft over de categorie of morfologische kenmerken van een woord, maar zich beperkt tot het bepalen van de stem. Daarnaast wordt de (taalkundige) accuratesse van het systeem beperkt doordat geen woordenboek wordt gebruikt. (Analyse van `klapschaatsen` geeft `klapschaats`, van `gemiddelde` geeft `middel`, van `varken` (en `varkens`) geeft `vark`, van `kinderen` geeft `kind`, van `verzekeren` geeft `verzekeer`.)
  Beschikbaarheid: Een on-line demo is beschikbaar, en ook de code van het programma is publiek gemaakt.

[27] `http://bach.arts.kuleuven.ac.be/fonilex/`
[28] `http://www.rxrc.xerox.com/research/mltt/Tools/morph.html`
[29] `http://www.inxight.com/enter.htm`
[30] `http://www-uilots.let.ruu.nl/~uplift/`

```
Morphologically analyzed text

<< We >>
we+Pron+Nom+MF+1P+Pl

<< gaan >>
gaan+Verb+Inf
gaan+Verb+PresInd+Pl

<< in >>
innen+Verb+PresInd+1P+Sg
innen+Verb+Imp+Sg
in+For
in+Prep

<< deze >>
deze+Pron+Dem

<< sectie >>
sectie+Noun+F+Sg

<< niet >>
nieten+Verb+PresInd+1P+Sg
nieten+Verb+Imp+Sg
niet+Noun+M+Sg
niet+Adv
nieten+Verb+Imp+Pl
nieten+Verb+PresInd+2P+Pl
nieten+Verb+PresInd+2P+Sg
nieten+Verb+PresInd+3P+Sg

<< in >>
innen+Verb+PresInd+1P+Sg
innen+Verb+Imp+Sg
in+For
in+Prep

<< op >>
op+Adj
op+Adv
op+Prep
```

Figuur 2.1: Uitvoer van de on-line demo van Xerox' morfologische analyseprogramma.

Part-of-speech taggers

- **Xerox.**
  Omschrijving: De gebruikte technieken worden niet beschreven, maar het is waarschijnlijk dat de Xerox tagger[31] ontwikkeld is volgens de statistische benadering die wordt beschreven in Chanod en Tapanainen (1995) en Cutting, Kupiec, Pedersen, en Sibun (1992). Deze methode maakt gebruik van een morfologisch analyseprogramma, in combinatie met een *guesser* die de mogelijke categorieën van onbekende woorden bepaald, en een statistisch model dat gebruik maakt van frequentieinformatie (*woord W is met n% kans een zelfstandig naamwoord*) en van informatie over de waarschijnlijkheid van woordklassecombinaties (*de kans dat een adjectief vooraf wordt gegaan door een lidwoord is n%*). De gebruikte *tagset* bevat 49 elementen. Een voorbeeld van de werking van de on-line demo-versie is te vinden in figuur 2.2.
  Beschikbaarheid: Er is een on-line demo versie. Voor commerciële licenties en toepassingen is de Xerox-dochter Inxight[32] verantwoordelijk.

- **Wotan.**
  Omschrijving: "De WOTAN tagger is ontwikkeld door Johan Berghmans (Berghmans 1994). De WOTAN *tagger* maakt gebruik van softwarecomponenten die door de TOSCA-groep zijn ontwikkeld voor het Engels, maar die voor het grootste gedeelte taalonafhankelijk zijn. De trainingsdata, die natuurlijk wel taalafhankelijk zijn, bestonden uit 1,5 miljoen woorden die in eerdere projecten getagd en gecontroleerd waren. De *tagger* bestaat uit vier componenten. Een eerste component is de tokenizer. Die maakt de onbewerkte tekst in ASCII-formaat klaar voor verwerking. Een tweede component is een woordvormenlexicon. Dit is met behulp van speciale software geëxtraheerd uit het trainingscorpus. Omdat er in de tagger geen morfologische regelcomponent is opgenomen, waaruit stamvormen zouden kunnen worden afgeleid, bevat het lexicon geen woordstammen, maar gehele woordvormen. Een derde component is het suffixenlexicon. Aangezien het mogelijk is dat een woord uit de te analyseren tekst niet in het woordvormenlexicon voorkomt (omdat het niet in het trainingsmateriaal aanwezig was), is er een suffixenlexicon voorzien. Dit moet voorkomen dat onbekende woorden niet van een tag zouden worden voorzien. Een vierde component tenslotte is van probabilistische aard. Deze component bepaalt op basis van de context welke tag aan een woord moet worden toegekend."[33] Figuur 2.3 geeft een voorbeeld van de analyse van een tekstfragment met de WOTAN-tagger.
  Beschikbaarheid: Geen gegevens.

- **MBT (KUB).**
  Omschrijving: Aan de KUB werd een *memory-based part-of-speech tagger-generator* ontwikkeld (Daelemans, Zavrel, Berck, en Gillis 1996b, Daelemans, Zavrel, Berck, en Gillis 1996a). Met behulp hiervan is o.a. een *tagger* voor het Nederlands ontwikkeld, die gebruik maakt van de vereenvoudigde WOTAN tagset (alleen woordsoorten, 13 tags). Het foutenpercentage voor deze tagger, die

[31] `http://www.rxrc.xerox.com/research/mltt/Tools/pos.html`
[32] `http://www.inxight.com/enter.htm`
[33] Omschrijving ontleend aan het ANNO-project.

```
Merk    merk    +NOUN
op      op      +PTKVA
dat     dat     +RELSUB
een     een     +ART
stemmer stemmer +NOUN
geen    geen    +INDDET
informatie      informatie      +NOUN
geeft   geven   +VVFIN
over    over    +PREP
de      de      +ART
categorie       categorie       +NOUN
of      of      +CON
morfologische   morfologisch    +ADJA
kenmerken       kenmerken       +NOUN
van     van     +PREP
een     een     +ART
woord   woord   +NOUN
,       ,       +CM
maar    maar    +CON
zich    zich    +PERS
beperkt beperken        +VVPP
tot     tot     +PREP
het     het     +ART
bepalen bepalen +VVINF
van     van     +PREP
de      de      +ART
stem    stem    +NOUN
.       .       +SENT
```

Figuur 2.2: Voorbeeld uitvoer van de Xerox tagger.


```
Nieuwsbericht       N(eigen,ev,neut)
BTRN                N(eigen,ev,neut)"
Radio               N(eigen,ev,neut)"
1                   N(eigen,ev,neut)"
:                   Punc(dub_punt)
zondag              N(eigen,ev,neut)
19                  Num(hoofd,bep,attr,onverv)
maart               N(eigen,ev,neut)
1995                Num(hoofd,bep,zelfst,onverv)
,                   Punc(komma)
19.00u              N(soort,ev,neut)
Zeven               Num(hoofd,bep,attr,onverv)
uur                 N(soort,ev,neut)
.                   Punc(punt)
^
Het                 Art(bep,onzijd,neut)
KMI                 N(soort,ev,neut)
verwacht            V(trans,ott,3,ev)
nog                 Adv(gew,geen_func,stell,onverv)
altijd              Adv(gew,aanw)
wisselvallig        Adj(attr,stell,onverv)
weer                Adv(gew,geen_func,stell,onverv)
met                 Prep(voor)
buien               N(soort,mv,neut)
van                 Prep(voor)
regen               N(soort,ev,neut)
of                  Conj(neven)
stofhagel           N(soort,ev,neut)
.                   Punc(punt)
^
```

Figuur 2.3: Voorbeeld van de uitvoer van WOTAN. Hetˆ-teken geeft hier een zinsgrens aan.

```
Language = Dutch

U kunt een aanvraag indienen door uw nummer en naam en
wachtwoord in te vullen en op de knop te klikken .

U/Pron kunt/V een/Art aanvraag/N indienen/V door/Prep uw/Pron
nummer/N en/Conj naam/N en/Conj wachtwoord/N in/Adv te/Prep
vullen/V en/Conj op/Prep de/Art knop/N
te/Prep klikken//V ./.
```

Figuur 2.4: Uitvoer van de MBT tagger.

werd getraind op het Eindhoven-corpus is 4% (3% voor bekende woorden, 28% voor onbekende woorden).[34] Een voorbeeld van de uitvoer is te vinden in figuur 2.4. Er bestaat ook een versie van MBT met de volledige WOTAN *tag set*.
Beschikbaarheid: Er is een *on-line* demo-versie[35] van de MBT *tagger* beschikbaar. Specifieke *taggers* zijn onder voorwaarden beschikbaar voor onderzoekdoeleinden.

- **DutchTale** (INL).
Omschrijving: "Om de teksten voor zijn Taaldatabank van het Hedendaagse Nederlands te taggen en lemmatiseren, gebruikt het INL o.a. de POS *tagger* Dutchtale. De tagger (voor het Nederlands) is volautomatisch, maar biedt wel de mogelijkheid om achteraf, met behulp van een intelligente tekstverwerker die daarvoor ontwikkeld is, de output manueel te controleren.

  Het systeem bestaat uit drie modules: In de eerste module (lexicale opzoek- en morfologische analysemodule) vindt lemmatisering en woordsoorttoekenning plaats. Als een woord niet in een tokenlexicon gevonden, dan vindt er morfologische analyse plaats. Het gaat hier om zowel samenstellings- als afleidingsanalyse. De tweede module desambigueert op basis van een beperkte regelcomponent. In een apart regelbestand staan enkele honderden desambigueringsregels opgesteld in een formele taal. De derde module is statistisch van aard en werkt met n-grammen. De meest waarschijnlijke desambiguering wordt gekozen op basis van de directe linker en rechter context van een ambigu token. Een woordsoorttag wordt toegevoegd indien er in de voorgaande modules geen lemma was toegekend."[36]
Beschikbaarheid: Geen gegevens.

- **D-Tale** (VU).
Omschrijving: De *dutch tagger lemmatizer* voor het Nederlands is een programma dat werd ontwikkeld door de afdeling lexicologie aan de VU, voor Van Dale. De output is een tekst voorzien van tags die woordklasse (zelfst. naamw, werkwoord) en eventuele aanvullende kenmerken aangeven (enkelvoud/ meervoud) en ook het lemma. Tijdens de analyse raadpleegt het programma het woordenboek (140.000 woordvormen), worden morfologische regels toegepast, worden mogelijke woordsoorten van onbekende woorden voorgesteld, en worden tenslotte met behulp van statistische technieken ambiguiteiten opgelost. Het foutenpercentage is 7%.[37]
Beschikbaarheid: Geen gegevens.

[34]Informatie ontleend aan een overzicht van Nelleke Oostdijk (KUN).
[35]`http://ilk.kub.nl`
[36]Informatie ontleend aan het ANNO-project. Meer informatie over de tagger is te vinden in van der Voort van der Kleij, Raaijmakers, Panhuijsen, Meijering, en van Sterkenburg (1994).

Syntactische analyse

- **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. 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. Het systeem is getest op verschillende documenten (o.a. juridische teksten, wetenschappelijke boeken en scripties, en (6 megabyte) nieuwsberichten).
Beschikbaarheid: Geen gegevens.

- **Amazon.**
Omschrijving: Amazon is een grammatica ontwikkeld door Peter-Arno Coppen (KUN). Het is een grammatica die het Nederlands redelijk breed afdekt: ongeveer 95% van de ingevoerde tekst kan door het systeem ontleed worden. Een *on-line* versie[38] van het systeem is te vinden op de AGFL (*affix grammars over finite lattices*) web-pagina. Deze versie werd gemaakt door Erik Oltmans (Oltmans 1994), en maakt gebruik van een woordenboek van ongeveer 300.000 woorden dat uit CELEX werd afgeleid. Voorbeeld uitvoer van het systeem is te vinden in figuur 2.5.
Beschikbaarheid: Geen gegevens.

Spraakherkenning

Er zijn geen producten voor spraakherkenning die zich speciaal op het Nederlands richten.

Spraaksynthese

- **Fluent Dutch.** Fluent Speech Technology[39] ontwikkelt *Fluent Dutch*, een programma voor spraaksynthese. Men maakt gebruik van de MBROLA (public domain) difoon spraaksynthesizer[40]. Er is een difoon-model voor het Nederlands beschikbaar (ook via MBROLA). Aan een programma voor grafeem-naarfoneem conversie wordt nog gewerkt:

[37]Informatie ontleend aan een overzicht van Nelleke Oostdijk (KUN).
[38]`http://www.cs.kun.nl/agfl/`
[39]`http://www-uilots.let.ruu.nl/~Arthur.Dirksen/fluent/fluent.htm`
[40]`http://tcts.fpms.ac.be/synthesis/mbrola.html`

```
de hieronder gegeven informatie is overgenomen uit
het boek$EOS$
---------------------------------------------------
parsing 1       (0.119 sec.)

  SE(+MAIN, P, EMPTY, NORM, EMPTY)
    TOP
      NP(SING, NORM, TOPP)
        NII(MAXPRO, SING, -NEU, NORM)
          DT(+DEF, SING, -NEU, ART)
            "de"
            SE SUB(VD, EMPTY, NORM, EMPTY)
                MI(NORM)
                      BW
                          "hieronder"
              CL(-MAIN, VD)
                  ww(VD, EMPTY)
                    "gegeven"
          NI(+CASE, SING, -NEU, aMOD)
            N(SING, -NEU)
              "informatie"
      ww(P, EMPTY)
        "is"
      MI(NORM)
              ww(VD, EMPTY)
                "overgenomen"
                PP(REST)
                  VZ(REST)
                    "uit"
                  NP(SING, NORM, MID)
                    NII(MAXPRO, SING, +NEU, NORM)
                      DT(+DEF, SING, +NEU, ART)
                        "het"
                      NI(+CASE, SING, +NEU, -MOD)
                        N(SING, +NEU)
                          "boek"
```

Figuur 2.5: Syntactische analyse in AMAZON.

" Fluent Dutch is a speech synthesis system for Dutch, which runs under Windows 3.1 or higher. It is not (yet) a full-fledged text-to-speech synthesizer, but generates synthetic speech of a superior quality from a phonetic transcription. The system can be used in multimedia applications such as "talking" dictionaries and educational CD-ROMS, as well as in dialogue systems of various kinds. At present, the system uses a male voice. A female voice is in preparation."
Beschikbaarheid: Een deel van het materiaal is vrij beschikbaar via Mbrola.

- **TreeTalk.**
  Omschrijving: TreeTalk is een programma voor grafeem-naar-foneem conversie dat is ontwikkeld binnen het *inductive language learning* project[41] aan de KUB (van den Bosch 1997, Daelemans en van den Bosch 1996). Het programma zet een reeks woorden om in reeksen fonemen in DISC notatie (de notatie die door CELEX wordt gebruikt). Het programma werkt op woordbasis en houdt dus geen rekening met prosodische effecten op zinsniveau.
  Beschikbaarheid: Er is een on-line demo-versie[42] van het programma beschikbaar.

Standaards

- **Fonetische Alfabetten.**
  Het meest gebruikte fonetische alfabet is IPA (*international phonetic association*. Voor codering van corpora en elektronische lexica is dit alfabet echter minder geschikt, omdat het gebruikt maakt van symbolen die geen deel uitmaken van het ASCII alfabet. Om dit probleem te omzeilen zijn verschillende ASCII fonetische alfabetten ontwikkeld. Een algemeen aanvaard alfabet is SAMPA (Wells 1987). In de CELEX database wordt bijvoorbeeld gebruik gemaakt van SAMPA. Daarnaast zijn binnen CELEX nog drie andere notaties beschikbaar (CELEX, CPA, en DISC), waarvan er één (DISC) zo is ontworpen dat ieder fonetisch symbool met precies één teken correspondeert (met name nuttig voor computationele toepassingen). Merk overigens op dat het hier steeds slechts notationele varianten van SAMPA en IPA betreft, zodat de verschillende notaties eenvoudig naar elkaar omgezet kunnen worden. Voor het FONILEX uitspraakwoordenboek werd gebruik gemaakt van YAPA (*yet another phonetic alphabet*). Dit alfabet wijkt op onderdelen af van de notatie die in CELEX primair is (DISC), o.a. doordat het rekening houdt met de uitspraak van (Franse) leenwoorden.

- **Woordsoorten.**
  Corpora die zijn voorzien van informatie over woordsoort en *part of speech taggers* die automatisch woordsoorten toekennen maken gebruik van een vooraf gedefinieerde verzameling woordsoorten, de *tagset*. In de WOTAN *tagger* wordt gebruik gemaakt van een *tagset* die is ontwikkeld door de TOSCA groep in Nijmegen. Deze tagset bestaat uit 243 verschillende tags (zie ook figuur 2.3). Uitgangspunt was de verdeling in hoofdwoordsoorten en hun onderverdelingen zoals die is te vinden in de ANS. Een sterk vereenvoudigde versie van deze tagset

[41] http://ilk.kub.nl/
[42] http://ilk.kub.nl/

maakt alleen een onderscheid in woordsoorten, en bevat slechts 12 elementen (ADJ, ADV, ART, CONJ, INT, MISC, N, NUM, PREP, PRON, PUNC, V). Deze wordt bijvoorbeeld gebruikt in de (online) MBT tagger. De Xerox tagger maakt gebruik van een verzameling van 49 tags (zie ook figuur 2.2). De INL corpora maken gebruik van een kleine tagset (dertien elementen: a(djectief), b(ijwoord), c(onjunctie), e(igennaam), l(idwoord), o(ngespecificeerd), p(ronomen), t(elwoord), v(oorzetsel), z(elfstandig naamwoord)), en soms van een grotere tagset, die werd ontwikkeld door de TOSCA groep (KUN) in het kader van het MECOLB-project. Deze tagset is waarschijnlijk identiek aan de WOTAN tagset.
Binnen het reeds vaker genoemde EAGLES-project[43] is een aanzet gegeven voor een standaard voor morfosyntactische annotatie van corpora[44]. Er worden vier typen tags onderscheiden:

Verplichte tags. Hiertoe behoren de hoofdwoordsoorten (Zelfstandig Naamwoord, Werkwoord, Voegwoord, ...

Aanbevolen tags. Hiertoe behoren de tags die algemeen gebruikte eigenschappen aanduiden als Geslacht, Getal, Persoon, ...

Optionele tags. Voor bepaalde doeleinden kan het nodig zijn meer specifieke woordsoorten, eigenschappen etc. te onderscheiden. Dit wordt gedaan met optionele tags.

Taalspecifieke tags. Bepaalde woordsoorten en eigenschappen komen slechts in enkele (Europese) talen voor. Hiervoor zijn aparte tags nodig. Voor het Nederlands zijn er bijvoorbeeld extra tags voorgesteld voor de waarden *De*, respectievelijk *Het*-woord bij Geslacht, en de waarden *Vol*, respectievelijk *Gereduceerd* bij Persoonlijk Voornaamwoord.

De volledige tagset is op bovengenoemde webpagina terug te vinden.

- **Syntactische en semantische annotatie.**
Binnen het hiervoor genoemde EAGLES-project is ook een voorstel gedaan voor een standaard voor syntactische annotatie[45]. Ook hier zijn er een aantal typen tags voorgesteld:

Verplichte tags. Er zijn vele redenen om een corpus syntactisch te annoteren. Er (b)lijkt geen type annotatie te zijn die voor alle doeleinden voldoet.

Aanbevolen tags: de bekende categorieën Zin, Niet-finiete Zin, Nominale Constituent, Verbale Constituent, Adjectivische Constituent, Adverbiale Constituent en Voorzetsel Constituent

Optionele tags. Er worden verschillende typen optionele tags onderscheiden waaronder:

Syntactische tags. Hieronder vallen tags voor bijvoorbeeld het nader benoemen van Niet-finiete Zinnen: afhankelijke zin, beknopte bijzin, enz. Of voor het aanduiden van de grammatische functie: Onderwerp,

[43] http://www.ilc.pi.cnr.it/EAGLES/home.html
[44] http://www.ilc.pi.cnr.it/EAGLES96/annotate/annotate.html
[45] http://www.ilc.pi.cnr.it/EAGLES96/segsasg1/segsasg1.html


```
(2 [CLS [CLS [NP $SUBJ ("Het" 1) ("KMI" 2) ]
[PRED ("verwacht" 3) ] [PP ("vooral" 4) ("in" 5)
("het" 6) ("westen" 7)]
[PP $POBJ ("van" 8) ("het" 9) ("land" 10) ]
[NP $DOBJ ("mooie" 11) ("opklaringen" 12) ] ] ("," 13)
[CLS ("elders" 14) [PRED ("is" 15) ] ("er" 16) ("af" 17)
("en" 18) ("toe" 19) [NP $SUBJ ("ook" 20) ("bewolking" 21) ]
[PP ("met" 23) ("vooral" 24) ("in" 25) ("de" 26)
("Ardennen" 27) [PP ("op" 30) ("nog" 28) ("kans" 29) ]
("lichte" 31) ("voorjaarsbuien"32) ] ] ] ("." 33) )
```

Figuur 2.6: Syntactische annotatie in ANNO.

```
[van,middelburg,wil,ik,reizen,naar,groningen],
"(origin.place.town.middelburg;
  user.wants.travel.destination.place.town.groningen)
",
"(
 (TS|(d1;d2)
    (MP|d1.d2 P|origin.place/van NP|town.middelburg/
                                              middelburg)
    (SO|d1.d2
        (SO|d2.d1 V|wants/wil PER|user/ik)
        (INFP|d1.d2 INF|travel/reizen
            (MP|d1.d2 P|destination.place/naar
            NP|town.groningen/groningen))))
"
```

Figuur 2.7: Syntactisch/semantische annotatie in OVIS.

Meewerkend Voorwerp, enz. De keuze van de tags is in al deze gevallen voor een deel taalspecifiek.

Semantische tags. Voor nadere tags bij een NC valt hier te denken aan Definiet en Indefiniet, bij een AdvC aan een nadere precisering als Tijd, Plaats, Hoedanigheid, etc.

Het zal duidelijk zijn dat deze EAGLES-standaard nog niet de status heeft van bijvoorbeeld die met betrekking tot morfosyntactische annotatie.
Er zijn niet of nauwelijks voorbeelden van syntactisch geannoteerde corpora voor het Nederlands. Binnen het ANNO project is hiertoe wel een aanzet gegeven (zie figuur 2.6 voor een voorbeeld). Binnen het OVIS project[46] is een deel van het corpus voorzien van syntactische en semantische annotatie (zie figuur 2.7).

[46] http://odur.let.rug.nl:4321/

2.1.4 Het internationale perspectief

In deze sectie noemen we een aantal hulpmiddelen die voor andere talen beschikbaar zijn, die van groot belang lijken voor onderzoek op TST-gebied, en die door onze interviewpartners genoemd werden als voorbeelden en als hulpmiddelen waarvan men graag een Nederlandstalige tegenhanger zou zien. Binnen het kader van dit onderzoek bleek het niet mogelijk een uitputtend onderzoek te doen naar hulpmiddelen voor talen anders dan het Nederlands, en ook onze gesprekspartners gaven meermalen aan niet over zo'n overzicht te beschikken. Wel lijkt men het eens te zijn over een aantal van de belangrijkste producten. Het onderstaande overzicht is gebaseerd op voorbeelden die in de interviews werden genoemd. Het is zeer onvolledig, en beperkt zich tot het Engels. Bij instanties als ELRA, LDC, het DFKI *software registry* is nog een veelvoud aan vergelijkbaar materiaal te vinden.

Corpora

- **British National Corpus.**
  Omschrijving: Dit is een groot (100 miljoen) corpus van gesproken en geschreven Engels.[47] Het corpus is samengesteld uit kortere teksten (maximaal 45.000 woorden) en fragmenten (van maximaal 45.000) woorden uit langere teksten. Ongeveer 10 miljoen woorden van het corpus zijn transcripties van gesproken Engels.
  Beschikbaarheid: Het corpus is beschikbaar op CD-ROM en via het WWW. De teksten uit het corpus mogen zelf niet verder gedistribueerd of openbaar gemaakt worden, maar alle resultaten die gebaseerd zijn op corpusonderzoek mogen gebruikt worden voor onderzoeksdoeleinden en daarop gebaseerde producten.

- **Brown Corpus, LOB Corpus.**
  Omschrijving: Het ICAME (*International Computer Archive of Modern and Medieval English*)[48] distribueert verschillende corpora, waaronder het Browncorpus (1 miljoen woorden) en het Londen-Oslo-Bergen-corpus (1 miljoen woorden). Beide corpora zijn beschikbaar in verschillende formaten, waaronder versies die voorzien zijn van woordsoort.
  Beschikbaarheid: Distributie door ICAME.

- **Penn Treebank.**
  Omschrijving: De Penn Treebank bestaat uit een aantal corpora (o.a. 1,6 miljoen woorden uit de Dow Jones nieuwsdienst, 1 miljoen woorden uit de Wall Street Journal, materiaal ontleend aan ATIS, MUC, en IBM handleidingen) die handmatig van constituentstructuur zijn voorzien.
  Beschikbaarheid: Het corpus wordt gedistribueerd door het Linguistic Data Consortium.[49] Het corpus is beschikbaar op CD-ROM, en kan ook gedeeltelijk via het Web geraadpleegd worden.

- **UN Corpus.**
  Omschrijving: Het UN-corpus bestaat uit parallele teksten voor het Engels, Frans, en Spaans met een omvang van 2,5 gigabyte tekst.
  Beschikbaarheid: Het corpus wordt gedistribueerd door het Linguistic Data Consortium.

- **Crater.**
  Omschrijving: Het CRATER-corpus is een 1 miljoen parallell (*aligned*) corpus voor het Engels, Frans, en Spaans. Het corpus is voorzien van annotatie (lemma en woordsoort, gecorrigeerd), en wordt geleverd met hulpmiddelen voor extractie en *alignment*.
  Beschikbaarheid: Distributie door ELRA.

- **ATIS.**
  Omschrijving: Het ATIS (*air travel information system*) corpus, verkrijgbaar bij LDC, bevat verschillende opnames van interactie tussen gebruikers en een (echt of gesimuleerd) systeem dat informatie geeft over verbindingen van luchtvaartmaatschappijen. De corpora zijn aangemaakt en gebruikt in het DARPA *spoken language systems* programma.
  Beschikbaarheid: Distributie door het Linguistic Data Consortium.

- **CSR.**
  Omschrijving: De CSR (*continuous speech recognition* corpora, verkrijgbaar bij LDC, bevatten gelezen fragmenten ontleend aan het Wall Street Journal corpus (zie Penn Treebank). De fragmenten zijn zo gekozen dat ze gebruik maken van een vocabulaire van 5.000 of 20.000 woorden. De corpora zijn gebruikt in het DARPA *Spoken Language Program*.
  Beschikbaarheid: Distributie door het Linguistic Data Consortium.

- **Overige LDC Corpora.** Naast bovengenoemde corpora zijn via het Linguistic Data Consortium nog een groot aantal andere spraakcorpora, zoals CALLFRIEND, CALLHOME, en SWITSCHBOARD en tekstcorpora, zoals de BROADCAST NEWS TRANSCRIPTS, het NORTH AMERICAN NEWS TEXT CORPUS, en TIPSTER, beschikbaar.

Lexica

- **COBUILD.**
  Omschrijving: Het Collins' Cobuild (*learners*) woordenboek[50] is gebaseerd op intensief gebruik van corpusgegevens (waarvoor een *Bank of English* is samengesteld met een omvang van meer dan 320 miljoen woorden).
  Beschikbaarheid: Distributie door Colbuild. Het corpus zelf is deels on-line te raadplegen, en wordt deels meegeleverd op de Cobuild CD-ROM, die het woordenboek, een grammatica, en verwijzingen naar het corpus bevat.

- **WordNet.**
  Omschrijving: WordNet is een semantische lexicale database voor het Engels (Miller *et al.* 1990), met een omvang van meer dan 100.000 woorden, waarin voor zelfstandige naamwoorden, adjectieven, en werkwoorden informatie over

[47] `http://info.ox.ac.uk/bnc/`
[48] `http://www.hd.uib.no/icame.html`
[49] `http://www.ldc.upenn.edu/`
[50] `http://titania.cobuild.collins.co.uk/`

synoniemen (*plank, board*), antoniemen (*rise, fall*), hyponiemen (*maple, tree*, ISA-relatie), en meroniemen (*tree, root*) (HASA-relatie).
Beschikbaarheid: De database is vrij beschikbaar via de Wordnet web-pagina.[51]

- **Comlex.**
Omschrijving: Comlex is een woordenboek met ongeveer 38.000 trefwoorden dat gedetailleerde informatie bevat over de syntactische eigenschappen van ieder woord, met name valentie.
Beschikbaarheid: Distributie door het Linguistic Data Consortium.

Halffabrikaten en overig

- **Brill-tagger.**
Omschrijving: De POS-tagger van Eric Brill (Brill 1995) maakt gebruik van *transformation-based error-driven learning* om uit een corpus voorzien van woordsoorten een *tagger* af te leiden.
Beschikbaarheid: De *tagger* is vrij beschikbaar via Brill's web-pagina.[52]

- **EngCG-2.**
Omschrijving: EngCG-2 van het Finse bedrijf Conexor is een snelle POS-tagger voor het Engels, die gebruik maakt van *constraint grammar*, een regelformalisme gebaseerd op *finite state* technologie (Samuelson en Voutilainen 1997).
Beschikbaarheid: Commerciële en academische licenties zijn verkrijgbaar via Conexor.[53]

- **XPOST.**
Omschrijving: De Xerox Part-of-Speech Tagger XPOST (Cutting *et al.* 1992) is een tagger geïmplementeerd in LISP, getraind op het Brown-corpus.
Beschikbaarheid: De tagger is beschikbaar voor onderzoeksdoeleinden.

- **XTAG.**
Omschrijving: XTAG[54] is een *wide-coverage* grammatica en voor het Engels, gebaseerd op *Tree Adjoining Grammar* ontwikkeld voor onderzoeksdoeleinden. Het systeem bevat een tagger getraind op het Wall Street Journal corpus, een woordenboek met meer dat 300.000 woordvormen, en meer dan 300 grammaticale regels (*lexicalized trees*).
Beschikbaarheid: De grammatica, documentatie, en bijbehorende software is vrij beschikbaar.

- **CLE.**
Omschrijving: De *Core Language Engine*[55] (Alshawi 1992) is een computationele grammatica, ontwikkeld door SRI Cambridge[56], die is bedoeld als *general purpose natural language processing system*. De CLE is gebruikt voor natuurlijke taal interfaces (ook voor gesproken taal), vertaalsystemen voor gesproken taal, toepassingen waarbij *controlled language* een rol speelt.
Beschikbaarheid: Licenties voor de CLE zijn mogelijk voor onderzoeksdoeleinden en voor commerciële toepassingen.

[51] http://www.cogsci.princeton.edu/~wn/w3wn.html
[52] http://www.cs.jhu.edu/~brill/home.html
[53] http://www.conexor.fi/index.html
[54] http://www.cis.upenn.edu/~xtag/
[55] http://www.cam.sri.com/ccsrc/cle.html
[56] http://www.cam.sri.com

- **EngLite en FDG.**
Omschrijving: De ENGLITE en FDG parsers van het Finse bedrijf Conexor zijn *wide-coverage* parsers die een *light (shallow) syntactic parse* c.q. een *full dependency parse* toe kennen aan zinnen.
Beschikbaarheid: Commerciële en academische licenties zijn verkrijgbaar via Conexor.[57]

- **TSNLP.**
Omschrijving: De TSNLP testsuite bevat geannoteerde data voor het Engels, Duits, en Frans, bedoeld om te worden gebruikt bij het testen en evalueren van taalverwerkende systemen. Per taal zijn meer dan 4000 items opgenomen.
Beschikbaarheid: Distributie via ELRA.

2.2 Immateriële infrastructuur

In deze sectie zal een overzicht worden gegeven van de instellingen (onderzoek, onderwijs, beleid) die actief zijn op het gebied van de Taal- en Spraaktechnologie. Hierbij wordt in een aparte sectie aandacht besteed aan de rol van de industrie. Tenslotte zullen een aantal initiatieven in het buitenland op een rijtje worden gezet.

2.2.1 Onderzoek

In **Nederland** wordt aan de meeste universiteiten onderzoek verricht op het gebied van TST voor het Nederlands. Bijvoorbeeld in:[58]

- Amsterdam, UvA: Alfa-informatica[59] (t), Instituut voor Fonetische Wetenschappen[60] (s)
- Amsterdam, VU: Lexicologie (t), Terminologie (t)
- Delft, TUD: Kennisgestuurde Systemen[61] (s)
- Eindhoven, TUE: Instituut voor Onderzoek naar Mens-Systeem Interactie[62] (s)
- Groningen, RUG: Alfa-informatica[63] (t)

[57] http://www.conexor.fi/index.html
[58] Een (s) betekent dat er voornamelijk onderzoek wordt verricht met betrekking tot spraak, een (t) voornamelijk met betrekking tot taal.
[59] http://earth.let.uva.nl/
[60] http://fonsg3.let.uva.nl
[61] http://www.kbs.twi.tudelft.nl
[62] http://www.tue.nl/ipo/
[63] http://www.let.rug.nl/alfa/

- Leiden, RUL: Algemene Taalwetenschap[64] (t)(s), Functieleer[65](t)
- Nijmegen, KUN: Taal- en Spraaktechnologie[66] (t)(s)
- Tilburg, KUB: Taal en Informatica[67] (t)
- Twente, UT: CTIT[68] (Centrum voor Telematica en Informatietechnologie) (t) (s)
- Utrecht, RUU: UIL/OTS[69] (Utrecht Institute of Linguistics/ Onderzoeksinstituut voor Taal en Spraak) (t)(s)

In **Vlaanderen** vindt er bijvoorbeeld onderzoek plaats aan de universiteiten in:

- Antwerpen, UIA: Centrum voor Nederlandse Taal en Spraak[70] (CNTS) (t)
- Brussel, VUB: ETRO[71] (s)
- Gent, UG: ELIS[72] (s)
- Leuven, KUL: Centrum voor computerlinguïstiek[73] (CCL) (t), ESAT[74] (s)

en eveneens aan de Katholieke Vlaamse Hogeschool (Departement Tolken en Vertalers) in Antwerpen.

Daarnaast zijn er in beide landen nog een aantal *universitaire onderzoekscentra* binnen andere faculteiten, die zich vooral op bijvoorbeeld het juridisch of het medisch taalgebruik richten, bijvoorbeeld ICRI[75] (Interdisciplinair Centrum voor Recht en Informatica, KUL) en Medische Informatica (Gent, Geneeskunde).

In **Nederland** zijn er een aantal *onderzoekscholen* opgericht, zelfstandige organisatorische eenheden met een eigen budgetverantwoordelijkheid. In zo'n onderzoekschool wordt onderzoek van een of meer universiteiten op een bepaald terrein gebundeld met het doel de kwaliteit van het onderzoek te verbeteren en tot een samenhangend onderzoeksprogramma te komen. De scholen kunnen interuniversitair zijn. Enkele voor TST relevante onderzoekscholen zijn:

- Onderzoekschool LOGICA[76],
- IPA[77] (Instituut voor Programmatuurkunde en Algoritmiek),

[64]http://www.leidenuniv.nl/ugids/h6/lv02.htm
[65]http://www.fsw.leidenuniv.nl/www/w3_func/research.htm
[66]http://www.let.kun.nl/onderzoek/05-02.html
[67]http://cwis.kub.nl/~fdl/research/ti/prog_mw.htm
[68]http://wwwctit.cs.utwente.nl/
[69]http://www-uilots.let.ruu.nl/
[70]http://ger-www.uia.ac.be/webger/ger/cnts/main.html
[71]http://etro.vub.ac.be/etro.html
[72]http://www.elis.rug.ac.be/
[73]http://www.ccl.kuleuven.ac.be
[74]http://www.esat.kuleuven.ac.be/~spch/
[75]http://www.law.kuleuven.ac.be/icri/
[76]http://www.wins.uva.nl/research/ozsl/lin
[77]http://www.win.tue.nl/cs/ipa/activities/FMcourse.04.1997.html

- BCN[78] (Behavioural and Cognitive Neurosciences),
- J.F. Schouten Institute for User-System Interaction Research[79] (opvolger van de onderzoekschool Perception and Technology),
- SIKS[80] (School voor Informatie- en KennisSystemen),
- LOT[81] (Landelijke Onderzoekschool Taalwetenschap),
- CLS[82] (Centre for Language Studies),
- HIL[83] (Holland Institute of Generative Linguistics),
- IFOTT[84] (Instituut voor Functioneel Onderzoek van Taal en Taalgebruik) [85]

In **Vlaanderen** werken de universiteiten van Antwerpen, Gent, Brussel (VUB) en Leuven samen in CLIF[86] (Computational Linguistics and Language Technology), een FWO-*onderzoeksgemeenschap*. CLIF heeft zich tot taak gesteld de Vlaamse taaltechnologie te coördineren en internationaal te verankeren. Daarnaast worden er hulpbronnen bijeen gebracht.

In **Nederland** zijn ook nog een aantal *niet-universitaire onderzoekscentra* werkzaam, veelal gefinancierd door overheid, bedrijfsleven en/of universiteiten samen:

- Telematica Instituut in Enschede[87]
- TNO[88], Nederlandse Organisatie voor Toegepast Wetenschappelijk Onderzoek

Het TELEMATICA INSTITUUT[89] is een consortium van bedrijven en kennisinstellingen, met financiële steun van de overheid. Er wordt vooral contractonderzoek uitgevoerd. De deelnemers zijn IBM, KPN, Lucent Technologies, ING en Rabofacet. Daarnaast contribueren ABP/USZO, Cap Gemini, Ericsson, Océ en Syllogic. De bedrijven ECT, Heidemij, NS, NOB, Origin en VNU zijn geassocieerd lid. Het Telematica Consortium is nog in onderhandeling met andere bedrijven over hun deelname. De Universiteit Twente, de Universiteit Delft, CWI, TNO, Multimedia en Telecommunicatie (TNO MET) nemen deel als kennisinstellingen. Het vroegere Telematica Research Centre (TRC) is in het Telematica Instituut opgegaan.

De vele TNO-onderzoeksinstituten voeren vooral contractonderzoek (zowel meer fundamenteel als toegepast) voor de overheid en het bedrijfsleven uit. TNO[90] is een

[78]http://BCN.bcn.rug.nl/bcn/
[79]http://www.tue.nl/ipo/school/index.html
[80]http://www.cs.ruu.nl/siks/
[81]http://www-uilots.let.uu.nl/relations/lot.htm
[82]http://cwis.kub.nl/~fdl/research/school/cls/index.htm
[83]http://oasis.leidenuniv.nl/hil/
[84] http://www.leidenuniv.nl/ugids/h12/z006.htm
[85]De laatste drie maken deel uit van LOT
[86]http://ger-www.uia.ac.be/webger/ger/clif/
[87]http://www.trc.nl/
[88]http://www.tno.nl/
[89]http://www.telin.nl/
[90]http://www.tno.nl/

semi-overheidsinstelling (het is opgericht door de overheid, maar kan tot op grote hoogte een eigen beleid voeren). Voor TST zijn vooral belangrijk het TNO Institute for Applied Physics[91] (TPD, Delft), het TNO Fysisch en Elektronisch Laboratorium[92] (Den Haag), het TNO Human Factors Research Institute[93] (Soesterberg) en, op beleidsniveau, TNO Strategie, Technologie en Beleid[94] (STB, Apeldoorn). Het CWI[95] (Centrum voor Wiskunde en Informatica) Amsterdam is meer zijdelings bij het TST-onderzoek voor het Nederlands betrokken.

Verder zijn er nog belangrijke instituten die vooral toegepast onderzoek verrichten, of *resources* en *tools* ter beschikking stellen:

- CELEX[96] (Centrum voor Lexicale Informatie), Nijmegen (Max Planck Instituut). In 1986 gesticht door 5 Nederlandse onderzoekscentra, waaronder het Max Planck Instituut en de Universiteit van Nijmegen (Taal en Spraak). Sinds 1989 is CELEX erkend als nationaal kenniscentrum. CELEX beschikt over een grote database met informatie met betrekking tot fonologie, morfologie, syntaxis en frequentie voor Nederlandse, Duitse en Engelse lemmata.

- SPEX[97], (*Speech Processing EXpertise Centre*), Nijmegen (KUN). Opgericht in 1987, deelnemende universiteiten: Amsterdam, Utrecht, Nijmegen, Leiden en Eindhoven. SPEX is een organisatie die zich bezighoudt met het ontwikkelen en beschikbaar stellen van software, tools en databases op het gebied van spraaktechnologie. Validatie (spraak) is momenteel een van de belangrijkste activiteiten (bijvoorbeeld voor ELRA).

- STDH[98] (Stichting Tekstcorpora en Databestanden in de Humaniora). Opgericht in 1990. De STDH is opgericht met als doel het onderzoek op het gebied van de tekstcorpora en databestanden in de humaniora te bevorderen en de kennis op dit gebied te verbreiden. Het is de intentie om de website van het STDH uit te bouwen tot een centraal informatiepunt op het gebied van het corpusonderzoek in Nederland en Vlaanderen. Ook wil de STDH de activitieiten op het gebied van corpusonderzoek bundelen en coördineren.

Het gaat in alle drie de gevallen om kleine organisaties (0.5 - 3 fte). Vaak worden ook de subsidies slechts voor een korte termijn toegezegd (zoals in geval van CELEX).

Hoewel de meeste onderzoekscentra puur Nederlands, dan wel Vlaams zijn, is er toch een gezamenlijk NEDERLANDS-VLAAMS initiatief, namelijk het INL (Instituut voor Nederlandse Lexicologie), te Leiden. Het INL is in 1969 opgericht. Relevant voor TST is de INL-Taalbank. Ook participeert het INL in Europese TST projecten, zoals PAROLE.

De belangrijkste onderzoeksprojecten op het gebied van de TST waren in Nederland het NWO *prioriteitsprogramma voor Taal- en Spraaktechnologie*[99] (1995-2000)

[91] http://www.tpd.tno.nl/TPD/smartsite.htm
[92] http://www.tno.nl/instit/fel/fel_nl.html
[93] http://www.tno.nl/instit/tm/index.html
[94] http://www.stb.tno.nl/
[95] http://www.cwi.nl/
[96] http://www.kun.nl/celex/
[97] http://lands.let.kun.nl/spex/
[98] http://CandL.let.ruu.nl/stdh/index.htm
[99] http://odur.let.rug.nl:4321/

en in Vlaanderen het *Korte termijn programma voor Taal- en Spraaktechnologie* (1994-1997). Er zijn tot dusverre ook een paar gezamenlijke onderzoeksinitiatieven geweest op het gebied van TST, namelijk EUROTRA (1982-1993) en het Corpus Gesproken Nederlands (1998-2003).

In Nederland houdt het NIWI[100] (Nederlands Instituut voor Wetenschappelijke Informatiediensten) een *databank* bij met informatie over wetenschappelijk onderzoek in Nederland. In Vlaanderen wordt dat gedaan in de IWETO-databank[101] (Inventaris van het Wetenschappelijk en Technologisch Onderzoek in Vlaanderen) door het Ministerie van de Vlaamse Gemeenschap, afdeling Wetenschap en Innovatie (AWI).

2.2.2 Onderwijs

In **Nederland** en in **Vlaanderen** kan aan alle universiteiten die in de vorige sectie werden genoemd als onderzoekscentra ook in een of andere vorm TST worden gestudeerd.

In **Vlaanderen** maken de *taal*technologische richtingen deel uit van de Letterenfaculteiten, terwijl de *spraak*technologische richtingen zijn ondergebracht bij Toegepaste Wetenschappen. In **Nederland** is er niet zo'n verdeling te maken. Daar zijn de meeste taal- èn spraaktechnologische richtingen ontstaan binnen de Letterenfaculteiten. In Delft, Eindhoven en Twente, de drie technische universiteiten, is TST ondergebracht bij de Faculteit (Technische) Informatica.

In **Nederland** kan men aan elk van de onderstaande universiteiten een studie in een TST-richting volgen. Men moet in de meeste gevallen eerst een propedeuse hebben afgelegd voor men kan overstappen.

- Amsterdam, UvA: Alfa-informatica[102] (t), Fonetische Wetenschappen[103] (s)
- Delft, TUD: Kennisgestuurde Systemen[104] (KGS) (s)
- Eindhoven, TUE: Mens-Systeem Interactie[105] (postdoctoraal) (s)
- Groningen, RUG: Alfa-informatica[106] (t), Technische Cognitie Wetenschap[107] (TCW) (t)
- Nijmegen, KUN: Taal, Spraak en Informatica[108] (t)(s)
- Tilburg, KUB: Taal en Kunstmatige Intelligentie[109] (t)
- Twente, UT: Linguistic Engineering[110] (t) (s)

[100] http://www.niwi.knaw.nl/
[101] http://www.vlaanderen.be/IWETO/DOCS/thema_pag.html
[102] http://earth.let.uva.nl/
[103] http://fonsg3.let.uva.nl
[104] http://www.kbs.twi.tudelft.nl
[105] http://www.tue.nl/ipo/usi/index.html
[106] http://www.let.rug.nl/alfa/
[107] http://tcw2.ppsw.rug.nl
[108] http://www.let.kun.nl/onderwijs/tsi.htm
[109] http://cwis.kub.nl/~fdl/voorl/tkibsch.htm
[110] http://wwwseti.cs.utwente.nl:/Parlevink/Students/studentsframe.html

- Utrecht, RUU: Documentaire Informatiekunde[111] (t), Taal- en Spraakautomatisering[112] (t)(s), Cognitieve Kunstmatige Intelligentie[113] (CKI) (t)

De Nederlandse onderzoekscholen (zie vorige sectie) verzorgen ook opleidingsprogramma's van AIO's, OIO's (Assistent in Opleiding, respectievelijk Onderzoeker in Opleiding: afgestudeerden die aan een proefschrift schrijven.), en bursalen. Dergelijke opleidingsprogramma's kunnen de vorm hebben van cursussen waaraan promovendi uit de aangesloten universiteiten deelnemen. Een onderzoekschool kan zowel binnen een universiteit worden opgericht als tussen meer universiteiten, soms ook samen met andere onderzoeksinstellingen, bijvoorbeeld TNO.

In Vlaanderen kan men Taal- en Spraaktechnologie studeren als postgraduaatstudie, zoals de *Master of Artificial Intelligence*[114] opleiding in Leuven. Aspecten van taal- en spraaktechnologie komen ook aan bod in de GGS Taalwetenschap (interuniversitair) en de GAS (Toegepaste) Informatica, oriëntatie Computerlinguïstiek.[115] Geen van de hier vermelde opleidingen is vergelijkbaar met de opleidingen in Nederland. Daarnaast is het soms mogelijk zich door middel van TST-keuzevakken gedurende de ingenieurs- of licentiaatstudie iets te specialiseren in spraak- of taaltechnologie (bijvoorbeeld zwaartepunt Taaltechnologie, Germaanse UIA of de module Taaltheorie en Computerlinguïstiek, Germaanse Leuven). In al die gevallen gaat het om een relatief beperkt aanbod van vakken.

Naast de universitaire opleiding is er dit jaar ook een specialisatiejaar *Taal en Informatica* (deeltijd-opleiding, 1 jaar) gestart door de Katholieke Hogeschool Zuid-West-Vlaanderen (KATHO), in samenwerking met Flanders Language Valley. Het onderdeel *Computerlinguïstiek* wordt door CLIF verzorgd.

2.2.3 Industrie

De industrie vervult een rol zowel met betrekking tot het onderwijs als met betrekking tot het onderzoek. Wat het onderwijs betreft gaat het vooral om het aanbieden van stageplaatsen. Daarnaast kan de industrie ook een grote rol vervullen met betrekking tot het aanbieden van het onderwijs zelf. Een treffend voorbeeld is de opleiding die sinds vorig jaar wordt aangeboden door de Katholieke Hogeschool Zuid-West-Vlaanderen (KATHO), in samenwerking met Flanders Language Valley[116].[117] Daarnaast speelt de industrie idealiter een rol bij het invullen van de grote lijnen van de TST-opleidingen. Ook met betrekking tot het onderzoek speelt de industrie een rol, als vragende partij ten opzichte van de universiteiten of, met hun eigen onderzoeksafdeling, als partner van die universiteiten in (nationale of Europese) onderzoeksprojecten.

Hieronder volgen een aantal van de bedrijven die in dit opzicht een rol spelen in Nederland en Vlaanderen:

[111] http://candl.let.ruu.nl/Teaching/Dik/main.htm
[112] http://www-uilots.let.ruu.nl/~Hans.Leidekker/gids/tsa.html
[113] http://www.phil.uu.nl/onderwijs/cki/index.html
[114] http://www.cs.kuleuven.ac.be/cwis/education/MAI/
[115] http://www.kuleuven.ac.be/onderwijs/oo_nl.html
[116] http://www.westhoek.be/algemeen/flv.htm
[117] FLV is specifiek bestemd voor bedrijven actief in de taal- en spraaktechnologie. Een van de stuwende krachten achter dit project, en achter de opleiding die door KATHO wordt aangeboden, is Lernout & Hauspie.

- CAP GEMINI[118]
- COMSYS[119]
- GETRONICS[120]
- FLUENCY[121] Speech Communications
- LANGUAGE AND COMPUTING[122]
- LANT[123] Machine Translation Systems
- LERNOUT & HAUSPIE[124]
- KPN[125] Research
- MEDIALAB[126]
- OCE[127] Technologies
- PHILIPS[128] Dictation Systems
- POLDERLAND[129]
- POLYDOC[130]
- SENTIENT MACHINE RESEARCH[131]
- SYLLOGIC[132]
- TELECATS[133]

2.2.4 Beleid

In Nederland wordt het beleid op het gebied van TST voornamelijk bepaald door

- Ministerie van Onderwijs, Cultuur en Wetenschappen, afdeling OWB (Onderzoek en Wetenschapsbeleid)

[118] http://tools.capgemini.nl
[119] http://www.comsys.nl/
[120] http://www.getronics.nl/
[121] http://www-uilots.let.ruu.nl/~Arthur.Dirksen/fluent/fluent.htm
[122] http://www.landc.be
[123] http://www.lant.be
[124] http://www.lhs.com/
[125] http://www.kpn.com/research/nl/schema.html
[126] http://www.medialab.nl/
[127] http://www.oce.nl/
[128] http://www.research.philips.com/
[129] http://polder.ubc.kun.nl/
[130] http://www.aex.nl/finance/alg/poldalg.html
[131] http://www.smr.nl/
[132] http://www.syllogic.nl/
[133] http://www.telecats.nl/

- Ministerie van Economische Zaken
- NWO[134], Nederlandse Organisatie voor Wetenschappelijk Onderzoek

Onder NWO hebben een aantal onderdelen[135] te maken met TST:

- TAAL, SPRAAK EN LOGICA, een van de clusters uit het gebied Geesteswetenschappen
- WSA[136] (Wetenschappelijk Statistisch Agentschap), een agentschap onder het Gebiedsbestuur voor de Maatschappij- en Gedragswetenschappen
- STW[137] (Stichting voor de Technische Wetenschappen), een zelfstandig onderdeel van NWO.[138] Het STW vormt de divisie Technische Wetenschappen.

In Vlaanderen zijn bij TST de volgende instanties betrokken:

- Ministerie van de Vlaamse Gemeenschap/Wetenschappelijk onderwijs
- Ministerie van de Vlaamse Gemeenschap/Admistratie Wetenschap en Innovatie (AWI)
- Kabinet van minister-president Van den Brande (Wetenschapsbeleid en Technologie)
- FWO-Vlaanderen[139] (Fonds voor Wetenschappelijk Onderzoek-Vlaanderen)
- IWT[140] (Vlaams Instituut voor de Bevordering van het Wetenschappelijk-Technologisch Onderzoek in de Industrie)

Het FWO stimuleert en financiert het fundamenteel wetenschappelijk onderzoek aan de universiteiten in de Vlaamse Gemeenschap en aan de instellingen voor wetenschappelijk onderzoek. Het stelt zich in dit opzicht strikter op dan de Nederlandse zuster-organisatie, die bijvoorbeeld het PRIORITEIT- programma Taal- en Spraaktechnologie heeft gefinancierd. Het Vlaamse Korte termijn programma ter zake is op ad hoc basis door het IWT begeleid.
Op Vlaams-Nederlands gebied zijn dan nog actief:

- NTU[141] (Nederlandse Taalunie)
- CLVV, de Commissie Lexicale Vertaalvoorzieningen
- COTERM (terminologie)

[134]`http://www.nwo.nl/`
[135]Sinds 1 augustus 1998 zijn de werkzaamheden van SION (Stichting Informatica-Onderzoek in Nederland) overgenomen door NWO, Gebiedsbestuur Exacte Wetenschappen.
[136]`http://129.125.158.28/`
[137]`http://www.stw.nl/`
[138]Het STW wordt mede gefinancierd door het Ministerie van Economische Zaken.
[139]`http://www.nfwo.be/`
[140]`http://www.iwt.be/`
[141]`http://www.taalunie.org/`

- VNC, het Vlaams-Nederlands Comité voor Nederlandse Taal en Cultuur.

De hier genoemde instanties laten ieder onderzoek op hun specifieke gebied uitvoeren door derden.

Op Europees vlak zijn de programma's met betrekking tot TST (MLIS[142], LE[143], ESPRIT[144]) ondergebracht bij het directoraat DG XIII van de Europese Commissie. Daarnaast is er nog het EUREKA-initiatief[145] dat '*market-driven*' onderzoek en ontwikkeling stimuleert. De fondsen komen van de nationale overheden, niet van de EG.

2.2.5 Buitenland

In een aantal landen levert de overheid grote inspanningen ten behoeve van een goede infrastructuur voor Taal- en Spraaktechnologie.
In Duitsland loopt al sinds 1993 (tot 2000) het megaproject VERBMOBIL[146]. In Denemarken en Griekenland zijn nationale centra voor Taal- en Spraaktechnologie opgericht, het Center for Sprogteknologi[147] (CST) in Kopenhagen en het Institute for Language and Speech Processing[148] (ILSP) in Athene. Zij krijgen een ruime betoelaging van de nationale overheid, althans gedurende de eerste jaren. Naast het CST is er in Denemarken ook nog de DSN[149] (Dansk Sprognævn). Deze instantie houdt zich bezig met taalplanning, terwijl het CST zich bezighoudt met taaltechnologie.
Ook in Spanje is er een instituut opgericht, het *Observatorio Español de Industrias de la Lengua*[150] (OEIL) door het Ministerio de Industria y Energía. Dit instituut in Madrid is onderdeel van het Instituto Cervantes en moet zorgen dat alle noodzakelijk basisvoorzieningen aanwezig zijn, het publiek voorlichten, TST-producten promoten, en contact houden met allerlei Europese initiatieven. Dit instituut verschilt van de beide hiervoor genoemde doordat het zelf geen onderzoeksprojecten uitvoert. Een instituut van weer een iets ander type is het *Research Institute for the Languages of Finland*[151] (RILF). Het is opgericht in 1976 om het beheer van een aantal reeds bestaande instituten te coördineren en te centraliseren. Het houdt zich momenteel bezig met onderzoek naar alle talen die in Finland worden gesproken, met de financiering van dergelijk onderzoek door derden, met het beheer van archieven/databestanden (al dan niet in elektronisch formaat) en met taaladviezen. Het RILF maakt deel uit van het Ministerie van Onderwijs. Voor de Franstalige landen zijn een aantal instanties[152] actief, waaronder AUPELF UREF[154] (Association des universités partiellement ou entièrement de langue française - Université des réseaux d'expression française) en RIOFIL[155] (Réseau International des Observatoires Francophones de l'Inforoute et du Traitement Informatique

[142]`http://www2.echo.lu/mlis/`
[143]`http://www2.echo.lu/langeng/en/lehome.html`
[144]`http://www.cordis.lu/esprit/home.html`
[145]`http://www.eureka.be/`
[146]`http://www.dfki.de/verbmobil/`
[147]`http://cst.ku.dk/general/enframeset.html`
[148]`http://www.ilsp.gr/GBINDEX.HTM`
[149]`http://www.dsn.dk/`
[150]`http://www.cervantes.es/internet/acad/oeil/mar_oeil.htm`
[151]`http://www.domlang.fi/english.html`
[152]Zie de webpagina's van INSTITUTIONS FRANÇAISES DE LA FRANCOPHONIE[153].
[154]`http://www.refer.fr/liban_ct/general/agence.htm`
[155]`http://www.riofil.org/index.html`

des Langues).[156] Meer op de Romaanse talen in het algemeen gericht is er de UNION LATINE[158].

In het buitenland zijn er een aantal grote organisaties die zich bezig houden met het verspreiden van (linguïstische) hulpbronnen, waaronder:

- BAS[159] (Bavarian Archive for Speech Signals), München. Het BAS heeft tot taak databases gesproken Duits op optimale (gestandaardiseerde) wijze beschikbaar te maken voor zowel wetenschap als industrie.
- CSLU[160] (Center for Spoken Language Understanding), Oregon. Het CSLU verzamelt en verspreidt spraakcorpora voor alle geïnteresseerden. Voor universiteiten zijn de corpora gratis beschikbaar.
- DEUTSCHE SPRACHARCHIV[161] (DSAv), Mannheim. Er worden 32 Duitstalige corpora beheerd, slechts een deel daarvan is voor externe onderzoeksdoeleinden beschikbaar (onder meer op juridische gronden).
- ELRA[162] (European Language Resources Association), Parijs. Naast het verzamelen en verspreiden van resources (zowel taal, spraak als terminologie) beschouwt ELRA vooral ook het valideren van resources als een belangrijke taak.
- LDC[163] (Linguistic Data Consortium), Pennsylvania. Anders dan ELRA produceert het LDC ook zelf corpora, databanken, lexica en andere hulpbronnen voor zowel onderzoek als ontwikkeling van producten.
- SPRÅKBANKEN[164] (Bank of Swedish) , Göteborg. De Språkbanken stelt zich tot taak linguistische data in machinaal leesbare vorm te verzamelen.
- ICAME Corpus Collectie[165]. Een verzameling corpora, vooral Engelstalige.

Instanties als ELRA, LDC, BAS zijn voor een groot deel afhankelijk van financiering door de overheid (ministeries, onderzoeksfondsen, Europese fondsen).

Naast instellingen die het verspreiden van hulpbronnen tot doel hebben, zijn er ook een aantal die vooral de eigen hulpbronnen ter beschikking stellen. Een belangrijke is:

- BRITISH NATIONAL CORPUS[166] (BNC). Een collectie van 100 miljoen woorden hedendaags Engels, zowel gesproken als geschreven

[156]Onder auspiciën van deze laatste instantie is een overzicht gemaakt van hetgeen er in Frankrijk op TST-gebied voorhanden is: *À la découverte de l'ingénierie linguistique en France*[157].
[158]http://www.vol.it/linguanet/HomePages/agilit_unione_latina.htm
[159]http://www2.phonetik.uni-muenchen.de/Bas/
[160]http://www.cse.ogi.edu/CSLU/corpora/corpora.html
[161]http://www.ids-mannheim.de/prag/dsav.html
[162]http://www.icp.grenet.fr/ELRA/home.html
[163]http://www.ldc.upenn.edu/ldc/noframe.html
[164]http://svenska.gu.se/lb/lbinfoeng.html
[165]http://www.hd.uib.no/corpora.html
[166]http://info.ox.ac.uk/bnc/

Wat de TST-opleidingen betreft, zijn er momenteel een aantal interessante ontwikkelingen, waaronder, binnen het SOCRATES programma, die voor een European Masters in Language and Speech, i.e. voor een opleiding die èn Taaltechnologie èn Spraaktechnologie omvat. Het gaat hier om een opleiding die in vele landen gevolgd kan worden, en die door gezaghebbende, internationale organisaties als ESCA[167] en EACL[168] zou moeten worden erkend.
Belangrijke netwerken voor TST zijn

- ELSNET[169] (Network of Excellence in Language and Speech)
- COMPULOGnet[170] (Network of Excellence in Computational Logic)
- I3NET[171] (European Network for Intelligent Information Interfaces)
- THEMATIC NETWORK IN SPEECH COMMUNICATION SCIENCES[172]
- ACO*HUM[173] (Thematic Network in Advanced Computing in the Humanities)
- THEMATIC NETWORK IN THE AREA OF LANGUAGES[174]

De drie eerste zijn ESPRIT netwerken, de andere SOCRATES netwerken. Er worden initiatieven ontplooit die op termijn mogelijk gevolgen hebben voor de TST-infrastructuur in het Nederlandse taalgebied. Ook zouden deze netwerken kunnen optreden als partners bij het aanvragen van projecten, bijvoorbeeld voor een basiscollectie taaltechnologische hulpmiddelen. Het BLARK (Basic LAnguage Resource Kit) initiatief van ELSNET en ELRA is hiervan een voorbeeld. Het idee hier is dat er onder het Vijfde Framework Programma voor gezorgd zou moeten worden dat men in alle Europese (EG en CEE) landen kan beschikken over een een minimale set hulpmiddelen, voorlopig gedefinieerd als een algemeen *tekst*corpus om alle soorten precompetatief onderzoek te kunnen verrichten, met een omvang van ongeveer 10 miljoen woorden, geannoteerd volgens een algemeen geaccepteerde standaard, iets soortgelijks voor een *spraak*corpus, en een collectie *tools* om met deze corpora om te kunnen gaan.
Andere internationale organisaties waarin mensen die werkzaam zijn in de taal- en spraaktechnologie zich hebben verenigd:

- ACL[175] (Association for Computational Linguistics)
- EACL[176] (European Chapter of the Association for Computational Linguistics)
- EAMT[177] (The European Association for Machine Translation)
- ESCA[178] (European Speech Communication Association)

[167]http://ophale.icp.grenet.fr/esca/esca.html
[168]http://issco-www.unige.ch/eacl/eacl.html
[169]http://www.elsnet.org/
[170]http://www.compulog.org/
[171]http://www.i3net.org/
[172]http://tn-speech.essex.ac.uk/tn-speech
[173]http://www.uib.no/acohum
[174]http://www.userpage.fu-berlin.de/~elc/ThematicNetworkProject/tnp-fram.htm
[175]http://www.cs.columbia.edu/~acl/
[176]http://issco-www.unige.ch/eacl/eacl.html
[177]http://www.lim.nl/eamt/
[178]http://ophale.icp.inpg.fr/esca/esca.html

- EURALEX[179] (European Association for Lexicography)
- FOLLI[180] (European Association for Logic, Language and Information)

Binnen ACL zijn er een aantal voor TST relevante SIGs (Special Interest Groups):

- SIGDAT[181] (Special Interest Group for linguistic data and corpus-based approaches to natural language processing)
- SIGLEX[182]
- SIGMEDIA[183] (Special Interest Group on Multimedia Language Processing)
- SIGNLL[184] (Special Interest Group on Natural Language Learning)
- SIGPHON[185] (Special Interest Group for Computational Phonology)

Nog zo'n SIG, maar dan door de EU gefinancierd is EAGLES[186] (Expert Advisory Group on Language Engineering Standards)

Een elektronische nieuwsbrief voor taal, spraak en logica wordt verspreid door COLIBRI[187], informatie over nieuwe TSTpublicaties (al dan niet officieel) is te vinden in CMP-LG[188] (Computation and Language E-Print Archive).

Tot slot nog enkele (organisatoren) van de belangrijkste conferenties en zomerscholen op TST-gebied:

- COLING[189] (International Conference on Computational Linguistics)
- ACL[190]
- EACL[191]
- ESCA[192] (European Speech Communication Association)
- ESSLLI[193] (European Summer Schools in Logic, Language and Computation)
- ELSnet[194]

[179] http://www.ims.uni-stuttgart.de/euralex/
[180] http://www.wins.uva.nl/research/folli
[181] http://www.cis.upenn.edu/~yarowsky/sigdat.html
[182] http://www.cis.upenn.edu/~mpalmer/sigpoint/index.html
[183] http://www.dfki.uni-sb.de/~andre/sigmedia/index.html
[184] http://www.cs.rulimburg.nl/~antal/signll/signll-home.html
[185] http://www.cogsci.ed.ac.uk/sigphon/
[186] http://www.ilc.pi.cnr.it/EAGLES/home.html
[187] http://colibri.let.ruu.nl/
[188] http://xxx.lanl.gov/cmp-lg/
[189] http://www.dcs.shef.ac.uk/research/ilash/iccl/index.html
[190] http://www.cs.columbia.edu/~acl/
[191] http://issco-www.unige.ch/eacl/eacl.html
[192] http://ophale.icp.inpg.fr/esca/esca.html
[193] http://www.wins.uva.nl/research/folli/
[194] http://www.elsnet.org/

Hoofdstuk 3

Evaluatie

3.1 Vooraf

Eerder werk

Het evalueren van software en producten die gebruikt worden in de taal- en spraaktechnologie is het onderwerp geweest van een aantal Europese projecten. Ook buiten Europa is aandacht besteed aan dit onderwerp. We noemen hier als bronnen van informatie EAGLES[1] (een project gericht op het beschikbaar maken van resources, het vaststellen van standaards en richtlijnen voor *resources*, en evaluatie. Met name het werk van de *spoken language* werkgroep (Gibbon, Moore, en Winski 1997)[2] verdient vermelding). Daarnaast zijn van belang: TSNLP[3] (*test suites for natural language processing* gericht op het evalueren van prestaties van software), en de *Survey of the State of the Art in Human Language Technology* (Cole, Mariani, Uszkoreit, Zaenen, en Zue 1998)[4] (waarin hoofdstuk 13 gewijd is aan evaluatie, met name gericht op software) ELSE[5], en GRACE[6] (beide met name gericht op het evalueren van *part-of-speech taggers*), en TEMAA[7] (gericht op de evaluatie van *authoring tools*: hulpmiddelen voor spelling- en grammatica-correctie). In vrijwel alle gevallen betreft het omvangrijke projecten, soms gericht op slechts een deel van de hulpmiddelen en software die in het voorafgaande is genoemd. Voorzover er daadwerkelijke evaluaties zijn uitgevoerd heeft het Nederlands overigens in geen van deze projecten, voor zover wij weten, een rol gespeeld.

Evaluatievormen

Evaluatie van corpora en lexica dient vooral gericht te zijn op de vraag voor welk doel een corpus of lexicon geschikt is. Gezien vanuit het perspectief van dit rapport is vooral de vraag van belang welke informatie niet of onvoldoende gerepresenteerd is in de verzameling van corpora en lexica zoals ze voor het Nederlands beschikbaar zijn.

[1] http://www.ilc.pi.cnr.it/EAGLES/home.html
[2] http://coral.lili.uni-bielefeld.de/~gibbon/EAGLES/
[3] http://clwww.essex.ac.uk/group/projects/tsnlp/
[4] http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html
[5] http://www2.echo.lu/langeng/en/le4/else/summary.html
[6] http://www.ciril.fr/~pap/grace.html
[7] http://www2.echo.lu/langeng/en/lre2/temaa.html

In plaats van een onderlinge vergelijking van materialen zijn we dus in de eerste plaats geïnteresseerd in de verzameling als geheel. Bij het evalueren van lexica en corpora zullen vooral de volgende criteria een rol spelen:

- Welke omvang heeft het materiaal?
- Welke taalkundige (fonologische, morfologische, syntactische, semantische) informatie is gecodeerd?
- Voor welk doel is het materiaal gemaakt (vertalen, frequentiegegevens, *information retrieval*, etc.)?
- Welke standaards worden gebruikt?
- In hoeverre is het materiaal bruikbaar in combinatie met andere hulpmiddelen?
- Hoe goed is het materiaal gedocumenteerd?

Bij het vergelijken van *tools* (software-modules zoals spraakherkenners, parsers en *taggers*) is vooral de relatieve prestatie van een product van belang. Dit is de meest gangbare vorm van evaluatie, bijvoorbeeld binnen de information retrieval (en waarvoor speciale evaluatierondes worden georganiseerd waarvan verslag wordt gedaan op conferenties als TREC[8] (*text retrieval conference*), en MUC[9] (*message understanding conference*), en binnen het ARPA programma (*Advanced Research Projects Agency*) (met name gericht op spraakherkenning en gesproken dialoogsystemen). Dergelijke evaluatiemethoden zijn niet alleen nuttig voor potentiële gebruikers van software, maar zijn ook een bron van informatie voor ontwikkelaars van software. Door systemen met elkaar te vergelijken kunnen de taken die voor iedereen moeilijk zijn geïdentificeerd worden, en kunnen de relatieve prestaties van verschillende benaderingen vergeleken worden. De voortdurende verbetering van de prestaties van spraakherkenning is bijvoorbeeld voor een niet gering deel te danken aan het feit dat objectieve criteria en vergelijkingsmethoden bestaan (en het feit dat we zeker weten dat deze producten beter worden is hier natuurlijk ook aan te danken). Daarnaast spelen de uitkomsten van deze evaluaties ook een niet te onderschatten rol bij het verwerven van fondsen voor verder onderzoek. Een vereiste voor dergelijke evaluaties is evenwel de beschikbaarheid van testdata in de vorm van (foutloos) geannoteerde corpora en testsuites. Dergelijke testsuites zijn, met uitzondering van het Engels, voor de meeste talen niet beschikbaar.

Een laatste vorm van evaluatie betreft het onderzoeken in hoeverre een toepassing een bepaalde taak naar behoren uitvoert. Dergelijke evaluaties zijn vergelijkbaar met een soort 'consumententest', en leveren meestal een winnaar op die als 'beste koop' gekarakteriseerd wordt. Deze vorm van evaluatie is met name geschikt voor complete systemen. Zo zou men bijvoorbeeld de afbreekroutines en spellingcorrectie (en eventueel grammaticacorrectie) van verschillende tekstverwerkers met elkaar kunnen vergelijken, en daarbij, naast precisie, ook de vraag of zinvolle suggesties voor verbetering worden gegeven en een criterium als gebruiksgemak mee kunnen nemen. Merk op dat dergelijke criteria niet objectief meetbaar zijn. Het belang van deze vorm van evaluatie is daarom sterk afhankelijk van de autoriteit van de uitvoerende instantie.

[8] `http://trec.nist.gov/`
[9] `http://www.tipster.org/muc.htm`

Werkwijze

Een uitvoerige evaluatie van beschikbare hulpmiddelen was binnen het kader van dit project niet mogelijk. De beschikbare documentatie is vaak gering, en veel van de genoemde materialen waren niet voor ons beschikbaar, of konden we slechts oppervlakkig inspecteren. Daar staat tegenover dat we uitvoerig met experts gesproken hebben (met name ook over de vraag wat hun oordeel is over de beschikbaarheid, bruikbaarheid, en kwaliteit van hulpmiddelen), en dat de interviews die we hebben afgenomen een groot aantal evaluatieve opmerkingen over bepaalde producten en hulpmiddelen bevatten.

Gezien het doel van dit onderzoek richten we ons hieronder vooral op de vraag wat er aan hulpmiddelen op het gebied van lexica en corpora beschikbaar is en hoe nuttig deze hulpmiddelen zijn voor het uitvoeren van een bepaalde taak. Het resultaat van deze vorm van evaluatie is dat men inzicht krijgt in de vraag welke vormen van onderzoek en productontwikkeling op basis van de huidige stand van zaken haalbaar zijn. We proberen dus een antwoord te geven op de vraag of de hulpmiddelen die nodig zijn voor de ontwikeling van bijvoorbeeld een automatisch vertaalsysteem, een *document retrieval* systeem, of een tekst-naar-spraak systeem, voorhanden zijn.

Bij de evaluatie van software moeten we zo mogelijk nog terughoudender zijn. Ten eerste is software, in de vorm van bijvoorbeeld grafeem-naar-foneem conversie programma's, *taggers*, of robuuste (*wide coverage*) grammatica's, schaars. Ten tweede ontbreekt het aan testmateriaal met behulp waarvan de prestaties van deze producten gemeten zouden kunnen worden. We geloven dat deze twee observaties niet los van elkaar staan. De ontwikkeling van software(-modules) op het gebied van taal- en spraaktechnologie die van algemeen nut zijn (dus ook buiten de specifieke context van het project waarbinnen ze worden ontwikkeld) is vrijwel onmogelijk wanneer er geen trainings- en testmateriaal beschikbaar is dat representatief is voor een bepaalde taak. Dergelijke *testsuites* ontbreken voor het Nederlands volledig.

3.2 Tekstcorpora

De corpora van het INL zijn in potentie ongetwijfeld de belangrijkste bronnen van informatie over taalgebruik, niet alleen door hun omvang, maar ook door het feit dat ze evenwichtig zijn samengesteld, voor een groot deel zijn voorzien van (ongecorrigeerde) annotatie, en worden onderhouden.

Daar staat tegenover dat vrijwel alle geïnterviewden die de corpora van het INL noemden, daarbij aantekenden dat de corpora slecht toegankelijk zijn. Momenteel zijn de corpora slechts via een telnet verbinding raadpleegbaar (met uitzondering van een vijf miljoen corpus op de ECI-MCI CD-ROM). Het raadplegen van het corpus gebeurt via een *query*-programma. Voor onderzoekers die in de eerste plaats een taalkundige belangstelling hebben, en die dus voornamelijk op zoek zijn naar voorbeelden en naar tamelijk oppervlakkige kwantitatieve gegevens, is het gebruik van een *query*-programma waarschijnlijk een nuttig hulpmiddel, omdat het het zoeken in de corpora vergemakkelijkt, zonder dat er programmeerkennis vereist is. Toch heeft zo'n programma ook nadelen. Ten eerste is het huidige programma lastig te bedienen, zeker wanneer men gebruik maakt van gebrekkige terminal-emulatie (waardoor veel functietoetsen niet overeenstemmen met de documentatie). Een web-gebaseerd programma (waarvan inmiddels voorbeelden bestaan) zou dit probleem kunnen oplossen. Ten tweede maakt het *query*-programma het lastig om de data automatisch te raadplegen, of om statistische gegevens te verzamelen die niet reeds in het programma voorzien zijn. Dit nadeel wreekt zich vooral bij computationeel taalkundig onderzoek. Het gebruik van corpora vereist hier vrijwel altijd dat de onderzoeker zelf kan bepalen hoe het corpus wordt doorzocht, en in welk formaat de resultaten van een zoekopdracht worden getoond, opgeslagen, of doorgegeven aan een ander programma. De beperkingen die een *query*-programma met zich mee brengt zijn daarom voor computationeel taalkundig onderzoek te knellend.

Een probleem van geheel andere orde is dat de INL-gegevens alleen voor niet-commercieel gebruik bedoeld zijn. Gezien de afspraken die het INL heeft gemaakt met de leveranciers van de data is het niet eenvoudig hiervoor een algemene oplossing te vinden.

Een tweede corpus dat vaak genoemd wordt is het Eindhoven-corpus. Als tekstcorpus is het van vrij geringe omvang, maar niettemin nuttig. Het grootste probleem is dat er geen instantie is die het beheer en de distributie van dit corpus voor haar rekening neemt, zodat onduidelijk is in hoeverre het corpus gedistribueerd mag worden.

Een derde corpus dat van belang lijkt is het ANNO-corpus (640.000 woorden, voorzien van woordsoort). Ook dit is een relatief klein corpus, maar het is wel voorzien van woordsoort. De onderhandelingen over de vrijgave van dit corpus met de instantie die de teksten leverde (BRTN) zijn nog gaande.

Voor de toekomst lijkt het PAROLE-corpus (dat op het INL wordt ontwikkeld in het kader van een Europees project) van belang (totale omvang twintig miljoen woorden, waarvan drie miljoen beschikbaar op CD-ROM, 250.000 woorden voorzien van woordsoort). Daarnaast kan het Corpus Gesproken Nederlands ook als tekstcorpus van nut zijn.

Tenslotte mag men verwachten dat een corpus volgens bepaalde, algemeen aanvaarde, standaards gecodeerd is en is voorzien van annotatie. Opvallend is dat er onder de geïnterviewden een zekere luchtigheid bestaat ten aanzien van standaards en coderingsconventies. Over het algemeen stelt men zich op het standpunt dat corpora (en woordenboeken) op een consistente manier samengesteld en gecodeerd moeten zijn, dat er sprake moet zijn van een instantie waar men met vragen terecht kan en die het noodzakelijke onderhoud pleegt, en dat het materiaal onder duidelijke voorwaarden beschikbaar moet zijn.

Conclusie

Voor corpus-gebaseerd onderzoek geldt dat men nooit genoeg data heeft. Vanuit dit perspectief gezien is er niet bijster veel beschikbaar. Naast de ECI-MCI CD-ROM (vijf miljoen woorden ruwe tekst) is er eigenlijk niets dat gemakkelijk verkrijgbaar of toegankelijk is. Niet alleen is de hoeveelheid beschikbare tekst gering, *de facto* zijn er geen corpora verkrijgbaar die zijn voorzien van woordsoort, om nog maar te zwijgen van corpora waarin rijkere vormen van annotatie (m.n. syntactische en semantische) zijn aangebracht. Hetzelfde geldt voor parallelle corpora, waarvan er momenteel slechts één beschikbaar is. De corpora die wellicht in de nabije toekomst beschikbaar komen of die in voorbereiding zijn, zullen in deze situatie maar weinig verandering kunnen brengen.

Een actie gericht op het beschikbaar maken van omvangrijke, geannoteerde, tekstcorpora, lijkt daarom zeker gerechtvaardigd. Tijdens de interviews zijn verschillende suggesties gedaan voor het verkrijgen van materiaal: de omroep, het Meertens Instituut (voor dialectologie), en de Verenigde Naties. Daarnaast lijkt de overheid zelf een potentiële leverancier van tekstdata.[10] De Taalunie zou een nuttige rol kunnen spelen bij het benaderen van leveranciers (die individuele onderzoekers of onderzoeksinstituten vaak niet als de juiste gesprekspartners zien) en het beheer van de data.

3.3 Spraakcorpora

Onderzoek op het gebied van spraakherkenning is in zeer sterke mate afhankelijk van corpora. Een aantal corpora zijn via ELRA beschikbaar, zoals het Polyphone-NL corpus. Het Vlaamse COGEN corpus zal in de nabije toekomst toegankelijk zijn. Met het opstarten van het project voor een Corpus Gesproken Nederlands een belangrijke stap gezet in de richting van een algemeen en omvangrijk corpus gesproken Nederlands. Dit corpus richt zich niet op specifieke toepassingen en is omvangrijk genoeg om een basis te vormen voor onderzoek op het gebied van spraakherkenning. Daarnaast zullen voor specifieke toepassingen altijd aanvullende corpora nodig zijn, waarin aan speciale eisen omtrent opname-condities en inhoud is voldaan. Voor een deel worden dergelijke corpora in Europees verband ontwikkeld (denk aan SPEECHDAT-CAR), voor een deel zal dit de verantwoordelijkheid blijven van de instellingen die betrokken zijn bij toegepast onderzoek.

Conclusie

Met het opstarten van het (Nederlands-Vlaamse) project voor een Corpus Gesproken Nederlands, waarbij ook de Taalunie als toekomstige beheersinstelling betrokken is, lijkt het momenteel niet noodzakelijk extra activiteiten op het gebied van gesproken corpora te ontwikkelen. In de interviews is er wel op gewezen dat dit corpus alleen maar van nut zal zijn voor de spraaktechnologie wanneer er voldoende afstemming is tussen de wensen van spraaktechnologen en de uitvoerders van het project, met name waar het gaat om codering, annotatie, en samenstelling van het corpus. Een effectieve manier om een dergelijke afstemming te bewerkstelligen is het opzetten van parallelle spraaktechnologische projecten waar, nog tijdens de looptijd van het project, gebruik wordt gemaakt van de voorlopige resultaten.

3.4 Lexica

Het meest genoemde hulpmiddel voor taaltechnologisch onderzoek is de lexicale database van CELEX. Deze database is een zeer nuttige bron van met name fonologische en morfologische informatie. Vrijwel zonder uitzondering is men positief over deze CD-ROM. Recentelijk is een versie van CELEX beschikbaar gemaakt die de nieuwe spelling bevat, en de FONILEX-database voorziet in Vlaams-gekleurde uitspraakgegevens. Binnen de taal- en spraaktechnologie is CELEX vooral gebruikt als hulpmiddel bij het

[10] Zo bestaat alleen al het rapport van de commissie Van Traa uit zo'n 5500 bladzijden, hetgeen ruwweg overeenkomt met 1,6 miljoen woorden.

ontwikkelen van afbreekroutines en grafeem-naar-foneem conversie. Voor spraaktechnologie kan verder gebruik worden gemaakt van het ONOMOMASTICA-woordenboek, dat de uitspraak van eigennamen en geografische aanduidingen bevat.

Naast CELEX is er weinig beschikbaar. Dit betekent dat voor die gebieden waaraan in CELEX geen of weinig aandacht is besteed (met name syntaxis (valentie) en semantiek) er concreet niets beschikbaar is. Er lopen momenteel wel verschillende projecten die wellicht in deze leemte zullen voorzien. Binnen het PAROLE-project ontwikkeld het INL een middelgroot woordenboek (20.000 trefwoorden) met o.a. valentie-informatie. Hetzelfde geldt voor het *Referentiebestand Nederlands* (RBN), dat in opdracht van de CLVV wordt ontwikkeld door een consortium bestaande uit de VU, het INL, de UU en de KU Leuven, en dat zich ook richt op het ontwikkelen van een woordenboek met o.a. valentie-informatie. EuroWordNet, tenslotte, richt zich op het ontwikkelen van een multilinguale conceptuele database (omvang 50.000 trefwoorden).

Tweetalige woordenboeken voor de economisch belangrijke talen zijn momenteel alleen beschikbaar bij Van Dale. Ondanks het feit dat het hier producten betreft die in de eerste plaats voor eindgebruikers bedoeld zijn, zijn er een aantal projecten waar men wel van deze woordenboeken gebruik heeft maakt (o.a. GLOSSER[11] en TWENTY-ONE[12]). Door de Commissie Lexicografische Vertaalvoorzieningen (CLVV) wordt gewerkt aan verschillende vertaalwoordenboeken, maar het betreft hier zonder uitzondering woordenboeken voor taalparen die commercieel niet interessant zijn. De terminologiedatabase van EURODICAUTOM, tenslotte, lijkt interessant, maar er zijn ons geen toepassingen van deze informatie voor TST (automatisch vertalen?, multilingual document retrieval?) bekend.

Conclusie

Het feit dat nu juist het meest succesvolle hulpmiddel voor taal- en spraaktechnologisch onderzoek (CELEX) een onzekere toekomst lijkt te hebben, is zorgelijk. Het lijkt zeker aan te bevelen te zoeken naar een constructie waarbij in ieder geval een minimale vorm van onderhoud en beheer van de data gewaarborgd is.[13]

Van Dale is in potentie een bron van zeer betrouwbare en uitgebreide lexicale informatie, maar stelt tot op heden haar data maar mondjesmaat beschikbaar. Wel lijkt men geïnteresseerd in mogelijkheden om de data in de toekomst op enigerlei wijze in te zetten bij de ontwikkeling van TST-producten.

In concreto zijn er momenteel geen lexica beschikbaar die voorzien in gedetailleerde syntactische en semantische informatie. Dit betekent dat de ontwikkeling van bepaalde toepassingen (met name *wide-coverage* grammatica's voor grammaticacorrectie, automatisch vertalen, dialoogsystemen, of IR) niet goed mogelijk zijn. Bij het beschikbaar komen van hulpmiddelen die in deze leemte zouden kunnen voorzien (PAROLE, RBN), maar die niet direct binnen de context van het TST-onderzoek zijn ontstaan, zou men nadrukkelijk aandacht moeten besteden aan de vraag hoe het materiaal beschikbaar moet worden gesteld. Ook zou onderzocht moeten worden of deze hulpmiddelen in alle behoeften voorzien, en of er nog aanvullende maatregelen, gericht op de ontwikkeling van syntactische en semantische lexicale databases, nodig zijn.

[11] http://www.let.rug.nl/~glosser
[12] http://twentyone.tpd.tno.nl/
[13] CELEX is voorlopig ondergebracht bij de KUN.

3.5 Overige hulpmiddelen

Er zijn weinig hulpmiddelen met een algoritmisch aspect (*tools*) voor het Nederlands beschikbaar. De enige sector waarbinnen verschillende hulpmiddelen genoemd worden, die ten dele ook on-line beschikbaar zijn, is de morfologie. Er zijn een aantal programma's die morfologische analyse uitvoeren en woordsoorten toekennen. Pogingen om te komen tot een algemene computationele grammatica en parser voor het Nederlands beperken zich tot CORRIE en AMAZON/CELEX. Daarnaast is wellicht binnen een enkel bedrijf iets beschikbaar (zoals de ROSETTA-parser van Philips en de METAL-parser van Siemens).

Op het gebied van standaards lijkt de DISC-notatie van CELEX de norm te zijn voor fonetische transcriptie. Verder valt op dat verschillende andere notaties die voor fonetische transcriptie gebruikt worden (SAMPA, YAPA) blijkbaar goed samengaan met DISC. Voor het annoteren van corpora met woordsoorten is de situatie minder duidelijk. Er zijn verschillende notaties in omloop, waarvan de WOTAN-tagset in ieder geval de meest genoemde is. Bij gebrek aan syntactisch en semantisch geannoteerde corpora kan niets gezegd worden over standaards voor dit niveau van annotatie.

Tenslotte is ons niets gebleken van algemeen aanvaarde en objectieve evaluatiecriteria die gebruik maken van *testsuites*, *tree-banks*, etc.

Conclusie

De situatie op het gebied van software is tamelijk zorgelijk. Op het gebied van morfologische analyse, tagging, en ook grafeem-naar-foneem conversie, bestaan er verschillende programma's, maar slechts enkele hiervan zijn gedocumenteerd en voor derden beschikbaar. Daarnaast is slechts een enkel programma gesignaleerd dat grammaticale analyse uitvoert.

Een verbetering van deze situatie kan eigenlijk alleen bereikt worden wanneer er corpora beschikbaar zijn die kunnen dienen als trainingsmateriaal en als testmateriaal.

Voor het vergelijken van de prestaties van verschillende part-of-speech taggers is het bijvoorbeeld dringend nodig dat er een standaard wordt ontwikkeld voor het annoteren van corpora en dat een aantal (gecorrigeerde) corpora beschikbaar komen waarin deze standaard wordt gehanteerd. Te verwachten valt dat met het opzetten van een dergelijke *testbench* verschillende groepen hun programma's beschikbaar zullen maken en dat, als gevolg van de mogelijkheid om resultaten te vergelijken, de foutenmarge van alle programma's zal verminderen.

Iets vergelijkbaars geldt voor grammaticale analyse. Om hier progressie te boeken dienen goede hulpmiddelen beschikbaar te zijn, zoals *tree-banks* (geannoteerd volgens een algemeen aanvaard schema), woordenboeken die voldoende syntactische informatie bieden, en formele beschrijvingen van de grammaticale regels van het Nederlands. Te verwachten valt dat met de beschikbaarheid van dergelijke hulpmiddelen de ontwikkeling van computationele grammatica's en parsers ook op gang zal komen.

Hoofdstuk 4

Interviews

4.1 Inleiding

In tabel 1 wordt een overzicht gegeven van de personen die we hebben gevraagd hun mening te gegeven over de positie van de taal- en spraaktechnologie voor het Nederlands, de kwaliteit en beschikbaarheid van hulpmiddelen, hun behoeften, de rol van de overheid en de Taalunie, en de wenselijkheid van een nieuwe instelling voor taalenspraaktechnologie. Ongeveer 80% van deze personen is door ons geïnterviewd, de rest reageerde per *e-mail*.

4.2 Toekomstige toepassingen

Er is een zeer grote concensus over toepassingen die in de toekomst van belang zullen zijn. Met name toepassingen waarbij spraaktechnologie een rol speelt (inclusief spraak-naar-spraak vertaling) en toepassingen die zich bevinden in het gebied van de *information/document retrieval/extraction* worden door een meerderheid van de respondenten genoemd. Daarnaast werd genoemd:

- auteursystemen voor SGML,
- proofing-tools (spelling- en grammatica-correctie),
- afbreken,
- *document-management* (*proofing*, en toepassingen op het gebied van *information* of *document retrieval*, ook multilinguaal),
- datamining (voor *document retrieval*, vertalen),
- '*knowledge management*' (*information retrieval* e.d.),
- terminologie,
- CALL (*computer-assisted language learning*),
- vertaalhulpmiddelen,
- samenvatten,
- natuurlijke taalinterfaces,
- spraakhulp,

Naam	**Instelling**	**NL/VL**	**u/b/o**	**e/i**
Prof. dr. G. Adriaens	Novell, CCL (KU Leuven)	VL	b/u	e
Dr. B. van Bakel	CTIT (UT)	NL	u	e
Dr. G. Bloothooft	OTS (RUU)	NL	u	i
Dr. W. Ceusters	Language and Computing	VL	b	i
Dr. V. Claes	RUCA	VL	u	i
Dhr. J. Colpaert	Didascalia	VL	b	i
Prof. dr. W. Decoo	Didascalia	VL	b	i
Prof. dr. Y. Dologlou	ESAT (KU Leuven)	VL	u	i
Drs. A Dijkstra	NWO	NL	o	i
Drs. P. van der Eijk	Cap Gemini	NL	b	i
Dr. H. van Halteren	Taal en Spraak (KUN)	NL	u	e
Drs. Th. van den Heuvel	Polderland	NL	b	i
Prof. dr. F. de Jong	CTIT (UT), TNO-TPD	NL	u/b	i
Dhr. P. van der Kamp	INL	NL/VL	u/o	i
Dr. U. Knops	LANT	VL	b	i
Drs. S. Krauwer	OTS (RUU)	NL	u,o	i
Dr. G. Kruyt	INL	NL/VL	u/o	i
Prof. ir. J. Landsbergen	OTS (RUU), IPO	NL	u/b	e
Prof. dr. J.-P. Martens	ELIS (U Gent)	VL	u	i
Prof. dr. W. Martin	Lexicografie (VU)	NL	u/o	i
Prof. dr. M. Moortgat	OTS (RUU)	NL	u	i
Dr. E. den Os	KPN	NL	b	i
Prof. dr. R. Scha	Afa-informatica (UvA)	NL	u	e
Drs. R. Piepenbrock	Celex (MPI/KUN)	NL	u/o	i
Drs. M. van Staden	TNO-STB	NL	o	i
Prof. dr. P. van Sterkenburg	INL	NL/VL	u/o	i
Dr. F. Steurs	KVH	VL	u	e
Prof. dr. F. Van Eynde	CCL (KU Leuven)	VL	u	i/e
Prof. dr. D. Van Compernolle	Lernhout & Hauspie, ESAT (KU Leuven)	VL	b,u	i
Dr. P. Vossen	Alfa-informatica (UvA)	NL	u	i
Prof. dr. D. Willems	Franse taalkunde (U Gent)	VL	u	i
Dr. J. Zuidema	Van Dale	NL	b	i

Tabel 4.1: Overzicht van respondenten. NL = werkzaam bij een Nederlandse instelling, VL = werkzaam bij een Vlaamse instelling, u = universitair, b = bedrijfsleven, o = overig (overheid, beleid, beheer van materialen), e = reactie per email, en i = interview. Er is één interview afgenomen met van Sterkenburg, Kruyt, en van der Kamp (INL) gezamenlijk. Hetzelfde geldt voor Decoo en Colpaert (Didascalia). Adriaens is geïnterviewd als vertegenwoordiger van Novell-Antwerpen. Dit bedrijf maakt intussen deel uit van Lernout & Hauspie.

- spraakherkenning (voor reisinformatie, *hands-free* bellen),
- spraaksynthese,
- sprekerverificatie,
- dicteersystemen (voor speciale beroepsgroepen),
- automatisch (spraakgebaseerd) vertalen,
- multimediale en multilinguale toepassingen,

4.3 Bestaande hulpmiddelen

De meeste respondenten vinden dat er weinig tot niets aan hulpmiddelen voor het Nederlands beschikbaar, respectievelijk gemakkelijk toegankelijk is. Het hulpmiddel dat veruit het meest genoemd wordt is CELEX. Bijna alle andere hulpmiddelen worden door slechts één of twee respondenten genoemd.

- CELEX,
- elektronisch groene boekje,
- CLVV woordenboeken,
- Referentiebestand Nederlands,
- EuroWordNet,
- FoniLex,
- Van Dale tweetalige woordenboeken,
- INL-corpora,
- Eindhoven-corpus,
- Parole,
- CoGen,
- ANNO,
- hulpmiddelen voor *information retrieval* van TNO en XEROX,
- ANS,
- CORRIE,
- MBROLA (spraaksynthese),
- Polyphone,
- ELRA-corpora (o.a. SPEECHDAT),
- SPRAAK (een programma voor de analyse van spraakdata, ontwikkeld door Paul Boersma, Fonetiek, UvA),

4.4 Het Nederlands in relatie tot andere talen

Bij de meeste respondenten bestaat de indruk dat er voor andere talen, en dan met name voor het Engels, aanzienlijk meer beschikbaar is dan voor het Nederlands. De vraag of het Nederlands er slechter voorstaat dan andere talen wanneer het Engels buiten beschouwing wordt gelaten wordt niet eenduidig beantwoord. Sommigen menen dat de situatie voor het Nederlands slechter is dan voor andere talen, terwijl anderen menen dat het Nederlands redelijk meekomt.

De meeste respondenten vonden het te ver gaan een uitputtend overzicht te geven van hulpmiddelen die voor andere talen bestaan, maar niet voor het Nederlands. Wel noemde men soms saillante voorbeelden:

- de Penn Treebank,
- vertaalprogramma's,
- spraakherkenning,
- tekst-naar-spraaksynthese, bijvoorbeeld voor *e-mail reading*,
- *learners' dictionary*,
- TSNLP (*testsuites*),
- XTAG grammatica project,
- Perseus-project (multimediale ontsluiting van Griekse en Latijnse teksten, geïntegreerd met woordenboeken en vertalingen),

4.5 Behoeften

Er blijkt een vrij algemene behoefte te bestaan aan grotere corpora, die rijk geannoteerd zijn. Daarnaast is er behoefte aan verschillende vormen van lexicale informatie. Een aanzienlijk aantal respondenten heeft behoefte aan meer formele en computationele beschijvingen en implementaties van de Nederlandse grammatica, hetzij als hulpmiddel bij het zelf maken van (domein- of applicatiespecifieke) *parsers*, hetzij als (modificeerbaar) onderdeel van een programma voor grammaticacorrectie, automatisch vertalen, *information retrieval*, etc. Een aantal respondenten wijst ook op het feit dat er niets beschikbaar is dat evaluatie van bestaande hulpmiddelen mogelijk zou maken.

Corpora

- Grotere corpora,
- Hele grote, rijk geannoteerde, corpora,
- Corpora met schrijffouten en correcties,
- Corpora voorzien van woordsoort,
- Corpora met gesproken dialogen,

- Corpora voorzien van topic-aanduiding (voor *information retrieval* e.d.),
- Grotere corpora met distributiegegevens gerelateerd aan domeingegevens,
- Verrijkte corpora,
- Parallelle corpora,
- (Parallel) corpus Nederlands-Vlaams,
- Gesegmenteerde (op fonetisch niveau) corpora (als basis voor spraaksynthese),
- Corpora met prosodische informatie,

Lexicale informatie

- Geformaliseerde, bilinguale, lexicale databases, gekoppeld aan monolinguale databases en thesauri,
- Equivalentierelaties voor alle betekenissen naar andere talen,
- Uitbreidingen van CELEX,
- Semantische hulpmiddelen (thesauri, semantische velden) voor *information retrieval*.
- Lexicons met volledige informatie m.b.t. spraak, morfologie, syntax, combinatoriek, semantiek, pragmatiek, gekoppeld aan corpora,
- Lexica met domeinmarkering,
- Lexica voor terminologie (ook specifiek voor Vlaanderen),

Halffabrikaten

- Het Groene Boekje als spellingchecker,
- 'NLP *middleware*' :modules zoals taggers, computationele grammatica's, (uitbreidbare) vertaalmodules, etc.
- '*Tunable NLE systemen*': systemen die gemakkelijk kunnen worden aangepast voor een specifieke toepassing,
- Vertaalsoftware,
- Grotere grammatica's,
- Beschrijvende elektronische grammatica,
- Formele elektronische grammatica,
- Veel preciezere versie van de ANS,
- Contrastieve Vlaams-Nederlandse grammatica,
- '*Large-coverage*' computationele grammatica, met informatie over fonologie, morfologie, syntaxis, en semantiek,
- Partiële en volledige constituentgrammatica's,
- Robuuste parser,
- Een groene grammatica, dan wel een door de Taalunie geautoriseerde *grammar checker*,

- Grammaticale (syntaxis en morfologie) en spraak-modules afgestemd op een groot woordenboek,
- Semantische *taggers* en *word-sense-disambiguation* hulpmiddelen,
- Semantische interpretatie van samengestelde woorden,
- Spraakhulpmiddelen,
- Tekst-naar-spraak module,
- Grafeem-naar-foneem conversie,
- Prosodie tool,
- Spraaksynthese,

Standaards en evaluatiemateriaal

- Trainingsmateriaal (geannoteerde corpora, *testsuites*).
- Testcorpora als de Penn Treebank,
- *Benchmarks*,
- Consensus over een EAGLES-compatibel, van de ANS afgeleid, annotatieschema, en bijbehorende software.

Overig

- Overzichten van beschikbaar materiaal,
- Software die draait onder DOS/Windows (en niet slechts op UNIX),
- hulpmiddelen voor *natural language processing* voor CALL,

Skeptische reacties

Respondenten gaven soms ook aan dat ze weinig belang hechten aan bepaalde hulpmiddelen:

- Minder behoefte aan corpora. (Voor specifieke applicaties of klanten is het nut van algemene corpora beperkt.)
- Het nut van rijker (syntactisch en semantisch) geannoteerde corpora is beperkt.

4.6 Basiscollectie

De meeste respondenten staan positief tegenover het ontwikkelen van een basiscollectie. Randvoorwaarden zijn dat het materiaal tegen een redelijke vergoeding beschikbaar moet zijn (ook voor het bedrijfsleven), en dat er garanties zijn voor continuïteit. Sommige respondenten benadrukken het belang van standaards (zoals EAGLES), terwijl anderen (met name in het bedrijfsleven) hier luchtiger over doen, en veronderstellen dat alles wat systematisch gedaan is nuttig zal zijn. Met name het bedrijfsleven verwacht dat met de beschikbaarheid van basisvoorzieningen de mogelijkheden om commerciele toepassingen te ontwikkelen zullen toenemen. Universitaire instellingen hebben vooral behoefte aan materiaal dat voor onderwijs en fundamenteel (*precompetitive*) onderzoek gebruikt kan worden.

Bij universitaire instituten bestaat ook de bereidheid mee te werken aan het opzetten van zo'n basiscollectie. Het bedrijfsleven ziet zichzelf in de eerste plaats als gebruiker van de basiscollectie, en lijkt ook niet bereid substantieel te investeren in de ontwikkeling hiervan.

De vraag wat deel zou moeten uitmaken van zo'n basiscollectie is niet systematisch aan de orde geweest. Een goede indruk van wat daar onder verstaan zou kunnen worden levert evenwel het overzicht van behoeften van de verschillende respondenten. Daarnaast leven er blijkbaar in de kringen van ELSNET ideeën voor de ontwikkeling van een *basic language resource kit* die voor iedere Europese taal beschikbaar zou moeten zijn.

Tenslotte benadrukken verschillende respondenten dat de overheid de ontwikkeling van dergelijke hulpmiddelen en *tools* niet aan het bedrijfsleven over moet laten. De overheid kan kwaliteitseisen stellen die wellicht verder gaan dan wat commerciële partijen als haalbaar beschouwen. Daarnaast wordt met het openbaar maken van hulpmiddelen zowel het fundamenteel onderzoek alsook het toegepaste (commerciële) onderzoek gestimuleerd.

4.7 Onderwijs en personeel

Men is over het algemeen somber over de mogelijkheden om gekwalificeerd personeel te vinden. Dit geldt vooral voor meer technisch onderlegde medewerkers. De laatste categorie vindt ook gemakkelijk in andere sectoren werk.

De instroom in (Nederlandse) letterenopleidingen is laag. Sommigen zien ruimte voor een meer interdisciplinaire benadering, waarbij ook op de middelbare scholen reeds de nodige bekendheid aan het vakgebied zou moeten worden gegeven. Anderen gaan nog verder, en menen dat taal- en spraaktechnologie vooral binnen de Informatica gelokaliseerd zou moeten worden. Voor Vlaanderen geldt dat specifieke TST opleidingen in de eerste en tweede cyclus ontbreken, en dat zelfs studierichtingen met voldoende aandacht voor formele taalkunde schaars zijn. Ook wat aanvullende of gespecialiseerde TST-studies betreft is het aanbod zeer schaars.

Daarnaast zou samenwerking tussen opleidingen het probleem van een lage instroom gecombineerd met een vrij uitgebreid vakgebied kunnen oplossen. Om de kwaliteit van de opleidingen te garanderen zou verder aansluiting bij bestaande (Nederlandse) 'kwaliteit en studeerbaarheids'-projecten en Europese initiatieven gezocht kunnen worden. Ook de samenwerking tussen Nederlandse en Vlaamse opleidingen zou verbeterd kunnen worden.

4.8 Een nieuwe beleidsinstelling

Er heerst een vrijwel algemene onvrede over het huidige beleid met betrekking tot TST. Men spreekt over gebrek aan coördinatie, en versnipperde initiatieven. Er is een behoefte aan een aanspreekpunt waar men op de hoogte is van beleidsinitiatieven, lopende projecten, subsidiemogelijkheden, en van bestaande hulpmiddelen.

In Nederland lijkt er vooral bij OCW (via onderzoeksfinancier NWO) wel aandacht voor TST te bestaan. Vanuit EZ is er weinig tot geen aandacht voor TST, alhoewel dat in het verleden wel het geval was (zie ook het Euromap-NL rapport (van Staden 1998)). Opvallend is dat er vanuit de technische hoek (TNO, Universiteit Twente) wel contacten met EZ lijken te bestaan.

In Vlaanderen speelt met name het IWT een rol bij het opstarten van TST-projecten. Van FWO (enkel ruimte voor fundamenteel onderzoek) verwacht men minder.

De roep om meer coördinatie, een betere samenwerking tussen bestaande instellingen, en een doorzichtiger beleid is vrij algemeen. Ook wordt er aangedrongen op het verruimen van de mogelijkheden om te komen tot Vlaams-Nederlandse samenwerking, bijvoorbeeld door structureel de mogelijkheid te bieden gezamenlijke TST-onderzoeksprojecten aan te vragen.

Voor wat betreft het beheer en onderhoud van hulpmiddelen wordt erop gewezen dat binnen het project voor een Corpus gesproken Nederlands een aantal van deze kwesties geregeld zal moeten worden, en dat men daarvan gebruik zou kunnen maken om te komen tot een meer structurele voorziening. Tegelijkertijd zou de rol van instanties als het INL, CELEX, SPEX, en, op het internationale vlak, ELRA bekeken moeten worden, die nu een deel van deze functie vervullen.

De meningen over het oprichten van een nieuw instituut zijn verdeeld. Enerzijds is men huiverig voor het oprichten van weer een nieuwe instantie. Men verwacht dat het mogelijk moet zijn om binnen de bestaande structuren te komen tot een beter beleid en betere samenwerking. Anderzijds wordt ook de situatie in Griekenland of Denemarken als voorbeeld genoemd, waar een nationale instelling verantwoordelijk is voor het beheer van TST-hulpmiddelen. Degenen die voor een nieuwe instelling zijn denken dan ook vooral aan een instituut dat het technisch beheer van hulpmiddelen op zich neemt, en niet direct aan een beleidsinstantie.

4.9 De rol van de Taalunie

De meningen over de mogelijke rol van de Taalunie op het gebied van TST zijn verdeeld. Als positieve punten worden genoemd het feit dat de Taalunie kan zorgen voor Vlaams-Nederlandse samenwerking, dat de Taalunie als belangenbehartiger van het Nederlands op kan treden en daarbij los staat van onderzoeksinstanties, dat de Taalunie kan zorgen voor meer structurele aandacht voor TST, en dat de Taalunie een rol kan spelen bij het oplossen van juridische kwesties. Als negatieve punten worden genoemd het feit dat de Taalunie te ver van het bedrijfsleven (de techniek, de markt) afstaat, dat de Taalunie te literair is en teveel een bewaker van het Nederlands is, en dat de Taalunie de juiste expertise mist.

Hoofdstuk 5

Aanbevelingen

Dit rapport is geschreven vanuit de gedachte dat het Nederlands een relatief kleine taal is en dat de ontwikkeling van een hoogwaardige infrastructuur voor taal- en spraaktechnologie in dit geval speciale aandacht vraagt.

Een overheid die het tot stand brengen van een dergelijke infrastructuur stimuleert door coördinerend op te treden en die, waar nodig, het daadwerkelijk tot stand brengen van hulpmiddelen of eindproducten voor TST ondersteunt, is hierbij noodzakelijk. In het inleidende hoofdstuk is reeds aangegeven dat communicatie in de eigen taal bijna altijd de voorkeur verdient boven communiceren in een andere taal. Naarmate informatie- en communicatietechnologie steeds geavanceerder wordt, neemt ook het belang van taal-en spraaktechnologie toe. Wanneer deze ontwikkeling voor bepaalde talen achterblijft, zal dat ertoe leiden dat die talen in toenemende mate in het defensief gedrongen worden. Een goede infrastructuur voor TST is daarmee dus in het algemeen belang.

Economische motieven alleen zullen er niet toe leiden dat TST-producten die zijn gebaseerd op het Engels ook automatisch voor het Nederlands op de markt zullen verschijnen. Bovendien bestaat er de vrees dat, voor zover Nederlandstalige producten wel verschijnen, de kwaliteit van deze producten achterblijft bij datgene wat voor andere talen mogelijk is. Europese programma's spelen een belangrijke rol bij het ontwikkelen van een infrastructuur voor TST. Dergelijke programma's gaan evenwel uit van het principe van subsidiariteit, en dus is er ook vanuit dit oogpunt gezien een rol weggelegd voor de nationale overheden. De situatie voor het Nederlands is tenslotte complex omdat het wordt gesproken in Nederland en Vlaanderen, en initiatieven van de overheid daarom ofwel slechts een deel van het taalgebied betreffen, ofwel speciale coördinatie vragen.

Naast een actieve rol van de overheid gaan we er bij deze aanbevelingen ook vanuit dat het wenselijk is dat hulpmiddelen voor TST zoveel mogelijk algemeen (tegen een redelijke vergoeding) beschikbaar zijn. Het zal duidelijk zijn dat de universiteiten en andere non-commerciële onderzoeksinstellingen belang hebben bij goed toegankelijke hulpmiddelen. Aan deze instellingen wordt een belangrijke rol toegedicht wanneer het gaat om het doen van fundamenteel, lange termijn, onderzoek, en wanneer het gaat om het opleiden van voldoende gekwalificeerd personeel. De universiteiten kunnen deze rol echter alleen vervullen wanneer de middelen die ze in huis hebben niet onderdoen voor datgene wat binnen het bedrijfsleven gangbaar is.

Bedrijven hebben belang bij een toegankelijke infrastructuur omdat het de drempel voor het ontwikkelen van Nederlandstalige TST-producten verlaagt. De soms geopperde gedachte dat algemeen toegankelijke hulpmiddelen de marktpositie van sommige commerciële instellingen ondermijnt moet ons inziens niet te serieus genomen worden. Er bestaat een grote afstand tussen de data en software die deel uitmaken van de infrastructuur en producten die geschikt zijn voor de consumentenmarkt of die op maat gemaakt zijn voor de behoeften van een afnemend bedrijf. Dit betekent dat er voldoende mogelijkheden overblijven voor commerciële exploitatie van de TST-markt.

Een laatste argument voor een algemeen toegankelijke infrastructuur is dat voor de hulpmiddelen die deel uitmaken van deze infrastructuur geldt dat het met name de som der delen is die maakt dat deze hulpmiddelen nuttig zijn. Een corpus algemeen Nederlands is bijvoorbeeld op zich weinig waardevol, maar kan wel dienen om data die aan een domeinspecifiek corpus zijn ontleend in het juiste perspectief te plaatsen. Een corpus voorzien van woordsoort wordt interessant wanneer er bijvoorbeeld ook een automatische *tagger* bestaat die werkt met dezelfde woordsoorten, en wordt nog interessanter wanneer er ook een programma voor syntactische analyse bestaat dat werkt met dezelfde woordsoorten. Op dezelfde manier kan een algemene grammatica van het Nederlands niet zonder een uitgebreid woordenboek, waaraan de informatie die voor het ontleden van concrete zinnen essentieel is, ontleend kan worden.

Op basis van deze twee overwegingen – de overheid dient een coördinerende en ondersteunende rol te spelen en de infrastructuur voor TST dient algemeen toegankelijk te zijn – komen we nu tot een aantal aanbevelingen. De aanbevelingen zijn gebaseerd op een karakterisering van de ideale infrastructuur voor TST, zoals we die hebben geschetst in hoofdstuk 1, het overzicht en de evaluatie van de bestaande infrastructuur, zoals we die hebben gegeven in hoofdstukken 2 en 3, en niet in de laatste plaats, op de interviews die we hebben gevoerd met deskundigen op het gebied van onderzoek, productontwikkeling, onderwijs, en beleid voor TST.

We willen hier tenslotte nog opmerken dat dit rapport tot stand is gekomen op het moment dat één van de belangrijkste initiatieven voor initiatieven voor TST voor het Nederlands, het project voor een Corpus Gesproken Nederlands, concrete vormen begint aan te nemen. Een aantal van onze gesprekspartners zijn ook bij dit project betrokken. De aanbevelingen die hieronder worden gedaan komen daarom op een belangrijk moment. Met de start van dit project wordt duidelijk dat aandacht voor TST niet slechts een incidentele zaak zou mogen zijn, en dat er een reële taak ligt voor een instantie die zich richt op het beheer van data en software (naast de resultaten van het CGN valt hierbij ook te denken aan de resultaten van een projecten als het Referentiebestand Nederlands en de lexicale databases van de Commissie Lexicale Vertaalvoorzieningen). Tenslotte moge duidelijk zijn dat een corpus gesproken Nederlands slechts één aspect is van een infrastructuur, en dat er nog een groot aantal hulpmiddelen niet, of slechts in zeer embryonale vorm, beschikbaar is voor het Nederlands.

Hieronder worden twee typen aanbevelingen geformuleerd: aanbeveling 1 waarbij de Nederlandse Taalunie een cruciale rol speelt vanwege de *monitor*[1] en *platform*functie[2] die ze vervult in het hele taalgebied, en de aanbevelingen 2 en 3 die

[1] Dit houdt in dat de NTU zorgt dat a) informatie met betrekking tot TST steeds ter beschikking staat voor alle belanghebbenden/belangstellenden en b) steeds wordt bijgewerkt.

[2] Dit houdt in dat de NTU zich tot taak stelt nieuwe initiatieven met betrekking tot TST te initiëren, te coördineren en te stimuleren, beleidsplannen en prioriteiten uit te denken enz.

buiten de directe competentie van de Taalunie vallen. De Taalunie zou ons inziens echter de bevoegde instanties in Nederland en Vlaanderen kunnen attenderen op hun verantwoordelijkheden in dezen.

5.1 Een Platform voor Taal- en Spraaktechnologie

> **Aanbeveling 1:** Het instellen van een Nederlands-Vlaams platform met als primaire taak het coördineren van activiteiten op het gebied van taal- en spraaktechnologie voor het Nederlands. De Taalunie zou hierbij als initiator en coördinator kunnen optreden.

Dit platform krijgt als centrale taken toebedeeld:

- Instellen van een Nederlands-Vlaamse werkgroep die tot taak heeft binnen een jaar een plan op te stellen voor het beheer, het onderhoud, en het beschikbaar stellen van materialen die kunnen worden ingezet bij het onderwijs, het onderzoek, en de ontwikkeling van producten op het gebied van de taal- en spraaktechnologie. De werkgroep zal moeten nagaan welke juridische constructie (bijvoorbeeld een stichting) het meest gechikt is en hoe een en ander kan worden gefinancierd.
  Met de volgende meer specifieke taken dient rekening te worden gehouden:

  - Het beheren van hulpmiddelen (met name lexicale databases en corpora) die zijn vervaardigd ten behoeve van onderzoek op het gebied van TST. Onder beheer wordt o.a. verstaan het eventueel aanpassen van het materiaal zodat het op verschillende computersystemen gebruikt kan worden, het waar mogelijk combineren van hulpmiddelen tot een nieuw hulpmiddel, het beschikbaar maken van documentatie, het beantwoorden van vragen van gebruikers, en eventueel het verlenen van assistentie bij het gebruik van deze hulpmiddelen.
  - Acquisitie van bestanden zoals woordenlijsten, tekstbestanden, en bestanden met spraak, die niet direct deel uitmaken van de TST-infrastructuur maar die daarbinnen een rol zouden kunnen spelen. Te denken valt aan tekstbestanden (fictie en nonfictie, wetenschappelijk, journalistiek, etc.) van uitgeverijen, van verschillende overheidsorganen, aan opnames van omroepen, en aan terminologische bestanden die binnen bedrijven, instellingen, en beroepsorganisaties worden aangelegd. Er moet worden onderzocht waar mogelijk waardevolle bestanden beschikbaar zijn, en er moeten afspraken met de leveranciers van het materiaal over auteursrechtelijke kwesties worden gemaakt. Daarnaast moet worden onderzocht wat in het buitenland beschikbaar is en hoe dat eventueel kan worden aangepast voor het Nederlands.
  - Onder onderhoud van bestanden moet onder meer worden verstaan het verbeteren van door gebruikers gesignaleerde fouten en omissies, het aanpassen aan nieuwe media, het eventueel uitbreiden van de dekking wanneer daaraan behoefte is. Daarnaast moet er regelmatig worden geëvalueerd

welke bestanden nog verder moeten worden onderhouden (hiervoor moeten criteria worden ontwikkeld) en door wie dat kan worden gedaan. Indien de oorspronkelijke makers hiervoor kunnen worden ingeschakeld verdient dit de voorkeur. Na verloop van tijd zal dit vaak niet meer mogelijk zijn en moet er een andere oplossing worden gezocht. Er moet ook worden onderzocht hoe bij toewijzing van een project het onderhoud in de eerstvolgende jaren al kan worden verzekerd.
  - Bij het beschikbaar stellen van hulpmiddelen spelen vaak juridische kwesties een rol (auteursrechten en dergelijke). Goede adviezen bij de start van een project zijn belangrijk. Verder moeten er overeenkomsten met de eigenaren van het materiaal worden gesloten over verdere distributie, het aanbrengen van verbeteringen, etc.

  Zo kan er een instelling in Nederland en Vlaanderen ontstaan die zich specifiek richt op het beheren van hulpmiddelen voor TST, een taak waarvoor momenteel geen enkele instantie verantwoordelijk is en een taak waar bovendien grote behoefte aan is.

- Instellen van een Vlaams-Nederlandse werkgroep die die tot taak heeft binnen een half jaar een plan op te stellen voor de versterking van de materiële infrastructuur voor de taal- en spraaktechnologie van het Nederlands. De werkgroep zal daarnaast de prioriteiten moeten vaststellen en moeten nagaan hoe een en ander kan worden gefinancierd.

  Te denken valt onder meer aan:

  - Het (daadwerkelijk en onder duidelijke voorwaarden) beschikbaar maken van materiaal dat in de afgelopen jaren aan diverse instituten is ontwikkeld.
  - Een omvangrijk (minstens 50 miljoen woorden) tekstcorpus dat uitgebalanceerd is qua genre, regio, en onderwerp.
  - Een woordenlijst conform de regels van de nieuwe spelling die qua omvang (minstens 500.000 woordvormen) en samenstelling (rekening houdend met het vocabulaire van speciale tekstsoorten en beroepsgroepen) geschikt voor automatische spellingcorrectie.
  - Een corpus, geannoteerd met woordsoort, dat gebruikt kan worden voor het trainen en testen van systemen voor het automatisch toekennen van woordsoorten (*taggers*).
  - Een *tagger* die de woordsoorten van bovengenoemd corpus hanteert en die een precisie heeft die conform internationale standaards is.
  - Een corpus, geannoteerd met constituentstructuur (*treebank*), dat gebruikt kan worden voor het trainen en testen van systemen voor automatische syntactische analyse.
  - Een *treebank* die de constituentstructuur van bovengenoemd corpus hanteert en die een precisie heeft die conform internationale standaards is.
  - Een parallel corpus (met name voor het taalpaar Nederlands-Engels).

– Tweetalige woordenboeken.
– Een formele en computationele, corpus-gebaseerde, beschrijving van de syntaxis van het hedendaagse Nederlands (*een computationele ANS* ofwel *de groene grammatica*) die kan dienen als uitgangspunt voor verder computationeel onderzoek naar de syntactische structuur van het Nederlands.
– Een elektronisch woordenboek dat gedetailleerde informatie bevat over syntactische fenomenen (met name valentie).
– Een elektronisch woordenboek dat semantische informatie bevat.
– Een set van materialen voor het 'waarderen' van Nederlandstalige NLP-producten en tools: standaarden, benchmarks, testsuite

Bij de overige taken van het platform zou men kunnen denken aan:

- Informatie verzamelen en verstrekken. Er wordt een overzicht bijgehouden van instellingen die actief zijn op het gebied van TST, van relevante projecten, en van subsidiemogelijkheden bij de EU en bij de nationale onderzoeksinstellingen. Dit zou kunnen gebeuren in een elektronische nieuwsbrief die regelmatig wordt verspreid.

- Instellen van een permanente adviescommissie voor TST. De commissie, waarin beleid, industrie, en universiteiten gerepresenteerd zijn, heeft als taak adviezen op te stellen die ertoe leiden dat de infrastructuur voor TST voldoet, respectievelijk blijft voldoen aan internationale maatstaven. De commissie kan bij nationale of Nederlands-Vlaamse projecten die bijvoorbeeld de ontwikkeling van woordenboeken, corpora, of grammatica's tot doel hebben, om advies worden gevraagd teneinde te garanderen dat optimaal rekening gehouden wordt met de eisen die TST-toepassingen stellen.

- Contacten tot stand brengen tussen overheid (nationaal en Europees), universiteit en bedrijfsleven.

- Organiseren van themabijeenkomsten.

- *Awareness* stimuleren, bijvoorbeeld door het organiseren van demonstraties van TST-producten en prototypes op beurzen, conferenties, en zomerscholen; door het opstellen en verspreiden van voorlichtingsmateriaal voor toekomstige studenten, etc.

- Uitgeven van een (elektronische) nieuwsbrief en/of een jaarverslag. Er vindt regelmatig rapportage plaats over de activiteiten van het platform en de adviescommissie in de vorm van een nieuwsbrief of jaarverslag, beschikbaar voor alle geïnteresseerden.

- Bevorderen van de Nederlands-Vlaamse samenwerking. Het platform en de adviescommissie onderzoeken actief hoe Nederlands-Vlaamse samenwerking op het gebied van TST bevorderd kan worden.

- Waar mogelijk en zinvol, het bevorderen van aandacht voor het Nederlands en betrokkenheid van Nederlandse partners in Europese initiatieven op het gebied van TST.

Uit het overzicht van instanties die verantwoordelijk zijn voor aspecten van het beleid voor TST en uit de interviews komt duidelijk naar voren dat er momenteel op beleids- en organisatorisch niveau sprake is van versnippering. Met het instellen van een platform voor TST kan deze versnippering worden tegengegaan. Daarbij moet bij een platform worden gedacht aan een structuur die weinig ambtelijke ondersteuning vraagt. Aan het geringe enthousiasme dat in sommige interviews doorklinkt voor het oprichten van weer een beleidsinstelling wordt op deze manier tegemoet gekomen.

Tegelijkertijd is een platform waarschijnlijk niet erg effectief wanneer het het zonder ondersteuning moet stellen. Bij ondersteuning kan men naast puur administratieve ondersteuning denken aan één wetenschappelijk medewerker die een wezenlijke inhoudelijke bijdrage kan leveren.

De Taalunie heeft een adviserende taak op het gebied van taalbeleid en taalpolitiek in Nederland en Vlaanderen. Daardoor sluit een platform voor TST dat adviseert in Nederland en Vlaanderen nauw aan bij de taakstelling en competentie van de Taalunie. De financiele middelen die nodig zijn voor de activiteiten van het platform zouden door het Comité van Ministers van de Taalunie ter beschikking moeten worden gesteld.

Anders dan bij de eerste aanbeveling kan de Taalunie bij de volgende twee aanbevelingen geen grote rol te spelen. Toch nemen we deze aanbevelingen in ons rapport op omdat zeker op de langere termijn goede onderwijs- en onderzoeksmogelijkheden op het vlak van TST doorslaggevend zullen zijn voor het succes van de TST-sector in Nederland en Vlaanderen.

5.2 Versterking van de positie van Onderzoek en Onderwijs

Aanbeveling 2: Het stimuleren van zowel fundamenteel als toegepast onderzoek op het gebied van TST

De positie van het universitair onderzoek in de taal- en spraaktechnologie kan verbeterd en verstevigd worden door de volgende maatregelen:

- Het beschermen van bestaande expertise in de universitaire onderzoeksgroepen (in Nederland), respectievelijk het geven van nieuwe impulsen aan dergelijke kernen van expertise (in Vlaanderen). Wetenschappelijk onderzoek (fundamenteel en toepassingsgericht) is het enige middel om nieuwe impulsen aan de industrie te kunnen geven. Vooral op langere termijn werpen investeringen hier hun vrucht af.

- Het zorgen voor continuïteit in onderzoeksfinanciering (in Nederland en Vlaanderen). De effecten van de door het beleid geleverde inspanningen zijn op langere termijn vaak gering omdat niet wordt geïnvesteerd in consolidatie en continuïteit waardoor de opgebouwde expertise snel weer verdwijnt.

- De mogelijkheden onderzoeken voor het tot stand brengen van een specifiek multidisciplinair fonds voor TST onderzoek. TST onderzoek valt vaak moeilijk bij een van de fondsen, of bij een van de afdelingen van die fondsen, onder te brengen. TST onderzoek bevat vaak zowel fundamentele als toegepaste aspecten en is daarbij ook nog vaak multidisciplinair (taalkunde, statistiek, informatica, psychologie). Sponsors van fundamenteel onderzoek (FWO, NWO) en toegepast onderzoek (IWT, STW) beschouwen TST projecten derhalve regelmatig als niet behorend tot hun domein.

- Het creëeren van een specifiek onderzoeksprogramma waarbij gezamenlijke, Vlaams-Nederlandse TST projecten kunnen worden ingediend. Gezamenlijke projecten komen nu nog vaak op ad hoc basis tot stand (cf. Corpus Gesproken Nederlands). De beschikbare middelen zouden beter kunnen worden besteed indien er meer structureel wordt samengewerkt. Het TST beleid in Nederland en Vlaanderen zou hiertoe meer op elkaar af moeten worden gestemd.

Aanbeveling 3: Het opzetten van een speciale (interuniversitaire) opleiding voor taal- en spraaktechnologie in Vlaanderen en het versterken van de opleidingen op dit gebied in Nederland.

De positie van het universitaire onderwijs in de taal- en spraaktechnologie kan verbeterd en verstevigd worden door de volgende maatregelen:

- Het opzetten, dan wel verder uitbouwen van een speciale opleiding voor taal- en spraaktechnologie in Vlaanderen, hetzij als specialisatie met een omvang van plusminus één jaar binnen bestaande programma's, hetzij als nieuwe opleiding binnen de geesteswetenschappen of binnen de informatica.
- Het bevorderen van de samenwerking tussen bestaande opleidingen in Nederland (en, wanneer geschikte partners zich aandienen, in Vlaanderen). Te denken valt aan overleg over curricula, uitwisseling van onderwijsmateriaal, gezamenlijk overleg met het bedrijfsleven, etc.
- Een onderzoek naar de mogelijkheden om te komen tot opleidingen die zich niet slechts richten op de taaltechnologie (computerlinguïstiek) of de spraaktechnologie (in Nederland meestal als onderdeel van de fonetiek, in Vlaanderen van elektrotechniek), maar die een evenwichtige combinatie zijn van beide. Gezien het technische karakter van een dergelijke opleiding ligt een exclusieve positie binnen de letteren niet voor de hand. Een interdisciplinaire opleiding, waarin zowel informatica als letteren deelneemt, verdient daarom wellicht de voorkeur.
- Het ontwikkelen van onderwijsmateriaal (teksten, oefeningen en opgaven, en software).
- Aansluiting bij Europese (Socrates) initiatieven op onderwijsgebied, zoals de ACOHUM-werkgroep voor computerlinguistiek en taaltechnologie, het netwerk voor fonetiek en spraakcommunicatie, en het voorgenomen *European Masters in Language and Speech*.

Bibliografie

Alshawi, H. (red.), (1992). *The Core Language Engine*. ACL-MIT press.

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

Van den Bergh, G., (1996). Taal is veel! In *Over de toekomst van het Nederlands*, Davidsfonds/Clauwaert, Leuven, pag. 9–24.

Berghmans, J. T., (1994). *WOTAN: een automatische grammaticale tagger voor het Nederlands*. Scriptie, Katholieke Universiteit Nijmegen.

Uit den Boogaart, P. C., (1975). *Woordfrequenties in geschreven en gesproken Nederlands*. Oosthoek, Scheltema & Holkema, Utrecht. Werkgroep Frequentie-onderzoek van het Nederlands.

van den Bosch, A., (1997). *Learning to pronounce written words. A study in inductive language learning*. Proefschrift, Universiteit Maastricht.

Bouma, G. en I. Schuurman, (1998). Intergovernmental language policy for Dutch and the language and speech infrastructure. In *Proceedings of the First international conference on language resources and evaluation*. Granada.

Brill, E., (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics* 21(4): 543–566.

Chanod, J.-P. en P. Tapanainen, (1995). Tagging French – comparing a statistical and a constraint-based approach. In *Seventh Conference of the European Chapter of the Association for Computational Linguistics*. Dublin, pag. 149–156.

Cherribi, O. en L. Sannen, (1998). De Nederlandse taal en de informatiemaatschappij. een korte notitie ten behoeve van de Interparlementaire Commissie van de Taalunie.

Cole, R. A., J. Mariani, H. Uszkoreit, A. Zaenen, en V. Zue (red.), (1998). *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, Cambridge.

Cutting, D., J. Kupiec, J. Pedersen, en P. Sibun, (1992). A practical part-of-speech tagger. In *Third Conference on Applied Natural Language Processing*. Trento, pag. 133–140.

Daelemans, W. en A. van den Bosch, (1996). Language-independent data-oriented grapheme-to-phoneme conversion. In *Progress in Speech Synthesis*. Springer Verlag, New York, pag. 77–90.

Daelemans, W., J. Zavrel, P. Berck, en S. Gillis, (1996a). MBT: A memory-based part of speech tagger-generator. In *Proceedings of the Fourth Workshop on Very Large Corpora*. Copenhagen, pag. 14–27.

Daelemans, W., J. Zavrel, P. Berck, en S. Gillis, (1996b). Memory-based part of speech tagging. In W. Daelemans, G. Durieux, en S. Gillis (red.), *CLIN 1995, Papers from the sixth CLIN Meeting 1995*. Universitaire Instelling Antwerpen, Antwerpen, pag. 41–62.

Dewallef, E., (1998). *Language Engineering in Flanders*. Ministerie van de Vlaamse Gemeenschap, departement voor Wetenschap, Innovatie, en Media, Brussel.

Geerts, G. en H. Heestermans, (1995). *Van Dale Groot woordenboek der Nederlandse taal*. Van Dale Lexicografie, Utrecht.

Gibbon, D., R. Moore, en R. Winski, (1997). *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin.

Haesereyn, W., K. Romijn, G. Geerts, J. De Rooy, en M. Van den Toorn, (1997). *Algemene Nederlandse Spraakkunst*. Martinus Nijhoff Uitgevers Groningen / Wolters Plantyn Deurne. Tweede, geheel herziene druk.

de Jong, E. D., (1979). *Spreektaal : woordfrequenties in gesproken Nederlands*. Bohn, Scheltema & Holkema, Utrecht. Werkgroep Frequentie-onderzoek van het Nederlands.

Kruyt, J. G., (1995). Nationale tekstcorpora in internationaal perspectief. *Forum der Letteren* 36(1): 47–58.

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.

Model, J., (1991). *Grammatische Analyse*. ICG Publications, Dordrecht.

Oltmans, E., (1994). AMAZON in AGFL. Een contextvrije herschrijfgrammatica voor de structurele component van het AMAZON/CASUS-systeem, beschreven in het AGFL-formalisme. Doctoraal scriptie, Universiteit Nijmegen.

Roukens, J., (1998). The Multilingual Information Society. *Elsnews* pag. 1.

Samuelson, C. en A. Voutilainen, (1997). Comparing a linguistic and a stochastic tagger. In *Proceedings of 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, pag. 246–253.

van Staden, M., (1998). *EUROMAP-Netherlands*. TNO Strategie, Technologie en Beleid, Den Haag.

Van Eynde, F., (1996). Het Nederlands en de nieuwe taaltechnologie. Vijf voor twaalf? In *Over de toekomst van het Nederlands*, Davidsfonds/Clauwaert, Leuven, pag. 25–40.

van der Voort van der Kleij, J., S. Raaijmakers, M. Panhuijsen, M. Meijering, en R. van Sterkenburg, (1994). Een automatisch geanalyseerd corpus hedendaags Nederlands in een flexibel retrievalsysteem. In L. Noordman en W. de Vroomen (red.), *Informatiewetenschap 1994. Wetenschappelijke bijdragen aan de derde STINFON-conferentie*. Tilburg, pag. 181–194.

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

Wells, J., (1987). Computer-coded phonetic transcription. *Journal of the International Phonetic Association* 17(2): 94–114.

Woordenlijst, (1996). *Woordenlijst Nederlandse Taal*. Sdu, Den Haag. (in opdracht van de Nederlandse Taalunie, met een voorwoord door Jan Renkema).