

*Nederlandse Taalunie*

**Vlaams-Nederlands meerjarenprogramma voor  
Nederlandstalige taal- en spraaktechnologie**

# **STEVIN**

*Spraak- en Taaltechnologische  
Essentiële Voorzieningen  
In het Nederlands*

STEVIN is een samenwerkingsverband tussen  
de Nederlandse Taalunie  
de administratie Wetenschap en Innovatie (AWI) van het Ministerie van de Vlaamse Gemeenschap (MVG)  
het Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie (IWT-Vlaanderen)  
het Fonds voor Wetenschappelijk Onderzoek (FWO-Vlaanderen)  
het Nederlandse Ministerie van Onderwijs, Cultuur en Wetenschap (OCW)  
het Nederlandse Ministerie van Economische Zaken (EZ)  
het Nederlandse EZ-agentschap SenterNovem  
de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO)  
het NWO gebied Exacte Wetenschappen (EW)  
het NWO gebied Geesteswetenschappen (GW)

15 september 2004



## Colofon

15 september 2004

De tekst van het meerjarenprogramma voor Nederlandstalige taal- en spraaktechnologie is opgesteld door een programmavoorbereidingscommissie bestaande uit:

- prof. dr. J. Odijk (Scansoft/Universiteit Utrecht) - voorzitter
- prof. dr. J.P. Martens (Universiteit Gent)
- prof. dr. F. van Eynde (Katholieke Universiteit Leuven)
- prof. dr. W. Daelemans (Universiteit Antwerpen/Universiteit van Tilburg)
- mw. D. Kenyon-Jackson (Polderland Language & Speech Technology BV)
- dr. P. Vossen (Irion Technologies)
- dr. A. van Hessen (Telecats BV/Universiteit Twente)
- prof. dr. L. Boves (Katholieke Universiteit Nijmegen)
- mw. dr. J. Beeken (INL/TST-centrale).

Deze commissie werd begeleid door

- prof. dr. K. Jaspaert (Nederlandse Taalunie)
- dhr. E. Dewallef (Ministerie Vlaamse Gemeenschap - administratie Wetenschap en Innovatie)
- ir. H. Kruihof (SenterNovem)
- mw. drs. A. Dijkstra (NWO).

*Nederlandse Taalunie  
Lange Voorhout 19  
Postbus 10595  
2501 HN Den Haag  
Nederland  
[www.taalunieversum.org](http://www.taalunieversum.org)*

<b>1. Samenvatting</b> .....	<b>3</b>
<b>2. Achtergronden</b> .....	<b>4</b>
2.1. Voorgeschiedenis	
2.2. Motivatie STEVIN	
2.3. Praktische aanloop tot STEVIN	
2.4. Relatie met andere lopende programma's	
<b>3. De kennisinfrastructuur</b> .....	<b>8</b>
3.1. Kennisinfrastructuur in Nederland en Vlaanderen	
3.2. Kennisinfrastructuur in de rest van Europa/de wereld	
3.3. Behoeften van de kennisinfrastructuur	
<b>4. Het bedrijfsleven</b> .....	<b>11</b>
4.1. Bedrijfsleven in Nederland en Vlaanderen	
4.2. Bedrijfsleven in de rest van Europa/de wereld	
4.3. Toepassingen	
4.4. Behoeften van het bedrijfsleven	
<b>5. Opzet integraal TST-meerjarenprogramma STEVIN</b> .....	<b>13</b>
5.1. De ketenbenadering	
5.2. Doelstellingen en integrale aanpak	
5.3. BaTaVo: data, tools, modules en onderzoeksthema's	
5.3.1 Spraaktechnologische basistaalvoorzieningen (laag 1)	
5.3.2 Spraaktechnologisch strategisch onderzoek (laag 2)	
5.3.3 Taaltechnologische basistaalvoorzieningen (laag 1)	
5.3.4 Taaltechnologisch strategisch onderzoek (laag 2)	
5.3.5 Taal- en Spraaktechnologische toepassingen (laag 3)	
5.4. Vraagstimulering	
5.5. Randvoorwaarden: IPR en standaarden	
5.6. Relatie met andere programma's	
5.7. Draagvlak: kennisinstellingen en industrie	
5.8. Beoordelingscriteria	
5.9. Aanvraagprocedure	
5.10. Evaluatie (nulmeting)	
<b>6. Kennisoverdracht, netwerkvorming en verankering</b> .....	<b>27</b>
6.1. Inleiding	
6.2. Geplande activiteiten	
6.3. Netwerkvorming en kennisuitwisseling	
6.4. Kennisoverdracht	
6.5. Zwaartepuntvorming en verankering	
6.6. Onderhoud, beheer en exploitatie van resultaten	
6.7. Internationalisering	

<b>7. Organisatie van het programma .....</b>	<b>31</b>
7.1. Inleiding en organigram	
7.2. De Programmacommissie, het Programmabestuur en het Programmabureau	
7.3. Projectleiders en onderzoekers	
7.4. Werkgroep Flankerend beleid	
7.5. Begeleidingscommissie	
7.6. Belangenverstrengeling	
<b>8. Financiën van het programma .....</b>	<b>36</b>
8.1. Middelen: MVG-AWI, EZ, NWO, OCW	
8.2. Kennisontwikkeling	
8.3. Flankerende activiteiten	
8.4. Programmamanagement en beheer	
8.5. Planning middelen in de tijd	
<b>9. Activiteitenplan 2004 - 2005.....</b>	<b>39</b>
<b>Bijlagen:</b>	
bijlage 1: Beleidsaanbevelingen (uit <i>Blauwdruk voor het beheer en onderhoud van met overheidsmiddelen gefinancierde digitale materialen</i> ) .....	40
bijlage 2: Intentieverklaring voor de versterking van de strategische samenwerking tussen Vlaanderen en Nederland op het vlak van innovatie .....	43
bijlage 3: Aanbevelingen Technologieverkenning .....	45
bijlage 4: Relatie met andere lopende (internationale) programma's en projecten.....	46
bijlage 5: Overzicht omvang kennisinfrastructuur .....	50
bijlage 6: Persoonlijke betrokkenheid bij aanvragen van leden van adviescolleges .....	52
bijlage 7: Overzicht taken en verantwoordelijkheden Programmabestuur, Programmacommissie, begeleidingscommissie en programmabureau .....	57
bijlage 8: Verklaring gebruikte afkortingen .....	59
De Taalunie is samen met NWO en SenterNovem begonnen met de formulering van:	
bijlage 9: Subsidieregeling Vlaams-Nederlands TST meerjarenprogramma	
bijlage 10: IPR regeling en gedragsregels kennisbescherming en octrooien	

## 1. Samenvatting

Het Vlaams-Nederlands meerjarenprogramma voor Nederlandstalige taal- en spraaktechnologie STEVIN (Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands) heeft als doel het geven van een stimulans aan een technologiesector om de innovatiecapaciteit van die sector te vergroten en tegelijkertijd de positie van het Nederlands in de moderne informatie- en communicatiewereld te behouden.

STEVIN heeft een hybride karakter waarbij een combinatie van stimuleringsinstrumenten wordt ingezet. De doelstellingen van het programma kunnen puntsgewijs als volgt worden samengevat:

- het realiseren van een adequate digitale taalinfrastructuur voor het Nederlands, gebaseerd op de BaTaVo-prioriteiten (BasisTaalVoorzieningen);
- het doen van strategisch onderzoek op het gebied van taal- en spraaktechnologie, met name op gebieden waar grote vraag naar is vanuit concrete toepassingen van de technologieën;
- het stimuleren van netwerk- en zwaartepuntvorming en verankering, het opleiden van nieuwe experts, het bevorderen van vraagstimulering en kennisoverdracht, en het adequaat regelen van intellectuele eigendomsrechten.

Een adequate digitale taalinfrastructuur van essentiële data en modules, gedragen door excellente onderzoekers en ontwikkelaars, is noodzakelijk voor zowel de kennisinfrastructuur als het bedrijfsleven. De kennisinfrastructuur heeft de data nodig als grondstof en de tools als hulpmiddel bij het beoefenen van de wetenschap; voor het bedrijfsleven zijn ze onmisbaar om complexe applicaties te kunnen ontwikkelen voor een aantal talen, maar primair voor het Nederlands, zodat de economische en culturele positie van het Nederlands in de mondiale informatie- en communicatietechnologie behouden blijft. Flankerende actielijnen zullen worden opgezet ter bevordering van de kennisoverdracht en vraagstimulering om op die manier innovatiemogelijkheden te optimaliseren.

STEVIN zal in hoofdzaak uitgevoerd worden door kennisinstellingen, maar met steun van en samenwerking met het bedrijfsleven zodat de ontwikkelde basistaalvoorzieningen en het uitgevoerde onderzoek nauw aansluiten bij daadwerkelijke behoeften van de TST-leveranciers, de applicatieontwikkelaars en hun klanten. Het adequaat regelen van intellectuele eigendomsrechten van de in te brengen en te produceren basistaalvoorzieningen en tools, de standaardisatie daarvan, het onderhoud, het beheer en de exploitatie zullen een integraal onderdeel vormen van het programma en in nauwe samenwerking met de TST-centrale worden gerealiseerd.

## 2. Achtergronden

De verregaande integratie tussen informatie- en communicatietechnologie (ICT) heeft mede geleid tot de meertalige informatiemaatschappij die wij vandaag de dag kennen. De instrumenten die de mens in staat stellen om met de hem omringende intelligente informatieomgeving te communiceren (zoals PC of GSM), moeten voortdurend kunnen inspelen op de snelle evoluties waaraan onze maatschappij onderhevig is. Een belangrijke technologische ontwikkeling daarbij is de communicatie tussen mens en machine met behulp van natuurlijke taal, waardoor de interactie tussen mens en machine steeds meer vanzelfsprekend wordt. Die interactie via natuurlijke taal is ondenkbaar zonder taal- en spraaktechnologie (TST).

*Taaltechnologie* verwijst naar de ontwikkeling van computersystemen die in staat zijn op een intelligente wijze natuurlijke taal te begrijpen, te verwerken en te reproduceren. Toepassingsgebieden van taaltechnologie kunnen gaan van de bekende spelling- en grammaticacontrole bij tekstverwerkingssoftware over automatische vertaalinstrumenten tot geavanceerde systemen voor informatie- en kennisbeheer en intelligente zoekmachines op het internet. Wat dit laatste betreft staat de technologische evolutie aan de vooravond van het "semantische web", dat niet enkel op trefwoorden maar ook op inhoud kan worden doorzocht. Bij *spraaktechnologie* gaat het om het herkennen en produceren van gesproken taal door machines. Spraaktechnologische toepassingen die meer en meer ingeburgerd raken, zijn dicteersystemen, interactieve leermethodes (vooral bij taalonderwijs), automatische informatiediensten, spraakgestuurde informatie- en navigatiesystemen in auto's en ontsluiting van multimediacdocumenten. Taal- en spraaktechnologie speelt ook een onmisbare rol bij het ontwikkelen van multimodale mens-systeeminteractie en de *ubiquitous and pervasive computing* die essentieel zijn voor wat bekend staat als *ambient intelligence*. Taal- en spraaktechnologie kan tevens in belangrijke mate het vermogen van visueel of motorisch gehandicapten bevorderen om met machines (of met behulp van deze machines met hun omgeving) te communiceren.

Naarmate ICT alsmear meer deel uitmaakt van het dagelijkse leven van de gebruiker, wordt het voor die gebruiker ook steeds belangrijker om in deze communicatie de eigen moedertaal te kunnen hanteren. Enkel op die manier worden taalbarrières opgeheven en komt een volwaardige participatie van alle burgers in de meertalige informatiemaatschappij binnen bereik. Elke taal die haar positie in deze maatschappij wil handhaven of versterken moet dan ook gelijke tred kunnen houden met de ontwikkelingen op het vlak van taal- en spraaktechnologie. Daarvoor zijn data, tools en menselijke expertise onontbeerlijk.

Het is niet de verantwoordelijkheid van de bedrijven die taal- en spraaktechnologische toepassingen ontwikkelen om de benodigde kennisinfrastructuur te bouwen en in stand te houden. Zij laten zich leiden door marktkansen en richten zich dan ook in hoofdzaak op de "grote" talen die navenante afzetmogelijkheden kunnen bieden. Het is wel een kernopdracht van de overheid. Met name in kleinere taalgebieden moet de overheid ervoor zorgen dat de benodigde digitale taalinfrastructuur voorhanden is en voldoende kwaliteit biedt. De digitale taalinfrastructuur van een bepaalde taal is het geheel van basistaalvoorzieningen (afgekort tot BaTaVo), in feite de "grondstoffen" die nodig zijn om taal- en spraaktechnologische toepassingen in die taal te kunnen ontwikkelen. Het gaat daarbij zowel om digitale gegevensbanken (i.e. corpora van geschreven en gesproken taal, elektronische woordenboeken en computationele lexicons) als om software en trainingsmateriaal voor het helpen aanmaken van de verschillende soorten verrijking van het desbetreffende taalmateriaal.

Het subsidiariteitsbeginsel<sup>1</sup> dat de Europese Commissie hanteert, gaat hier ook van uit: de Commissie wil zelf wel bijdragen tot het creëren van een meertalige taalinfrastructuur om zo de meertalige communicatie te garanderen, maar elk taalgebied in Europa is verantwoordelijk voor de opbouw van een eigen digitale taalinfrastructuur. Zo ook het Nederlandse taalgebied, dat de nodige inspanningen moet leveren om een volwaardige speler te blijven in de meertalige Europese en wereldwijde informatiemaatschappij.

---

<sup>1</sup> Het subsidiariteitsbeginsel is als algemeen beginsel bij het Verdrag van Maastricht (1997) in het Verdrag van de Europese Gemeenschap opgenomen. Het beginsel bepaalt dat de Unie - behalve op terreinen waar zij een exclusieve bevoegdheid heeft (zoals over de euro) - alleen optreedt wanneer haar optreden efficiënter is dan een optreden op nationaal, regionaal of plaatselijk niveau.

## 2.1. Voorgeschiedenis

Nederland en Vlaanderen hebben de afgelopen jaren steeds intensiever samengewerkt op het gebied van zowel onderzoek als ontwikkeling in Nederlandstalige taal- en spraaktechnologie. Gezien het gezamenlijke belang en de kosten die op deze manier gedeeld kunnen worden, is dat een uitstekende ontwikkeling. Door een gezamenlijke Nederlands-Vlaamse inspanning van overheid, kennisinfrastructuur en bedrijfsleven kan versnippering van kennis, ervaring en middelen geminimaliseerd worden. Voorbeelden van deze samenwerking zijn:

- Het Vlaams-Nederlandse programma *Corpus Gesproken Nederlands* (CGN)<sup>2</sup> is gericht op de aanleg van een databank van het hedendaags Standaardnederlands zoals dat wordt gesproken door volwassenen in Nederland en Vlaanderen. Dit corpus is onontbeerlijk voor het onderzoek naar de Nederlandse taal en voor de verdere ontwikkeling van de taal- en spraaktechnologie in bijvoorbeeld Nederlandstalige automatische spraakherkenners en dialogsystemen. Het corpus kan daardoor bijdragen aan het behoud van de economische en culturele positie van het Nederlands in Europa. Het CGN-programma startte in 1998 en is begin 2004 afgerond. Het werd gefinancierd door OCW in Nederlandse (Ministerie van Onderwijs, Cultuur en Wetenschap) en MVG in Vlaanderen (Ministerie van de Vlaamse Gemeenschap), en praktisch gecoördineerd door NWO (Nederlandse Organisatie voor Wetenschappelijk Onderzoek).
- Het *Vlaams-Nederlands platform voor het Nederlands in Taal- en Spraaktechnologie* (TST-platform) is in 1999 opgezet en werd gefinancierd door MVG-AWI (Administratie Wetenschap en Innovatie), OCW, EZ (Ministerie van Economische Zaken) en NWO. Dit platform heeft het opstellen van een gezamenlijke Vlaams-Nederlandse beleidsagenda voor de Nederlandstalige taal- en spraaktechnologie mogelijk gemaakt met als oogmerk dat het Nederlands een volwaardige rol kan (blijven) spelen in de meertalige informatiemaatschappij.<sup>3</sup> De coördinatie was in handen van de Nederlandse Taalunie.
- In de Vlaams-Nederlandse participatie in het Europese *EUROMAP*-consortium, gefinancierd vanuit het KP5-IST-Programma (Vijfde Kaderprogramma, Information Society Technologies), voerden Nederland en Vlaanderen gezamenlijk activiteiten uit om a) de kennis bij bedrijven, organisaties en gebruikers over de mogelijkheden van taal- en spraaktechnologie te vergroten; b) technologieoverdracht van onderzoekinstellingen naar de markt te bevorderen; en c) *community building* binnen specifieke domeinen te bevorderen. In de tweede fase van dit project - vanaf 2000 - hebben Nederland en Vlaanderen gezamenlijk als partner deelgenomen in dit project. De coördinatie werd uitgevoerd door de Nederlandse Taalunie in samenwerking met SenterNovem.
- In 2000 heeft de Nederlandse Taalunie na een uitgebreide tender-procedure aan Systran<sup>4</sup> de opdracht gegeven tot het uitvoeren van het *NL-Translex*-project. Dit project is gezamenlijk door de Vlaamse en Nederlandse overheid met financiering van het Europese MLIS-project (Multilingual Information Society) gerealiseerd. In het kader van NL-Translex zijn componenten (Nederlandse lexica en andere taalspecifieke modules) ontwikkeld voor systemen voor automatisch vertalen die vertalingen mogelijk maken van en naar het Nederlands op het hoogste prestatieniveau. Deze systemen kunnen worden gebruikt binnen de instellingen van de Europese Unie en door de overheden van de lidstaten. Het Nederlands geldt als bron- én doeltaal. De andere talen in de vertaalparen zijn Engels en Frans.
- Het uitbouwen van de *makel- en schakelfunctie* van de Nederlandse Taalunie kwam voort uit een van de drie actielijnen van het TST-platform. De voornaamste initiatieven die in dit kader werden opgezet of uitgebouwd waren: het opzetten van een TST-infodesk<sup>5</sup>; het bevorderen van kennisoverdracht tussen verschillende actoren via het organiseren van conferenties, workshops en seminars; gezamenlijke vertegenwoordiging van Nederland en Vlaanderen in diverse internationale projecten en netwerken (EUROMAP, ENABLER, voorbereiding LANGNET). Na het beëindigen van de actielijn worden de activiteiten door de Taalunie in opdracht van het Comité van Ministers voortgezet.
- De Nederlandse Taalunie heeft in 2004 een *taal- en spraaktechnologiecentrale* (TST-centrale) opgezet, die tot doel heeft het onderhoud, het beheer en de exploitatie van digitale taalmaterialen voor het Nederlands te coördineren. Het Instituut voor Nederlandse Lexicologie

---

<sup>2</sup> Zie <http://lands.let.kun.nl/cgn/>

<sup>3</sup> Zie Actieplan Nederlands in Taal- en Spraaktechnologie: <http://taalunieversum.org/taal/technologie/>

<sup>4</sup> Zie <http://www.systransoft.com/>, met de gratis mogelijkheid voor korte vertalingen.

<sup>5</sup> Zie <http://www.taalunieversum.org/taal/technologie/>.

(INL) heeft de opdracht gekregen de TST-centrale op te zetten en werkt hiervoor samen met verschillende expertisecentra, wetenschappelijke instellingen én partners uit de particuliere sector in Nederland en Vlaanderen. Ook deze activiteit is voortgekomen uit een van de actielijnen van het TST-platform.

- De Nederlandse Minister van EZ en de Vlaamse Minister van Financiën en Begroting, Ruimtelijke Ordening, Wetenschap en Technologische Innovatie hebben op 7 april 2004 een *Intentieverklaring voor de versterking van de strategische samenwerking tussen Vlaanderen en Nederland op het vlak van innovatie* (zie *bijlage 2*) ondertekend.

## 2.2. Motivatie STEVIN

STEVIN is voortgekomen uit de actielijnen die in de afgelopen jaren zijn uitgevoerd onder verantwoordelijkheid van het TST-platform. Daarnaast is een belangrijke bron van inspiratie geweest de *Technologieverkenning Nederlandstalige Taal- en Spraaktechnologie*, die in opdracht van het Nederlandse Ministerie van EZ is opgesteld.<sup>6</sup>

Om een extra impuls te geven aan het verstevigen van de positie van het Nederlands in taal- en spraaktechnologie en dus op de wereldwijde informatiesnelweg, werd in 1999 het TST-platform opgericht. In dit platform zetelen vertegenwoordigers van de Vlaamse overheid (MVG-AWI, IWT-Vlaanderen<sup>7</sup>, FWO-Vlaanderen<sup>8</sup>) en de Nederlandse overheid (OCW, EZ, SenterNovem en NWO) en de Nederlandse Taalunie. Het TST-platform stelde een *Actieplan voor het Nederlands in TST* op, met als doel de juiste randvoorwaarden te creëren voor een succesvol beleid met betrekking tot taal- en spraaktechnologie. Het TST-platform heeft drie actielijnen uitgevoerd:

1. De uitbouw van een make- en schakelfunctie voor de taal- en spraaktechnologie ten behoeve van actoren (kennisinstellingen, bedrijven, beleidsorganisaties en gebruikers) op het vlak van taal- en spraaktechnologie in Vlaanderen, Nederland en Europa.
2. Het opstellen van een BaTaVo-prioriteitenlijst. Met het oog op de verdere uitbouw van de Nederlandstalige digitale taalinfrastructuur werd een lijst opgemaakt met de basisvoorzieningen voor het Nederlands die prioritair moeten worden ontwikkeld voor taaltechnologie en spraaktechnologie.
3. Het opstellen van een blauwdruk voor doeltreffend beheer, onderhoud, distributie en beschikbaarstelling van met overheidsmiddelen ontwikkelde digitale materialen voor het Nederlands. Deze actielijn heeft geleid tot het instellen van de TST-centrale.

In opdracht van het Nederlandse Ministerie van EZ werd vervolgens in 2003, met de BaTaVo-prioriteitenlijst als uitgangspunt, een uitgebreide technologieverkenning uitgevoerd. Dit gebeurde op advies van de Stuurgroep IOP<sup>9</sup> na beoordeling van een eerste voorstel voor een Vlaams-Nederlands TST-meerjarenprogramma. In de *Technologieverkenning Nederlandstalige Taal- en Spraaktechnologie* worden de hiernavolgende conclusies en aanbevelingen opgesomd.

1. Taal- en spraaktechnologie kan een aanzienlijke bijdrage leveren tot duurzame economische groei.
2. Het model van het dynamisch innovatiesysteem is goed bruikbaar voor het formuleren van een programmatische aanpak.
3. Een programmatische aanpak is de aangewezen manier om een extra injectie aan de taal- en spraaktechnologie te geven.
4. Er wordt geadviseerd om over te gaan tot het opzetten van een hybride onderzoeksprogramma voor Nederlandstalige taal- en spraaktechnologie dat de verschillende lagen van het innovatiesysteem omvat.

## 2.3. Praktische aanloop tot STEVIN

Alle participerende partijen zijn van mening dat Vlaanderen en Nederland hun samenwerking op het gebied van taal- en spraaktechnologie moeten continueren. Door een gezamenlijke Nederlands-

---

<sup>6</sup> Deze verkenning werd uitgevoerd door het Nederlands-Vlaamse Consortium M&I/Partners - Montemore NV. De volledige tekst van de aanbevelingen is opgenomen in *bijlage 3*.

<sup>7</sup> Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen.

<sup>8</sup> Fonds voor Wetenschappelijk Onderzoek.

<sup>9</sup> EZ-subsidieinstrument Innovatiegerichte Onderzoeksprogramma's, zie <http://www.senter.nl/iop/>.

Vlaamse inspanning van overheid, kennisinfrastructuur en bedrijfsleven kan versnippering van kennis, ervaring en middelen geminimaliseerd worden en kennisuitwisseling tussen academisch onderzoek en bedrijfsleven geoptimaliseerd worden. Zowel Vlaanderen als Nederland beschikken over voldoende potentieel om een adequate digitale taalinfrastructuur voor het Nederlands uit te bouwen, om daarmee taal- en spraaktechnologische toepassingen te ontwikkelen waarmee Nederlandstalige burgers optimaal kunnen participeren in de meertalige informatiemaatschappij. Verscheidene bedrijven in Vlaanderen en Nederland hebben een vooraanstaande positie weten te verwerven op de mondiale markt van taal- en spraaktechnologie, en de Vlaamse en Nederlandse kennisinstellingen hebben wereldwijd een sterke reputatie weten op te bouwen in diverse domeinen van het taal- en spraaktechnologische onderzoek. STEVIN bevordert synergie en samenwerking tussen de verschillende spelers in de taal- en spraaktechnologische sector, zodat het aanwezige potentieel optimaal wordt aangewend. Daarbij is ook de onderlinge samenwerking tussen Vlaanderen en Nederland van groot belang, want een gezamenlijke aanpak kan de slagkracht om de belangen van het Nederlandse taalgebied in een meertalig Europa te behartigen alleen maar verhogen.

In 2002 en 2003 is door verschillende partijen in Nederland en Vlaanderen gezocht naar de noodzakelijke financiering voor een Vlaams-Nederlands onderzoeksprogramma voor Nederlandstalige taal- en spraaktechnologie (met als uitgangspunt de BaTaVo-prioriteitenlijst). Voor de financiering wordt de standaardverdeling aangehouden: Vlaanderen 1/3 deel en Nederland 2/3 deel. Voor de definitieve toezegging van de financierende partijen bleek een bijstelling van de plannen noodzakelijk. Met name werd de behoefte gesignaleerd om de activiteiten in te bedden in de innovatieketen. Daarnaast gaf de Stuurgroep IOP een aantal specifieke praktische adviezen o.a. ten aanzien van de organisatie van STEVIN waar een sterke regie, goede bestuurbaarheid, concrete richting en doelstellingen en haalbaarheid belangrijke basisprincipes moeten vormen.

STEVIN is opgesteld op basis van input van alle financierende partijen. Ook zijn meegenomen de adviezen van de Vlaams-Nederlandse Commissie Terminologie (CoTerm)<sup>10</sup> en de Adviescommissie Lexicografische Vertaalvoorzieningen (ALVV, voortgekomen uit de CLVV<sup>11</sup>). Uiteindelijk heeft dit geleid tot een ambitieus programma dat past binnen de doelstellingen van alle financiers, en waarbinnen sprake is van daadwerkelijke *cross-border funding*. Ook past het programma binnen de *Intentieverklaring* van de Nederlandse Minister van EZ en de Vlaamse Minister van Financiën en Begroting, Ruimtelijke Ordening, Wetenschappen en Technologische Innovatie (zie *bijlage 2*). Tevens sluit het uitstekend aan bij de grensoverschrijdende ambities van de wetenschapsfondsen in Nederland en Vlaanderen.

In de volgende hoofdstukken is een nadere toelichting te vinden op het betrokken veld, respectievelijk de kennisinfrastructuur en bedrijfsleven in Nederland en Vlaanderen (hoofdstuk 3 en 4), de opzet van STEVIN (hoofdstuk 5), de beoogde activiteiten ten aanzien van kennisoverdracht, netwerkvorming en verankering (hoofdstuk 6), de organisatie van het programma (hoofdstuk 7) en de financiën (hoofdstuk 8).

## 2.4. Relatie met andere lopende programma's

STEVIN is gerelateerd aan een aantal nationale en internationale lopende onderzoeksprogramma's. In Nederland lopen reeds een aantal grote programma's die zich deels op hetzelfde onderzoeksgebied richten dan wel op aanverwante onderzoeksgebieden. Sterk gerelateerde programma's in Nederland zijn het NWO *IMIX*-onderzoeksprogramma (Interactieve Multimodale Informatie-eXtractie), het IOP *MMI* programma (Mens-Machine Interactie), het BTS<sup>12</sup>-project *Waterland*, en het CIC<sup>13</sup>-programma *Pidgin*. Daarnaast worden binnen de BSIK<sup>14</sup>-programma's MultimediaN en ICIS een aantal aanverwante projecten uitgevoerd (zie sectie 5.6).

IWT-Vlaanderen heeft in het kader van de SBO-regeling (strategisch basisonderzoek) twee gerelateerde projecten gefinancierd: *FlaVoR* en *AtraNoS*.

Daarnaast is er reeds een lopend project waarin Vlaanderen en Nederland samenwerken dat wordt gefinancierd door FWO-Vlaanderen en NWO via het VNC<sup>15</sup>-programma *PROSIT*.

<sup>10</sup> Zie: [http://taalunieversum.org/taalunie/commissie\\_terminologie\\_coterm/](http://taalunieversum.org/taalunie/commissie_terminologie_coterm/).

<sup>11</sup> Zie [http://taalunieversum.org/taalunie/commissie\\_lexicografische\\_vertaalvoorzieningen\\_clvv/](http://taalunieversum.org/taalunie/commissie_lexicografische_vertaalvoorzieningen_clvv/).

<sup>12</sup> Bedrijfgerichte Technologische Samenwerkingsprojecten, EZ-subsidieinstrument, zie <http://www.senter.nl/asp/page.asp?id=i000008&alias=technologischesamenwerking>.

<sup>13</sup> Concurrenieren met ICT Competenties, programma ontwikkeld door EZ en OCW met het doel ICT-doorbraken te stimuleren, zie <http://www.cic-online.nl/>.

<sup>14</sup> Besluit subsidies investeringen kennisinfrastructuur (voorheen ICES/KIS), zie <http://www.senter.nl/bsik/>.

<sup>15</sup> Vlaams Nederlands Comité, programma voor Vlaams Nederlandse samenwerking, zie <http://www.nwo.nl/vnc/>.

Ook zijn Vlaamse en Nederlandse onderzoekers als partner betrokken (geweest) bij een groot aantal internationale onderzoeksprogramma's. De belangrijkste nog lopende EU projecten daarvan zijn MUMIS (Multimedia and Searching Environment), SMADA (Speech-driven Multimodal Automatic Directory Assistance), COMIC (Conversational Multimodal Interaction with Computers), M4 (MultiModal Meeting Manager), BIOMINT (Biological Text Mining), MUSA (Multilingual Subtitling of Multimedia Content), 2002-METIS (Statistical Machine Translation using Monolingual Corpora), PASCAL: (Pattern Analysis, Statistical Modelling and Computational Learning), SAFIR (Speech Automatic Friendly Interface Research), en AMI (Augmented Multi-party Interaction).

Meer informatie over de bovengenoemde Nederlandse, Vlaamse, bilaterale en Europese projecten is te vinden in *bijlage 4*.

### 3. De kennisinfrastructuur

#### 3.1. Kennisinfrastructuur in Nederland en Vlaanderen

In 2003 is door M&I/Partners in samenwerking met Montemore in opdracht van EZ een Technologieverkenning Nederlandstalige Taal- en Spraaktechnologie uitgevoerd. Onderdeel daarvan was een overzicht van de omvang van het taal- en spraaktechnologische onderzoeksveld. De belangrijkste gegevens zijn in de hieronder staande tabellen met enige kleine rekentechnische correcties samengevat. Verdere details zijn opgenomen in *bijlage 5*.

*Omvang en financiering TST-onderzoeksgroepen in aantal fte (fulltime equivalenten onderzoekscapaciteit):*

Fte in Taal en Spraak	Taal			Spraak		
	Vlaanderen	Nederland	som	Vlaanderen	Nederland	som
voltijds studenten <sup>16</sup>	4,0	36,0	<b>40,0</b>	3,0	7,0	<b>10,0</b>
(wetenschappelijke) staf	30,8	85,5	<b>116,3</b>	19,0	34,7	<b>53,7</b>
	18%	50%	<b>68%</b>	11%	20%	<b>32%</b>

*Het financiële volume<sup>17</sup> (in miljoen euro per jaar) dat hier naar schatting mee is gemoeid:*

Financiële omvang TST	Taal			Spraak		
	Vlaanderen	Nederland	som	Vlaanderen	Nederland	som
(wetenschappelijke) staf	<b>2,1</b>	<b>4,7</b>	<b>6,8</b>	<b>1,3</b>	<b>2,0</b>	<b>3,3</b>

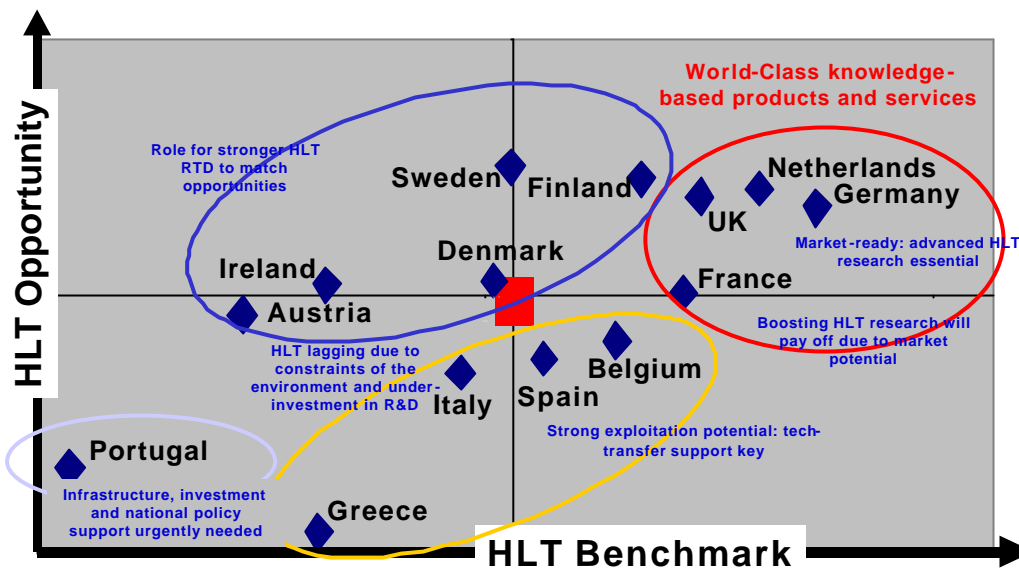
#### 3.2. Kennisinfrastructuur in de rest van Europa/de wereld

In Europa wordt veel hooggekwalificeerd onderzoek op het gebied van taal- en spraaktechnologie uitgevoerd dat zich kan meten met de grote HLT-initiatieven (Human Language Technologies) in de Verenigde Staten en Japan. In een in het kader van het KP5-programma EUROMAP uitgevoerde *benchmark study* worden Nederland en België aangewezen als landen waar het taal- en spraaktechnologisch onderzoek op zeer hoog niveau staat. Dit wordt weerspiegeld in de bovengemiddelde deelname van Nederlandse en Vlaamse wetenschappers aan conferenties en in Europese samenwerkingsprojecten, als ook in de positie van Nederland en België op de Europese HLT-scorecard (figuur 1). Beide landen worden naast Duitsland, Groot-Brittannië, Frankrijk en Finland aangewezen als de "more technologically advanced members" die in staat zijn de technologie op de markt te brengen. Dit laatste wordt bevestigd in de M&I-Technologieverkenning.

<sup>16</sup> Dit aantal is stijgende en daarnaast is er een behoorlijk aantal anderen dat TST-vakken volgt (zie bijlage 5).

<sup>17</sup> Cijfers overgenomen uit het M&I/Partners rapport: Financieel volume is niet exact, maar afgeleid op basis van aantal en type fte's.

Hoewel er een aanzienlijke taalspecifieke component in de technologie zit, zijn met name de algoritmen waarmee die kennis verworven en toegepast kan worden, in ruime mate taalafhankelijk. De basiskennis voor geavanceerde taal- en spraaktechnologie is op alle natuurlijke talen toepasbaar. Het onderzoek in Nederland en Vlaanderen kan voortbouwen op en bijdragen aan het internationale onderzoek.



Figuur 1: European HLT Scorecard (overgenomen uit Benchmarking HLT Progress in Europe, 2003)

Het onderzoek op het gebied van de taal- en spraaktechnologie wordt sterk gedomineerd door een focus op het (Amerikaanse) Engels. Daarnaast is de financiering voor het onderzoek traditioneel zeer sterk in de Verenigde Staten, met name het werk in de onderzoekslaboratoria van de grote telefoonmaatschappijen (vooral AT&T) en overheidsinstanties als DARPA en NSA. Ook de grote computer- en softwarefabrikanten en dienstverleners zoals IBM, Microsoft en General Electric doen grote investeringen in de ontwikkeling van taal- en spraaktechnologie, gericht op het ontsluiten en toegankelijk maken van informatie in gesproken en geschreven documenten. De meest vooraanstaande instituten zijn momenteel Microsoft Research, IBM Research, BBN Technologies, Carnegie Mellon University, MIT, SRI International, en International Computer Science Institute. In Japan hebben MITI en ATR en bedrijven als NTT, NEC, Fujitsu, Mashusitsa en Sony grote belangen in de ontwikkeling van taal- en spraaktechnologie. Daarbij ligt de nadruk vaak sterker op de automatische verwerking van gesproken taal, omdat het typen van Japanse karakters erg lastig is. Om dezelfde reden wordt veel onderzoek gedaan naar de automatische verwerking van (gesproken en geschreven) Chinees, Koreaans en andere zogenaamde "16-bit talen".

In Europa wordt veel onderzoek gedaan in Engeland, waar de corpora voor het Amerikaanse Engels relatief gemakkelijk aangepast en opnieuw gebruikt kunnen worden. De belangrijkste instituten zijn Cambridge University, Sheffield University, Edinburgh University, naast een aantal industriële onderzoekslaboratoria van o.a. Canon, Microsoft, Toshiba, en de Britse Intelligence diensten. Ook de Franse overheid investeert op grote schaal in taal- en spraaktechnologie, niet alleen ter ondersteuning van het Franse bedrijfsleven, maar ook voor de versterking van de positie van de Franse taal in internationaal verband. De meest vooraanstaande instituten zijn France Télécom R&D, LIMSI, INRIA, Université de Grenoble en Université d'Avignon. In Duitsland zijn het DFKI, de universiteiten van Aachen, München, en Nürnberg-Erlangen de meest vooraanstaande instituten. In Duitsland spelen ook de onderzoekslaboratoria van Philips, Siemens, en DaimlerChrysler een belangrijke rol.

Internationaal gezien neemt het Nederlandse onderzoek een vooraanstaande plaats in op een aantal TST-domeinen. Daartoe behoren zeker toepassingen als *Information Extraction* en Multimodale Interfaces. Daarnaast speelt Nederland een vooraanstaande rol op het gebied van de basistechnologie, vooral waar het gaat om de combinatie van regelgebaseerde en statistische verwerkingstechnieken. Nederland speelt ook een vooraanstaande rol op het gebied van automatische ontleding. In de "European HLT Scorecard" (figuur 1) opgesteld door het EUROMAP-consortium, komt Nederland naar voren als een van de Europese landen met de beste technologie en de beste kansen op de markt.

Vlaanderen scoort beter dan de positie “Belgium” doet vermoeden: de resultaten van Vlaanderen worden negatief beïnvloed door de resultaten van Wallonië. De Vlaamse TST-groepen zijn, met steun van het FWO-Vlaanderen, verenigd in de wetenschappelijke onderzoeksgemeenschap CLIF<sup>18</sup>. Mede door de samenwerking in dit kader zijn de verschillende groepen de laatste jaren erg succesvol geweest in competitieve fondsenwerving bij de nationale en Europese fondsen (IWT-Vlaanderen, FWO-Vlaanderen, de Europese Kaderprogramma’s). Het onderzoek van de groepen wordt internationaal goed onthaald, en de specialisaties van de verschillende centra zijn complementair (bijvoorbeeld statistische methodes in Antwerpen, grammaticaformalismen en spraakherkenning in Leuven, spraaksynthese in Gent).

### 3.3. Behoeften van de kennisinfrastructuur

In 2001 en 2002 is in opdracht van het TST-platform een uitgebreid onderzoek uitgevoerd om de prioriteiten vast te stellen voor de ontwikkeling van basisvoorzieningen voor de taal- en spraaktechnologie voor het Nederlands (de zogenaamde basis-taalinfrastructuur BaTaVo).<sup>19</sup> In dit kader is een uitgebreide inventarisatie en evaluatie gemaakt van bestaande basistaalinfrastructuur en ook een lijst met zaken die bij voorkeur zo spoedig mogelijk ontwikkeld zouden moeten worden. De kosten voor het realiseren van de elementen van deze prioriteitenlijst werden begroot op ruim 11 miljoen Euro. De prioriteitenlijst is weergegeven in figuur 2.

<i>Voor taaltechnologie:</i>
1a. Geannoteerd corpus geschreven Nederlands (een treebank met syntactische, eventueel morfologische structuren).
1b. Syntactische analyse. Robuuste herkenning van de structuur van zinnen in tekst. De te gebruiken technologie hiervoor kan klassiek zijn (grammatica en <i>parser</i> ), gebaseerd op statistische modellen, of op <i>shallow parsing</i> (constituentendetectie en toekennen van grammaticale relaties).
1c. Semantische annotaties voor de treebank (1a).
2. Robuuste modulaire tekstvoorverwerking: <i>tokenisation</i> (herkenning van voorkomens van woorden), indeling van tekst in zinnen, <i>named entity recognition</i> voor verschillende teksttypes.
3. Vertaalequivalenten in belangrijkste talen voor basislexicon.
<i>Voor spraaktechnologie:</i>
1. Automatische spraakherkenning, inclusief robuuste herkenning, herkenning van niet-standaard taalgebruik, adaptatie van de spraakherkenner, taalmodellen en prosodieherkenning.
2. Spraakcorpora voor specifieke applicaties zoals bijvoorbeeld customer care en Computer Assisted Language Learning (CALL). Multimediale spraakcorpora: corpora die naast spraak van radio en TV ook informatie van andere media bevatten (bijvoorbeeld teksten en figuren van WWW, kranten of tijdschriften).
3. Tools voor (semi-)automatische annotatie van spraakcorpora (labels op verschillende niveaus: segmenteel, prosodisch, syntactisch, semantisch, pragmatisch).

*Figuur 2: behoeften van de kennisinfrastructuur (overgenomen uit het TST platform rapport Het Nederlands in taal- en spraaktechnologie: prioriteiten voor basistaalvoorzieningen, 2002)*

In het kader van de M&I-Technologieverkenning is de kennisinstellingen gevraagd naar de toekomstige ontwikkelingen die zij zien op het gebied van taal- en spraaktechnologie, welke toepassingsmogelijkheden een grote vlucht zullen gaan nemen, en welke “quick wins” er te behalen zijn. Daarnaast is ook de vraag gesteld op welke gebieden de ontwikkelingen achterblijven, en wat

<sup>18</sup> Computational Linguistics in Flanders, zie <http://pcger33.uia.ac.be/clif/>.

<sup>19</sup> Zie : *Het Nederlands in taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen*, eds. Walter Daelemans & Helmer Strik: <http://taalunieversum.org/taal/technologie/docs/daelemans-strik.pdf>

de oorzaken daarvan zijn. De visies van de verschillende kennisinstellingen hierop kwamen sterk overeen.

Toekomstige ontwikkelingen ziet men op een aantal terreinen:

- semantische analyse van teksten ten behoeve van een betere ontsluiting: informatie-extractie, (content-based) information retrieval, kennismanagement, robuuste classificatie van teksten, automatische vertaling;
- vraag-antwoord systemen, dialoogsystemen (zowel op taal- als spraaktechnologisch gebied);
- ontsluiting van gesproken audio-archieven, multimediale ontsluiting;
- robuuste spraakherkenning (mobiele toepassingen via GSM, PDA);
- robuuste spraaksynthese (IVR, Interactive Voice Response);
- sprekerherkenning ten behoeve van toegangscontrole en beveiliging (authenticatie).

TST-gebieden waarop ontwikkelingen achterblijven, liggen vooral op het gebied van de Nederlandstalige taal- en spraaktechnologie: herkenning van Nederlandse spraak, syntactische/semantische analyse van het Nederlands, en spraaksynthese voor het Nederlands. Voor de achterblijvende ontwikkelingen ziet men een drietal oorzaken:

- Het onderzoek is gefragmenteerd, en wordt op veel verschillende plaatsen uitgevoerd, waarbij te weinig gebruik wordt gemaakt van elkaars onderzoeksresultaten.
- Er is geen commercieel belang voor het Nederlands, het is een te klein taalgebied. Daardoor zijn er ook te weinig hulpmiddelen, zoals corpora en *benchmarks*.
- De industrie kan niet meer investeren, en universiteiten zijn te sterk afhankelijk van korte termijn financiering.

Daarnaast wordt het belang van de combinatie van fundamenteel taalkundige en datageoriënteerde statistische methoden benadrukt, en vindt men onderzoek daarnaar ook belangrijk.

## 4. Het bedrijfsleven

### 4.1. Bedrijfsleven in Nederland en Vlaanderen

In het technologieoverdrachtmodel wordt een onderscheid gemaakt tussen "technologieontwikkelaars" en "applicatieontwikkelaars". Technologieontwikkelaars zijn bedrijven die taal- en spraaktechnologische modules en/of halffabrikaten ontwikkelen en commercialiseren. Soms kunnen deze als zelfstandig product in de markt gezet worden, maar meestal is er sprake van *enabling software* die ingebouwd wordt in grotere applicaties en/of diensten. Applicatieontwikkelaars zijn bedrijven die taal- en spraaktechnologie in hun (eind)producten en/of diensten integreren, en zijn dus in feite "intermediaire gebruikers" van deze technologie. Deze bedrijven zijn vooral actief in de sectoren van de kantoorautomatisering (inclusief kennismanagement), informatie- en communicatietechnologie en de telecommunicatie. In de Technologieverkenning Nederlandstalige Taal- en Spraaktechnologie van M&I/Partners wordt deze indeling verder uitgewerkt in een model waarin het innovatieproces als een gelaagde keten wordt gerepresenteerd (zie verder sectie 5.1). In Nederland en Vlaanderen zijn ruim 50 bedrijven min of meer direct betrokken bij één of meer lagen van het TST-innovatieproces. Een overzicht staat op:

<http://taalunieversum.org/taal/technologie/ontwikkelaars.php>.

Een aantal Nederlandse MKBs heeft zich samen met een aantal kennisinstellingen met expertise op het gebied van taal- en/of spraaktechnologie verenigd in de Stichting NOTaS<sup>20</sup>. Deze stichting heeft tot taak de belangen te behartigen van bedrijven en kennisinstellingen die actief zijn op het terrein van taal- en spraaktechnologie en de sector een eigen gezicht te geven in binnen- en buitenland.

### 4.2. Bedrijfsleven in de rest van Europa/de wereld

Volgens de EUROMAP *benchmark study* zijn er circa 300 commerciële bedrijven in Europa die taal- en spraaktechnologische producten op de markt brengen. In het rapport wordt een twintigtal bedrijven

---

<sup>20</sup> Nederlandse Organisatie voor Taal- en Spraaktechnologie, zie <http://www.stichtingnotas.nl/>.

opgevoerd als *showcase*; drie daarvan zijn Nederlandse dan wel Vlaamse bedrijven. De website van ELSNET (European Network of Excellence in Language and Speech Technology), opgezet met subsidie uit verschillende Kaderprogramma's bevat circa 2500 organisaties die op dit gebied actief zijn.<sup>21</sup> Het grootste deel daarvan, circa 1500, bevindt zich in de Verenigde Staten. Onder deze bedrijven is een aantal grote multinationale computer- en software fabrikanten, zoals IBM, Microsoft en General Electric en een aantal grote telefoonmaatschappijen, zoals AT&T, BT, France Télécom en NTT. Daarnaast zijn er vele technologieontwikkelaars waaronder BBN, SRI, Scansoft en SYSTRAN en applicatieontwikkelaars zoals Philips, Bosch, Siemens en DaimlerChrysler AG, waarbij opgemerkt moet worden dat de scheiding tussen deze categorieën niet altijd even scherp getrokken kan worden.

Met name in de ontwikkeling van de spraaktechnologie is een slingerbeweging te zien: veel bedrijven hebben hun strategie afwisselend gericht op alleen maar technologieontwikkeling of op de gecoördineerde ontwikkeling van basistechnologie en applicaties. Momenteel lijkt er overeenstemming te bestaan over het feit dat verschillende soorten toepassingen (bijvoorbeeld dicteren, telefonische informatiediensten, *command & control* in de auto) zulke specifieke eisen stellen aan de technologie dat het niet mogelijk is om bijvoorbeeld een spraakherkenner te bouwen die voor alle toepassingen geschikt is. Met andere woorden: ook binnen bedrijven is het inzicht gegroeid dat een integrale ketenbenadering van cruciaal belang is.

### 4.3. Toepassingen

Eindgebruikers komen vooral in aanraking met complete toepassingen waarin taal- en spraaktechnologie is ingebed, zoals automatische vertaling op internet, automatische vertaalhulpsystemen voor o.a. lokalisatie van software en gebruiksaanwijzingen, dicteerpakketten, spraakinterfaces naar informatiesystemen via de telefoon, systemen voor *Computer Assisted Language Learning*, voorleessoftware voor visueel gehandicapten, automatische ondertiteling, transcriberen van en zoeken op audio-archieven, spraak als onderdeel van *ambient technology*, informatie-extractiesystemen, kennismanagementsystemen, etc. Dit zijn complexe systemen die gebruik maken van verschillende taal- en spraaktechnologische componenten (vaak *lingware* genoemd) als lemmatisering en woordsoort-disambiguering, en van dataverzamelingen (*resources*) als lexica en corpora. Verschillende van deze "halffabrikaten" (op zichzelf hebben ze slechts beperkt commercieel belang) spelen een meer of minder belangrijke rol in de ontwikkeling van verschillende toepassingen. In veel gevallen geeft taal- en/of spraaktechnologie een toegevoegde waarde aan een product. Spellingcheckers zijn daarvan het meest bekende voorbeeld. Grammaticacheckers en stijlcheckers zijn ook vaak ingebouwd, maar leveren niet dezelfde gebruikerstevredenheid.

In bepaalde gevallen wordt de TST-component intern ontwikkeld door het bedrijf dat ook de volledige toepassing maakt, vaak echter ook - en vooral bij de "kleinere talen" - wordt de betreffende *lingware* aangekocht van een externe partij. Microsoft maakt bijvoorbeeld zijn eigen spellingcheckers voor de grootste talen, maar koopt ze aan voor het Nederlands. Vanwege de functie die taal- en spraaktechnologie heeft in toepassingen, wordt het ook wel een *enabling technology* genoemd. In dit verband is het relevant te vermelden dat Nederlandse en Vlaamse bedrijven tot de belangrijkste leveranciers van multilinguale *lingware* behoren. Na de recente consolidatie in de spraaktechnologie is Scansoft - met een belangrijke vestiging in Vlaanderen - wereldwijd met afstand de grootste onafhankelijke leverancier van spraaktechnologie. Polderland behoort tot de belangrijkste leveranciers van taaltechnologie voor Microsoft. Een uitgebreide beschrijving van taal- en spraaktechnologische toepassingen, modules en componenten is te vinden in het reeds eerder genoemde rapport "*Het Nederlands in Taal- en Spraaktechnologie*" dat in opdracht van het TST-platform is opgesteld. Ook in de M&I-Technologieverkenning wordt een groot aantal toepassingen opgesomd. De meeste hiervan zijn hierboven genoemd.

### 4.4. Behoeften van het bedrijfsleven

In de afgelopen jaren is een aantal verkenningen uitgevoerd naar activiteiten en wensen van het bedrijfsleven in de taal- en spraaktechnologie. In 2001 en 2002 is in opdracht van het TST-platform een uitgebreid onderzoek uitgevoerd om de prioriteiten vast te stellen voor de ontwikkeling van basisvoorzieningen voor de taal- en spraaktechnologie voor het Nederlands (de zogenaamde basis taalinfrastructuur).<sup>22</sup> In dit kader is een groot aantal *stakeholders* bevroegd. Daarnaast is in ongeveer

---

<sup>21</sup> Zie <http://www.elsnet.org/expertframes.html>.

<sup>22</sup> Zie: *Het Nederlands in taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen*, eds. Daelemans & Strik: <http://taalunieversum.org/taal/technologie/docs/daelemans-strik.pdf>

dezelfde periode een *benchmark study* uitgevoerd in het kader van het Europese Euromap project.<sup>23</sup> Tot slot heeft M&I/Partners een behoorlijke groep bedrijven ondervraagd in het kader van de technologieverkenning die ze hebben uitgevoerd in opdracht van het Nederlandse Ministerie van Economische Zaken.

Uit de laatstgenoemde verkenning komen onderstaande conclusies naar voren, waarbij opgemerkt wordt dat door de beperkte retourzending van de enquêteformulieren door bedrijven en het niet prijsgeven van vertrouwelijke informatie omtrent hun strategische richtingen het onmogelijk is de huidige prioriteiten voor het volledige TST-veld in Vlaanderen en Nederland nauwkeurig en gedetailleerd te schetsen. Op basis van de ingevulde enquêteformulieren en andere informele feedback kunnen de volgende tendensen wel afgeleid worden, zoals weergegeven in figuur 3.

<i>Voor taaltechnologie:</i>
<ul style="list-style-type: none"> <li>• met stip: semantische analyse</li> <li>• grafeem-foneemomzetting</li> <li>• tekstvoorverwerking dialogsystemen.</li> </ul>
<i>Voor spraaktechnologie:</i>
<ul style="list-style-type: none"> <li>• met stip: robuuste spraakherkenning;</li> <li>• adaptatie;</li> <li>• betrouwbaarheidsmaten</li> </ul>
<i>Toepassingen taaltechnologie:</i>
<ul style="list-style-type: none"> <li>• retrieval NAW gegevens;</li> <li>• retrieval van teksten;</li> <li>• vertalen.</li> </ul>
<i>Toepassingen spraakherkenning en -synthese:</i>
<ul style="list-style-type: none"> <li>• Interactive Voice Respons Systemen;</li> <li>• Call Center toepassingen.</li> </ul>
<i>Waar zien bedrijven toegevoegde waarde van kennisinstellingen? Dat is bij:</i>
<ul style="list-style-type: none"> <li>• ontwikkeling van corpora;</li> <li>• <i>benchmarking</i> en standaardisatie;</li> <li>• onderzoek naar gebruik statistiek en heuristiek.</li> </ul>

Figuur 3: behoeften van het bedrijfsleven (tabel overgenomen uit M&I Technologieverkenning, 2004)

De Programmacommissie is van mening dat naast de in de M&I-verkenning genoemde behoeften ook onderzoek en ontwikkeling van *resources* ten behoeve van multimedia-informatie-extractie en multimodale interactie een belangrijke toegevoegde waarde hebben. Naast de technisch-inhoudelijke onderwerpen hecht het bedrijfsleven grote waarde aan een goede toegankelijkheid van bestaande data en tools, en aan een goede IPR-regeling voor nieuw te ontwikkelen hulpbronnen.

## 5. Opzet integraal TST-meerjarenprogramma STEVIN

### 5.1. De ketenbenadering

In de Technologieverkenning Nederlandstalige Taal- en Spraaktechnologie van M&I/Partners wordt uitgebreid uiteengezet hoe het TST-innovatieproces kan worden beschouwd als een gelaagde keten met vier lagen (zie figuur 4). Op al deze lagen zijn verschillende soorten bedrijven actief. Voor zover het gaat om de basistaalvoorzieningen (laag 1) betreft het private partijen die bestanden bezitten gericht op een specifiek marktsegment, maar ook woordenboekfabrikanten. In het TST-ontwikkelingsproces (laag 2) zijn twee typen bedrijven te onderscheiden: de TST-componenten-

<sup>23</sup> Zie *Benchmarking HLT Progress in Europe*: <http://www.hltcentral.org/>

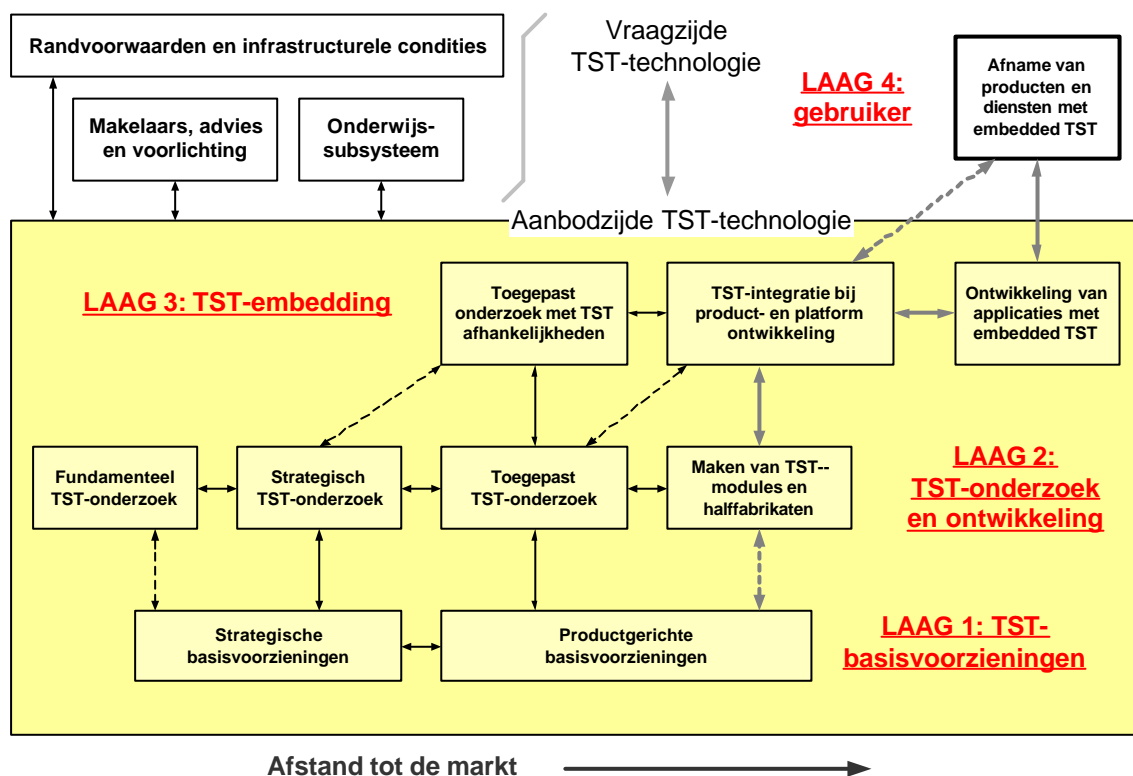
bouwers die halffabrikaten vervaardigen die ze weer doorleveren aan platformbouwers en integratoren. Daarnaast zijn ook applicatieontwikkelaars actief die platformen en TST-componenten aankopen en daarmee in eigen beheer of in opdracht producten en/of diensten maken. Tot slot, in het commercialisatieproces (lagen 3 en 4) zijn de volgende actoren actief: de platformbouwers, die hun eigen commerciële producten of diensten willen verbeteren door het inbouwen van TST, de distributeurs van producten en de uitbaters van diensten (o.a. banken, de overheid, call centres).

Soms komt dit onderscheid echter vrij kunstmatig over en is er in de realiteit sprake van een diffusie tussen de lagen, aangezien sommige bedrijven enerzijds taal- en spraaktechnologie van andere bedrijven gebruiken, en anderzijds deze technologieën ook zelf (verder) ontwikkelen.

In figuur 4 staat langs de horizontale as de afstand tot de markt uit met spelers in de keten die dichter of verder van de markt afstaan. Verticaal zijn de volgende lagen te onderscheiden:

1. de basistaalvoorzieningen, uitdrukkelijk beperkt tot de data zelf en het aanmaakproces ervan;
2. TST-onderzoek en -ontwikkeling, uiteindelijk resulterend in beschikbare TST-componenten;
3. TST-embedding;
4. de vraagzijde.

Lagen 1-3 vormen samen het aanbodsysteem. In hoofdstuk 6 wordt in meer detail uiteengezet hoe ook binnen het aanbodsysteem sprake is van leverancier-klant relaties, en hoe die van invloed zijn op de organisatie van STEVIN.



Figuur 4: Gelaagd model voor het TST-innovatiesysteem (figuur overgenomen uit de *Technologie-verkenning Nederlandstalige taal- en spraaktechnologie*, M&I/Partners BV en Montemore NV (2004)

#### Laag 1. TST-basisvoorzieningen

Basistaalvoorzieningen zijn ofwel generisch van aard (voor een bepaalde taal) en veralgemenend ofwel zijn ze gericht op de verdere ontwikkeling van bepaalde toepassingen. Ze zijn essentieel zowel voor het onderzoek als voor een brede waaier van productontwikkelingen. Vanaf het begin moet herbruikbaarheid over een langere periode nagestreefd worden. Gezien het arbeidsintensieve karakter moet dit type basistaalvoorziening bij voorkeur slechts één keer (met gezamenlijke inspanning) vervaardigd worden en vervolgens breed beschikbaar gemaakt worden tegen een redelijke prijs. Het betreft vooral geannoteerde lexicons, thesauri en corpora.

#### Laag 2. TST-onderzoek en ontwikkeling

Laag 2 heeft betrekking op het fundamentele en strategische onderzoek dat door kennisinstellingen en bedrijven wordt uitgevoerd en aan de ene kant noodzakelijk is voor de ontwikkeling van bepaalde

basistaalvoorzieningen in laag 1 en aan de andere kant voor de ontwikkeling van de TST-componenten in laag 2. In deze laag horen thuis:

- regels en grammatica's zoals o.a. fonologische regels voor grafeem-foneem omzetting; grammatica's voor syntactische analyse; morfologische regels etc.
- modules en basiscomponenten zoals grafeem-foneemomzetting; tekstvoorverwerking (o.a. het detecteren van zinsgrenzen, datums, eigennamen, tijdstippen, afkortingen); morfologische analyse; syntactische analyse; semantische analyse; tekstgeneratie; spraaksynthese; spraakherkenning; foneem-grafeemomzetting; prosodiegeneratie; prosodieherkenning; sprekerherkenning; taal- en dialectidentificatie.

De componenten kunnen onderverdeeld worden in twee klassen:

- Modules zijn componenten die verdere ontwikkeling ondersteunen. De gebruiker is bij voorkeur een andere TST-expert die de zwaktes kan relativeren en weet te omzeilen. Modules zullen veelal puur ondersteunend werken in het ontwikkelingsproces van een andere TST-component. Anderzijds kunnen ze ook dienen als basistechnologie waaruit een commercieel product wordt ontwikkeld.
- Halffabrikaten zijn afgewerkte producten die klaar zijn voor integratie in andere halffabrikaten of rechtstreeks in eindproducten. De klant is mogelijk een andere TST-ontwikkelaar die dit halffabrikaat in een groter geheel inbouwt, maar wellicht frequenter is het een niet-TST-specialist die deze kant-en-klare module in zijn toepassing inbouwt.

In de overgrote meerderheid van de toepassingen vormt taal- en spraaktechnologie slechts een component(je) van een veel groter geheel en soms is de eindgebruiker zich weinig bewust van de onderliggende technologie. Voorbeelden zijn legio: spellingcheckers in wordprocessors; spraaksynthese die gebruikt wordt om informatie die is opgeslagen in een database weer te geven over de telefoon; spraakherkenning die een informatiesysteem of een boekingsstelsel stuurt.

### Laag 3. Applicatieontwikkeling (TST-embedding)

Laag 3 heeft betrekking op toegepast onderzoek naar en ontwikkeling van toepassingen voor eindgebruikers waarin gebruik gemaakt wordt van een TST-component. Verschillende toepassingen zijn reeds eerder beschreven in sectie 4.3. Het profiel van de integratoren is zeer uiteenlopend. Het betreft onder meer: integratoren in de telefonie (Logica-CMG, VOXTRON) die werken in opdracht van de dienstensector (bijvoorbeeld banken, openbaar vervoer, telecom operatoren); ontwikkelaars van consumentendiensten en/of professionele apparatuur (Philips, Nokia, Bosch, Siemens); software-ontwikkelaars (content management, tekstverwerking, vertaler *workbenches*, taalleersoftware).

## **5.2. Doelstellingen en integrale aanpak**

De hoofddoelstelling van STEVIN is het geven van een stimulans aan een technologiesector met een aanzienlijke economische potentie met de bedoeling om de innovatiecapaciteit van die sector te vergroten en tegelijkertijd de positie van het Nederlands in de moderne informatie- en communicatiewereld te behouden.

In de technologieverkenning van M&I/Partners is een sterkte-zwakte-analyse gemaakt van het TST-innovatiesysteem. Daaruit is naar voren gekomen dat de wetenschappelijke kwaliteit van zowel het Nederlandse als Vlaamse TST onderzoek internationaal gezien op hoog niveau staat (zie ook de EUROMAP *benchmark study*). Zowel Nederland als Vlaanderen kennen een traditie waarin samenwerkingsprojecten worden opgezet tussen de "taalindustrie" en de kennisinstituten. Aangezien de betrokken industrie voornamelijk kleine tot middelgrote partijen betreft kan men zich geen samenwerking puur voor de etalage permitteren maar wordt zakelijk bekeken wat samenwerking concreet kan opleveren. Ook is er een aantal zwakke punten aan te wijzen: er is een gebrek aan basistaalvoorzieningen; de organisatie van de TST-sector behoeft verbetering om meer continuïteit te bewerkstelligen; IPR-zaken, standaardisatie, onderhoud, beheer en exploitatie dienen meer structureel aangepakt te worden; de vraagkant dient nader ontwikkeld te worden.

Deze situatie vraagt om een integrale aanpak waarvan de effecten zorgvuldig gemonitord moeten worden. STEVIN zal daarom een hybride karakter hebben waarin een combinatie van stimuleringsinstrumenten wordt ingezet. De doelstellingen van het programma kunnen puntsgewijs als volgt worden samengevat:

- het realiseren van een adequate digitale taalinfrastructuur voor het Nederlands, gebaseerd op de BaTaVo-prioriteitenlijst;

- het doen van strategisch onderzoek op het gebied van taal- en spraaktechnologie, met name op gebieden waar grote vraag naar is vanuit concrete toepassingen van de technologieën;
- netwerk- en zwaartepuntvorming en verankering, het opleiden van nieuwe experts, kennisoverdracht, vraagstimulering en het adequaat regelen van intellectuele eigendomsrechten.

Het programma zal in hoofdzaak uitgevoerd worden door kennisinstellingen maar met steun van en samenwerking met het bedrijfsleven zodat de ontwikkelde basistaalvoorzieningen en het uitgevoerde onderzoek nauw aansluiten bij daadwerkelijke behoeften van de TST-leveranciers, de applicatieontwikkelaars en hun klanten.

Het adequaat regelen van intellectuele eigendomsrechten van de in te brengen en te produceren basistaalvoorzieningen en tools, de standaardisatie daarvan, het onderhoud, het beheer en de exploitatie zullen een integraal onderdeel vormen van STEVIN en in nauwe samenwerking met de TST-centrale worden gerealiseerd. Voor de organisatie van het programma is uitgegaan van de volgende basisprincipes: sterke regie, goede bestuurbaarheid, concrete richting en doelstellingen, en haalbaarheid van het onderzoek.

### 5.3 BaTaVo: data, tools, modules en onderzoeksthema's

In deze sectie wordt een onderscheid gemaakt tussen basisvoorzieningen (laag 1), strategisch onderzoek (laag 2) en toepassingen (laag 3). Dit spoort met de indeling die wordt gehanteerd in de Technologieverkenning Nederlandstalige Taal- en Spraaktechnologie, en het correspondeert ook met de driedeling die wordt gebruikt in de Human Language Technologies Roadmap<sup>24</sup>, waarin men spreekt van *Language Resources*, *Language Processing* en *LangTech Applications*.

De TST-basisvoorzieningen zijn enerzijds bedoeld ter ondersteuning van de verdere ontwikkeling (door spraak- en/of taaltechnologen) van *state-of-the-art* TST voor het Nederlands (= *technology push*), en anderzijds ter ondersteuning van bedrijven die deze technologie willen integreren in nieuwe toepassingen (= *demand pull*). Het betreft hierbij zowel data als tools. Gevolg gevend aan de aanbevelingen in de M&I-Technologieverkenning worden deze twee types basisvoorzieningen dan ook duidelijk van elkaar onderscheiden, maar het lijkt het meest vruchtbaar om data en tools in samenhang te ontwikkelen. Weliswaar is een initiële verzameling data nodig om tools te ontwikkelen (zodat er sprake lijkt te zijn van een vicieuze cirkel), maar anderzijds kunnen deze tools vervolgens ingezet worden om de overige benodigde data op de meest efficiënte en consistente manier te produceren - bijvoorbeeld door de tools als *bootstrapper* te gebruiken, gevolgd door manuele verificatie (mogelijk ook ondersteund door tools), en formele validatie van de aangebrachte correcties met behulp van tools. De noodzaak voor een dergelijke opzet wordt groter naarmate de complexiteit van de annotaties bij de data toeneemt. Met een dergelijke aanpak zal de kwaliteit, de formele correctheid en de consistentie van de geproduceerde data en tools op efficiënte wijze het best gegarandeerd kunnen worden en zal de schijnbare vicieuze cirkel omgezet worden in een zichzelf versterkende ontwikkelingspiraal.

In deze secties worden alle basistaalvoorzieningen genoemd waarvan bekend is (o.a. uit de BaTaVo-prioriteitenlijst en de M&I-Technologieverkenning) dat ze noodzakelijk of nuttig zijn voor de verdere ontwikkeling van de Nederlandstalige taal- en spraaktechnologie, inclusief deze waarvan men mag aannemen dat ze reeds voor een deel beschikbaar zijn bij bedrijven of kennisinstellingen. Om duplicatie bij de creatie van basistaalvoorzieningen te vermijden en om te beletten dat de opname van data in de BaTaVo de concurrentiepositie van TST-bedrijven in gevaar zou kunnen brengen (zie ook sectie 5.4) zal er enerzijds een actie opgezet worden om direct aan de bedrijven in het veld te vragen welke basistaalvoorzieningen zij in hun bezit hebben en ter beschikking willen stellen voor opname in de BaTaVo, en tegen welke voorwaarden (zulks in nauwe samenwerking met de TST-centrale). Anderzijds zal ook geëist worden dat aanvragers zich terdege oriënteren op de beschikbaarheid van basistaalvoorzieningen die relevant zijn voor hun projecten.

Voor al deze basistaalvoorzieningen geldt dat voor de ontwikkeling ervan een specificatie opgesteld dient te worden, evenals validatiecriteria, en dat validatie door een externe partij reeds tijdens de ontwikkelingsstijd en natuurlijk ook na oplevering een integraal onderdeel dient te vormen van ieder projectvoorstel. Dat is een beproefde methode om de kwaliteit van de opgeleverde basistaalvoorzieningen te garanderen.

Het strategisch onderzoek is ondermeer nodig om ervoor te zorgen dat *state-of-the-art* technologie effectief ook voor het Nederlands ter beschikking komt, en dat er voldoende expertise in Vlaanderen

---

<sup>24</sup> Deze is te vinden op: <http://elsnet.dfki.de>.

en Nederland wordt opgebouwd om een optimale kennisoverdracht tussen kennisinstellingen, TST-bedrijven en TST-gebruikers te bewerkstelligen.

De basistaalvoorzieningen en het strategisch onderzoek worden hieronder beschreven volgens hun technologisch karakter (spraak- versus taaltechnologie). Voor de toepassingen waarin vaak (meestal) beide aspecten aan bod komen, is dit onderscheid niet gemaakt.

### 5.3.1 Spraaktechnologische basisvoorzieningen (laag 1)

#### Te ontwikkelen data die geen voorafgaand onderzoek vergen

Het gaat hier in het bijzonder over corpora die nodig zijn voor de ontwikkeling van spraaktechnologie in toepassingen die door TST bedrijven als belangrijke testcases voor het demonstreren van de mogelijkheden van TST worden beschouwd. Voorbeelden hiervan zijn:

1. spraakcorpora en multimodale corpora die nodig zijn in het kader van toepassingen zoals CALL (Computer Assisted Language Learning), NAW (Naam Adres Woonplaats) of CCQA (vragen en antwoorden in call centers), educatieve toepassingen voor kinderen, etc.;
2. multimediale corpora (spraak + andere media) die belangrijk zijn voor de ontwikkeling van toepassingen waarin men b.v. BNT (Broadcast News Transcription) (audio + video) of persoonsidentificatie (spraak + beeld) nodig heeft;
3. tekstcorpora die nodig zijn voor de ontwikkeling van stochastische taalmodellen zoals vereist voor continue spraakherkenning, een belangrijke component in de toepassingen genoemd onder 1 en 2.

De prioriteit die men toekent aan de realisatie van een corpus hangt samen met de prioriteit die men geeft aan de toepassing waarvoor het corpus bedoeld is. Voor de projecten in dit luik gelden de volgende aanbevelingen:

1. De voorkeur zal uitgaan naar de ontwikkeling van corpora ter ondersteuning van die toepassingen waarvoor veel interesse (en bereidheid tot het effectief ontwikkelen ervan) vanuit de bedrijfs wereld bestaat.
2. Bij de ontwikkeling van spraakcorpora dient men aandacht te hebben voor de samenstelling van goed gebalanceerde testsets waarmee men de ermee te ontwikkelen spraaktechnologie kan evalueren. Meerdere testsets zijn ondermeer nodig om verschillende aspecten van robuustheid te kunnen evalueren.
3. Ook bij de ontwikkeling van tekstcorpora dient men testsets samen te stellen waarop men methodes voor het ontwikkelen van taalmodellen kan evalueren.

**Projectsoort:** de projecten die men in dit luik verwacht, bevatten geen noemenswaardige onderzoekscomponent en hebben een zeer specifiek doel, namelijk de creatie van welomschreven corpora.

#### Te ontwikkelen voorzieningen waarvoor nog wel onderzoek nodig is

De hier te ontwikkelen voorzieningen zijn generiek van aard en dienen twee doelstellingen:

1. niet-spraaktechnologen de mogelijkheid geven om spraaktechnologische systemen te bouwen die goed beantwoorden aan hun onderzoeks- en/of integratienoden;
2. spraaktechnologen de mogelijkheid geven om bepaalde deelaspecten van de spraaktechnologie voor het Nederlands verder te ontwikkelen zonder daarbij ook alle andere componenten zelf te moeten ontwikkelen.

Het gaat hier om een modulair opgezette spraakherkenner die uit verschillende losse componenten bestaat. Niet-spraaktechnologen kunnen hiermee een gewenste configuratie samenstellen, terwijl spraaktechnologen een component door een ander, met een gewijzigde configuratie, kunnen vervangen. Verwacht wordt dat de aanmaak van goede voorzieningen nog heel wat onderzoeksinspanningen zal vergen (zie ook laag 2).

Uit verkennend onderzoek naar de mogelijke toepassingen van taal- en spraaktechnologie en naar de tekortkomingen van de huidige technologie voor het Nederlands volgt dat er (in volgorde van prioriteit) nood is aan voorzieningen voor de ontwikkeling van robuuste *spraakherkenning*, automatische *annotatie van corpora*, en *spraaksynthese*. Hieronder volgt een niet-exhaustieve

opsomming van voorzieningen waarvan men vindt dat ze dienen te worden ontwikkeld of publiek beschikbaar gemaakt.

#### 1. *Tools en data voor de ontwikkeling van spraakherkenners voor het Nederlands:*

- Akoestische modellensets (Vlaamse en Nederlandse) voor de herkenning van respectievelijk (i) geïsoleerde woorden over de telefoon (b.v. voor NAW), (ii) continue breedbandspraak (b.v. voor BNT) en (iii) continue telefoonspraak (b.v. voor Customer Care).
- Uitspraaklexica (Vlaamse en Nederlandse) voor de herkenning van verschillende types spraak zoals gelezen spraak, spontane spraak, snelle spraak, etc. Het gaat hierbij over lexica met uitspraakvarianten (en hun kansen van optreden) die zijn aangepast aan het spraaktype.
- G2P-tools voor (multiple) transcripties van Nederlandse en veel voorkomende niet-Nederlandse woorden.
- Een stochastisch taalmodel voor het Nederlands dat als basis kan dienen voor continue spraakherkenning in verschillende toepassingen (zelfs al is het model verkregen op basis van materiaal uit een beperkt domein).
- Een tool voor het bouwen (niet het ontwikkelen) van spraakherkenners die gebruik maken van de hiervoor vermelde taalmodellen, uitspraaklexica en akoestische modellen, en die in staat zijn meerdere hypothesen met hun confidenties te genereren. Verder dienen ook scripts te worden aangemaakt voor het configureren van sleutelwoordherkenning (1 of meerdere woorden in een draaguiting) en continue spraakherkenning (herkenning van volledige uitingen) met deze tool.
- Tools voor het aanmaken van formele grammatica's voor het bouwen van eenvoudige dialoogsystemen.
- Tools voor het aanmaken van meer generieke stochastische taalmodellen voor continue spraakherkenning, dit zijn modellen die aan een nieuw domein zijn aan te passen zonder dat daarbij alle taalmodelwaarschijnlijkheden opnieuw dienen getraind te worden.

Bij de ontwikkeling van een tool voor het bouwen van een spraakherkenner zal men zoveel mogelijk gebruik maken van publiek beschikbare software.

#### 2. *Tools voor de automatische annotatie van spraakcorpora:*

- Tools voor de automatische segmentatie van orthografisch getranscribeerde spraak in prosodische eenheden (b.v. de prosodische frasen die men bij synthese wenst te gebruiken) en voor de prosodische labeling van deze prosodische eenheden (b.v. prominentie van woorden, grenssterktes tussen opeenvolgende woorden, soort intonatiepatroon, etc.).
- Tools voor de automatische segmentatie van spraak met een gekende orthografische en POS transcriptie in respectievelijk syntactische, semantische en pragmatische eenheden (b.v. NP, PP, etc.).

#### 3. *Tools voor de ontwikkeling van spraaksynthese in het Nederlands:*

Dat de ontwikkeling van spraaksynthese slechts in derde prioriteit gerangschikt is, heeft vooral te maken met het feit dat er reeds goede commerciële spraaksynthesizers voor het Nederlands bestaan, en dat het weinig zin heeft om onafhankelijk van bedrijven binnen BaTaVo een *open source* spraaksynthesizer (zoals Nextens er een is) te willen ontwikkelen. De idee is dan ook dat projecten die het beschikbaar stellen van een spraaksynthesizer beogen, gehonoreerd kunnen worden op voorwaarde dat ze geruggensteund zijn door bedrijven met goede commerciële systemen.

**Projectsoort:** de projecten in dit luik zullen een onderzoekscomponent bevatten, maar ook een duidelijk engagement ten aanzien van de aanmaak van specifieke voorzieningen.

### 5.3.2 **Spraaktechnologisch strategisch onderzoek (laag 2)**

Op het gebied van spraakherkenning is er wat betreft het Nederlands op een aantal gebieden zoals robuuste herkenning, duidelijk een achterstand op de *state-of-the-art* voor de "grote talen". Er is op dit vlak dan ook nog veel strategisch onderzoek vereist om op zijn minst weer in de pas te lopen. Hoewel een groot deel van dit onderzoek in ruime mate taalonafhankelijk is, is het toch onontbeerlijk om het in het kader van dit TST-meerjarenprogramma te ondersteunen. Het is namelijk de enige manier om ervoor te zorgen dat *state-of-the-art* technologie ook effectief voor het Nederlands wordt aangewend en geëvalueerd. In wat volgt worden (in volgorde van prioriteit) een aantal belangrijk geachte onderzoeksthema's vermeld.

1. *Robuustheid*: onderzoek naar methodes die de spraakherkenning in een of ander opzicht kunnen verbeteren (betere modellen op de verschillende niveaus, behandeling van spontane spraakfenomenen, het genereren van zinvolle output voor woorden buiten het lexicon, etc.), en in het bijzonder ook methodes die specifiek zijn voor het Nederlands en verwante talen (gebruik van morfemen als lexicale basiseenheden, efficiënte behandeling van verbuigingen, etc.).
2. *Uitvoer*: onderzoek naar methodes voor het aanvullen van de ruwe uitvoer van de spraakherkenner, zoals het aanbrengen van punctuatie en hoofdletters, het aanbrengen van uitingtags (b.v. bevestiging, vraag om verduidelijking, etc.), o.a. door gebruik te maken van prosodie.
3. *Confidentiematen*: onderzoek naar betere maten voor de betrouwbaarheid van gegenereerde hypothesen, o.a. door de evaluatie van hypothesen op basis van formele taalkundige kennis (woordsoortinformatie, syntactische structuur, semantiek, etc.).
4. *Adaptatie*: onderzoek naar methodes voor de snelle en accurate detectie van situationele parameters (akoestiek, bandbreedte, spreekstijl, geslacht van spreker, accent, etc.) en voor de automatische aanpassing van voorhanden zijnde modellen (akoestische, lexicale, taalkundige) aan deze situaties.
5. *Lattices*: Bij het classificeren en *retrieven* van *Spoken Documents* kunnen de *lattices* de noodzakelijke extra informatie bieden. Onderzoek naar de gewenste diepte van de die *lattices* is echter noodzakelijk om een goede mix tussen overkill en snelheid te bepalen.

De voorkeur zal uitgaan naar strategisch onderzoek dat gebeurt in het kader van goede samenwerkingsverbanden. In het bijzonder is het aan te bevelen dat Nederlandse en Vlaamse onderzoeksgroepen rond bepaalde thema's met elkaar samenwerken.

### 5.3.3 Taaltechnologische basisvoorzieningen (laag 1)

In het domein van de taaltechnologie is er grote nood aan een drietal basisvoorzieningen (in volgorde van prioriteit): een Corpus Geschreven Nederlands, een multifunctioneel elektronisch lexicon en parallelle corpora.

#### 1. Corpus Geschreven Nederlands

Voor de ontwikkeling van een groot aantal TST-componenten is de beschikbaarheid van een elektronisch corpus geschreven Nederlands een absolute noodzaak. Om bruikbaar te zijn, moet het corpus voldoen aan een aantal vereisten in verband met omvang, representativiteit en taalkundige annotatie.

Qua omvang en representativiteit wordt gedacht aan 500 miljoen woorden, samengesteld uit verschillende soorten teksten, zowel qua register en genre als qua teksttype: niet alleen gedrukte teksten, maar ook webpagina's, getranscribeerde spraak, teksten van kinderen, e.d. In het corpus zullen ook, voor zover het juridisch mag en technisch kan, reeds bestaande corpora geheel of gedeeltelijk worden opgenomen.

Een vijfde van het corpus (100 miljoen woorden) dient te worden voorzien van een taalkundige annotatie die minstens de volgende lagen omvat: (1) POS tagging, (2) lemmatisering, (3) meerwoord-eenheden, (4) morfologische analyse van samenstellingen, en (5) semantische tagging.

Een klein deel van het corpus (10 miljoen woorden) dient ook voorzien te worden van een *tree bank*. Hiervoor dient de accuraatheid van de eerder genoemde annotaties nagenoeg perfect te zijn, wat in de praktijk zal betekenen dat er handmatige correcties noodzakelijk zullen zijn.

Een nog kleiner deel van het corpus (1 miljoen woorden) dient geselecteerd te worden uit een heel specifiek (nader te bepalen) domein. Dit subcorpus - indien rijk geannoteerd - zal onderzoek naar methodes voor snelle en efficiënte fijnafstemming van taal- en spraaktechnologie op een specifiek domein - een belangrijke factor voor de succesvolle inzet van deze technologieën - mogelijk maken.

Wat de annotatierichtlijnen voor de verschillende lagen betreft is het van belang dat aansluiting wordt gezocht bij internationale standaarden (b.v. TEI, EAGLES) en bij datgene wat reeds beschikbaar is voor het Nederlands (b.v. CGN, Parole).

#### 2. Elektronisch lexicon (incl. terminologiebank)

Naast het Corpus Geschreven Nederlands is ook een multifunctioneel Nederlands lexicon absoluut noodzakelijk voor taal- en spraaktechnologie. Het lexicon dient alle woorden af te dekken die in het Corpus Geschreven Nederlands voorkomen. Voor elk woord moet minimaal die informatie beschikbaar zijn die relevant is voor de annotatie van het corpus: frequentie (mogelijk op meerdere

niveaus), spelling, woordsoort, morfo-syntactische features, meerwoordige uitdrukkingen, collocatie, semantische features, verwijzing naar woordbetekenis zoals opgenomen in EuroWordNet, e.d.

Ook dient de relatie tussen *entries* uit het lexicon en daadwerkelijke voorkomens van woordvormen in het corpus (incl. spellingsvarianten, typefouten en *tokenization*-fouten) behandeld te worden. Het lexicon dient niet alleen de algemene woordenschat af te dekken, maar met name voor het domeinspecifieke subcorpus ook de relevante (mogelijk meerwoordige) terminologie van dit domein.

Het verdient aanbeveling om te starten met de reeds beschikbare elektronische lexica en terminologische databanken, zowel monolinguale als multilinguale (CELEX - CGN, RBN, Van Dale, NL-Translex, CLVV, Nederlandse EuroWordNet, etc.). Het bruikbaar maken van de in die lexica opgenomen informatie voor taal- en spraaktechnologische toepassingen is de centrale prioriteit en zij dienen verder uitgebreid en verbeterd te worden in functie van het Corpus Geschreven Nederlands.

Het is mogelijk het lexicon modulair in te richten, zodat er meerdere deellexicons zijn ieder met deeleigenschappen van de relevante *entries*, mogelijk ook geproduceerd door verschillende onderzoeksgroepen. Bij een dergelijke benadering dient echter wel gegarandeerd en regelmatig getoetst te worden dat de deellexicons naadloos op elkaar aansluiten en dat blijven doen tijdens de ontwikkeling ervan.

Gelet op de rol van het lexicon in functie van het Corpus Geschreven Nederlands is het van groot belang dat er voldoende overleg is tussen de consortia die verantwoordelijk zijn voor de bouw van het lexicon en de bouw van het corpus.

### 3. *Parallele corpora*

Bij de ontwikkeling van multilinguale toepassingen, zoals automatische vertaling en meertalige informatievergaring (IR), is het van belang om te kunnen beschikken over parallelle corpora. Die corpora dienen uiteraard het Nederlands te omvatten en moeten ook zijn opgelijnd, minstens op zinsniveau. De keuze van de talen, de omvang, de samenstelling en de aard van de annotaties zijn afhankelijk van de beoogde toepassing en dienen dan ook vanuit de noden van die toepassing te worden gemotiveerd.

**Projectsoort:** de projecten in dit luik zullen een onderzoekscomponent bevatten, maar tevens ook een duidelijk engagement ten aanzien van de aanmaak van specifieke voorzieningen.

## 5.3.4 Taaltechnologisch strategisch onderzoek (laag 2)

Binnen de taaltechnologie is het duidelijk dat er grote nood is aan strategisch onderzoek in de volgende domeinen (in volgorde van prioriteit): (1) semantische analyse, (2) tekstvoorverwerking (in de ruime zin van het woord), (3) morfologische analyse en (4) syntactische analyse. In termen van volume vergt tekstvoorverwerking het minste en semantische analyse het meeste onderzoek. Een goede tekstvoorverwerking is van kapitaal belang voor de spraaktechnologie (bijvoorbeeld voor de ontwikkeling van generieke stochastische taalmodellen). De andere taalkundige analyses kunnen uiteraard ook worden ingezet voor de verwerking van spraak, mits ze in probabilistische termen (zowel naar input als naar output) te formuleren zijn. Ze zijn in elk geval van groot belang voor de ontwikkeling van het Corpus Geschreven Nederlands (daar kunnen ze gebruikt worden ter versnelling van de corpusontwikkeling en kunnen ze meteen ook grondig gevalideerd worden).

### 1. *Semantische analyse*

Het onderzoek naar semantische analyse situeert zich voornamelijk op de volgende vlakken :

- *Semantische tagging.* Er dient te worden gezocht naar een *tagset* voor lexicaal-semantische analyse. Die neemt de vorm aan van een hiërarchisch gestructureerd systeem van onderscheidingen. Die moeten bruikbaar zijn voor de differentiatie van de verschillende betekenissen van eenzelfde woord (homonymie en polysemie) en moeten worden voorzien van precieze en operationeel bruikbare criteria voor de toekenning van *tags* aan woorden-in-context (desambiguering, selectierestricties). Het werk dient aan te sluiten bij internationale standaarden (Senseval, EuroWordNet, Simple, Framenet, propbank, etc.) en bij wat er reeds voor het Nederlands beschikbaar is (EuroWordNet-Dutch, etc.). Naast de *tagset* zelf moet er ook onderzoek gedaan worden naar methoden voor automatisch *taggen*, in het bijzonder word-sense-desambiguation en domeinclassificatie methoden voor het Nederlands. Samenstellingen en analyses van samenstellingen spelen daarbij een bijzondere rol, evenals het detecteren van nieuwe betekenissen of het semantisch duiden van onbekende woorden en namen.

- *Integratie van semantische, morfologische en syntactische modules.* De informatie die bij semantische *tagging* beschikbaar komt is hoofdzakelijk van lexicale aard. Een belangrijke aanvulling betreft de modellering van de betekenisopbouw in het geval van samengestelde woorden en zinnen. Hoe is de betekenis van een geleed woord opgebouwd uit de betekenissen van zijn samenstellende delen (integratie met morfologie) en hoe is de betekenis van de zin opgebouwd uit de betekenissen van de woorden en woordgroepen (integratie met de syntaxis) ? Om de bruikbaarheid van een compositioneel-semantische module te kunnen evalueren is het van belang dat hij - net als de parser - wordt geïntegreerd in een taal- of spraaktechnologische toepassing (laag 3).

## 2. Tekstvoorverwerking (in de ruime zin van het woord)

Tekstvoorverwerking betreft onder meer *tokenization*, detectie van zinsgrenzen, herkenning en classificatie van eigennamen (*named entities*), spellingcorrectie, behandeling van figuren en tabellen, extractie en behandeling van multi-woordeenheden, etc. Deze modules zijn ondermeer te gebruiken en te valideren bij de opbouw van het Corpus Geschreven Nederlands.

## 3. Morfologische analyse

Aangezien de samenstellingsregels in het Nederlands (bijna) even productief zijn als de regels voor de vorming van woordgroepen is de behandeling ervan van groot strategisch belang voor taal- en spraaktechnologie. Modules die automatisch samenstellingen kunnen opsporen, splitsen, ontleden en genereren kunnen o.m. grote diensten bewijzen bij het herkennen van nieuwe woorden, bij het verifiëren van de consistentie en accuraatheid van lexica en bij informatievergaring ("werkbij" is een soort bij, maar "voorbij" niet). Voor spraakherkenning is het ook van belang om te kunnen bepalen of twee opeenvolgende woorden al dan niet een samenstelling vormen.

## 4. Syntactische analyse: een robuuste parser voor het Nederlands

In het recente verleden zijn er enerzijds *tree banks* gemaakt (CGN, Alpino, etc.) en zijn er anderzijds parsers ontwikkeld met een behoorlijke *coverage* (b.v. Amazon, Alpino). De tijd is rijp voor een evaluatie van de bestaande parsers en voor de ontwikkeling van een plan dat moet uitmonden in de creatie van een robuuste, publiek beschikbare en *performante parser* voor het Nederlands. Dit houdt o.m. in dat criteria worden ontwikkeld voor het meten van de performantie. Speciale aandacht dient uit te gaan naar de interactie tussen de parser en de elektronische lexica. Vooral met betrekking tot de behandeling van multi-woordeenheden is dit cruciaal. Om de bruikbaarheid van de parser te kunnen evalueren is het van belang dat hij wordt geïntegreerd in een taal- of spraaktechnologische toepassing (laag 3)

### 5.3.5 Taal- en spraaktechnologische toepassingen (laag 3)

In dit onderdeel van STEVIN beoogt men vooral de uitwerking van toepassingen waarin TST-componenten worden aangewend, maar waarvan wordt aangenomen dat er nog onderzoek vereist is om tot een succesvolle realisatie te kunnen komen. Men verwacht hier in de eerste plaats de toepassingen die al in een vroeg stadium als potentieel interessant werden aangestipt (zie b.v. spraakcorpora op laag 1), maar ook toepassingen die later, in de loop van STEVIN, als het resultaat van vraagstimulering (laag 4) naar voren komen. De hieronder vermelde lijst van mogelijke onderwerpen die in dit luik aan bod zouden kunnen komen, is dus louter illustratief en niet beperkend.

- Extractie van betrouwbare informatie uit automatische transcripties van audio, gemaakt door een spraakherkenner. De uitvoer van een spraakherkenner verschilt van standaard tekst in die zin dat hij deels foutief is (indien slechts 1 oplossing werd gegeven), of meerdere mogelijkheden impliceert. Zijn de standaard IR (*Information Retrieval*) modellen die voor de extractie van informatie uit standaard tekst werden ontworpen, dan nog altijd optimaal?
- Detectie van accent en identiteit van de spreker. Het eerste aspect is b.v. belangrijk indien de toepassing zowel in Vlaanderen als in Nederland dient te werken. Het laatste aspect is belangrijk in alle toepassingen waarin men wil weten wie wat heeft gezegd (b.v. bij de automatische notulering van vergaderingen, bij IR uit transcripten van nieuwsberichten, etc.).
- Extractie van informatie uit monolinguale (IR) of multilinguale (CLIR) teksten: het opzoeken en extraheren van informatie over een bepaald domein kan nauwkeuriger en gericht gebeuren indien men morfologische en semantische analyses kan uitvoeren.
- Semantisch web: Dit moet toelaten dat ondermeer software agents de inhoud van web-pagina's kunnen begrijpen om hun taak te vervullen. Ontologie neemt hierbij een centrale plaats in.

- Dialogsystemen en Q&A oplossingen die gebruik maken van IR, CLIR of het Semantische web, in het bijzonder betrekking hebbend op multimodale data: tekst, spraak en visueel.
- Automatische samenvattingen en tekstgeneratie, waarbij cohesie beter kan worden gemoduleerd en meer natuurlijke expressies kunnen worden gegenereerd.
- Automatische vertaling: dit is een toepassing waarvoor alle bovenvermelde basis-taalvoorzieningen en modules relevant zijn; de integratie ervan in een werkend vertaalsysteem is en blijft de uitdaging bij uitstek in de taaltechnologie.
- Educatieve systemen, gebaseerd op multimodale interactie, voor een breed scala aan vakken, maar met een focus op aanvankelijk lees- en schrijfonderwijs. In dit kader kan ook gedacht worden aan TST-gebaseerde hulpmiddelen voor het onderwijzen en leren van het Nederlands als tweede taal.
- Educatieve systemen in de vorm van computerspelletjes.

De voorkeur zal hier uitgaan naar projecten met een duidelijke inbreng van TST-bedrijven en met als finaliteit de effectieve realisatie (in de vorm van een prototype of *demonstrator*) van een toepassing. Het is nochtans ook mogelijk om projecten in te dienen die niet tot een dergelijke finaliteit leiden, maar die een belangrijke stap voorwaarts in die richting kunnen betekenen. Verder strekt tot aanbeveling dat de projecten resultaten (data, tools en onderzoek) aanwenden die eerder in het kader van STEVIN werden verkregen, en dat ze samenwerking tussen TST bedrijven en taal- en spraaktechnologen uit kennisinstellingen inhouden.

#### 5.4 Vraagstimulering

Om aan vraagstimulering te kunnen doen dient de industriële kennisbehoefte geïdentificeerd en gebundeld te worden. Vervolgens kunnen op basis daarvan stimuleringsacties worden bepaald. Uiteraard heeft de overheid hier zijn beperkingen en kan het alleen de eigen vraag van overheidspartijen beïnvloeden. Vooral dient ervoor gewaakt te worden dat de overheid bij stimulering geen ongewenste verstoring van het marktmechanisme veroorzaakt. In ieder geval is de beoogde basistaalvoorziening een essentieel ingrediënt voor een volwassen voorwaardenscheppend beleid van de zijde van de overheid. Middels de basistaalvoorziening moeten de bedrijven in staat worden gesteld voldoende slagkracht te ontwikkelen op de snel internationaliserende TST-markt.

De Programmacommissie heeft reeds in de voorbereidende fase een aantal vruchtbare discussies over dit onderwerp gevoerd waarin een aantal mogelijke activiteiten aan de orde zijn gekomen:

- **Demonstrators:** TST is *enabling* Het is zelf strikt genomen geen toepassing, maar stelt andere applicaties en/of diensten in staat iets te doen, b.v. om te gaan met meer natuurlijk taalgebruik op basis waarvan de gebruikersvriendelijkheid vergroot kan worden. TST moet beter zichtbaar gemaakt worden voor de ontwikkelaars van applicaties en diensten en vervolgens bij de beslissers over toekomst van zulke applicaties. Dit kan gerealiseerd worden door het maken en demonstreren van voorbeeldapplicaties voor specifieke doelgroepen.
- **Gebruikersvoorlichting:** Op informatiedagen voor specifieke doelgroepen kan zowel de vraagarticulatie nader uitgewerkt worden als ook voorlichting gegeven worden over TST-mogelijkheden en onmogelijkheden.
- **Markteducatie:** Naarmate de toepassingen van TST duidelijker en bekender worden, ontstaat bij bedrijven de behoefte aan mensen (maar niet noodzakelijk specialisten) die kennis van zaken hebben. Gedacht wordt aan het opzetten van "*Master Classes TST*" b.v. over metadata tagging. Het gaat hierbij nadrukkelijk niet om een extra wetenschappelijke opleiding, maar om een cursus van een aantal dagen/avonden waarbij deelnemers een uitvoerig overzicht wordt geboden van bestaande en "komende" TST technieken en manieren waarop deze toegepast zou kunnen worden.
- **Launching customers:** De overheid treed zelf op als voorbeeldklant in projecten die tot de verbeelding spreken. Gedacht kan worden aan gemeentes, omroepverenigingen, politie, zorginstellingen, etc. Deze instellingen hebben met elkaar gemeen dat ze veel (geschreven en gesproken) data hebben die meestal nog niet goed (genoeg) ontsloten zijn. Eenmaal gerealiseerde projecten kunnen, gegeven het publieke karakter van de instellingen, dikwijls rekenen op enige vorm van publiciteit.

Een aantal van deze activiteiten wordt in hoofdstuk 6 al enigszins nader uitgewerkt. Een aantal commissieleden is al enthousiast bezig om met name de plannen met betrekking tot markteducatie

nader uit te werken. Het stimuleren van de ontwikkeling van demonstrators zal vanaf de start van STEVIN actief door de Programmacommissie worden ondersteund. In mei 2005 zal de Programmacommissie nog een nader uitgewerkt activiteitenplan ten aanzien van andere activiteiten rond vraagstimulering die zij wensen te ondersteunen, ter goedkeuring voorleggen aan het Programmabestuur. De Programmacommissie zal voor die tijd nader overleg voeren met vertegenwoordigers uit het bedrijfsleven. Ook zal gekeken worden of het Innovatieplatform in dit kader een rol zal kunnen spelen. Verder zal bekeken worden welke lering getrokken kan worden uit het advies van de commissie-Riseeuw op basis waarvan in de ICT-sector een *corporate academy* is opgezet. Via die *academy* krijgen de ICT-professionals toegang tot een competentiegericht opleidingsaanbod en scholingsmogelijkheden binnen hun vakgebied.

## 5.5 Randvoorwaarden: IPR en standaarden

IPR-beleid is een onvoorwaardelijk onderdeel van STEVIN. Immers, het is de bedoeling dat de basis taalvoorzieningen voor het Nederlands, op de eerste plaats alle resultaten van het programma, op niet-discriminatieve wijze beschikbaar te stellen aan alle belanghebbende partijen. Het is op zich al een uitdaging om een voor alle partijen bevredigend IPR beleid te formuleren en te implementeren voor nieuwe voorzieningen die met steun van publieke middelen ontwikkeld worden. In het onderhavige programma wordt het IPR-beleid nog gecompliceerd door het feit dat het niet alleen gaat om de intellectuele rechten op nieuw te creëren basisbestanden, maar ook op bestanden die in het verleden zijn gecreëerd met (mede)financiering van de nationale of Europese overheden en waarvan achteraf is gebleken dat de rechten onvoldoende duidelijk zijn geregeld.

Het uitgangspunt voor het IPR-beleid in STEVIN is dat alle basisvoorzieningen - nieuw en bestaand - beheerd en onderhouden zullen worden door de TST-centrale van de Nederlandse Taalunie. "Beheer en onderhoud" omvat het toegankelijk maken van de voorzieningen, en het beschermen van de aan die voorzieningen verbonden rechten. Het ligt daarom voor de hand om de IPR regels voor het onderhavige programma te baseren op de regels en richtlijnen die momenteel door de TST-centrale ontwikkeld worden. Die regels en richtlijnen zijn mede gebaseerd op de ervaringen die opgedaan zijn in het kader van het zopas voltooide Vlaams-Nederlandse project *Corpus Gesproken Nederlands*. In dat verband is nauw samengewerkt met de European Language Resource Association (ELRA) en het Amerikaanse Linguistic Data Consortium (LDC). Zowel ELRA als LDC - die overigens ook nauw samenwerken met elkaar - hebben IPR-standaarden ontwikkeld die de combinatie van bestaande en nieuw te ontwikkelen voorzieningen mogelijk maken en die brede steun vinden bij alle betrokken partijen, overheden, onderzoeksinstellingen en bedrijven.

Om de regeling van IPR-zaken in dit programma zo eenvoudig en doorzichtig mogelijk te maken, wordt de eis gesteld dat slechts aan de uitvoering van een project begonnen kan worden als bij dat project betrokken partijen contractueel vastgelegd hebben dat, en op welke wijze, de resultaten van het project toegankelijk gemaakt zullen worden voor belanghebbenden. De hier bedoelde contracten zullen afgeleid worden van de regels en richtlijnen die door de TST-centrale ontwikkeld worden. Als een project voortbouwt op bestaande voorzieningen waarop commerciële partijen rechten kunnen claimen, moet voor de start van het project contractueel vastgelegd zijn dat die bestaande voorzieningen tegen redelijke voorwaarden beschikbaar gesteld zullen worden, analoog aan de manier waarop *pre-existing knowledge* behandeld wordt in de IPR-regelingen voor projecten in het 6de Kaderprogramma van de EU (zie de *best practice guide* ([www.cordis.lu/fp6/find-doc.htm#ipr](http://www.cordis.lu/fp6/find-doc.htm#ipr))).

De herbruikbaarheid van sommige basistaalvoorzieningen die in het verleden ontwikkeld zijn werd sterk beperkt doordat idiosyncratische formaten en datastructuren gebruikt werden. Gelukkig zijn er binnen de taal- en spraaktechnologie in de afgelopen jaren aanzetten gegeven tot de ontwikkeling van standaarden die hergebruik vergemakkelijken. Een deel daarvan is ontwikkeld binnen Europese samenwerkingsprogramma's zoals EAGLES en ISLE. Andere zeer belangrijke organen die zich bezighouden met normalisatie zijn, o.a. ISO/T37, W3C alsook de Localising Industry Standards Association (LISA). Voor de Nederlandse TST-industrie, met zijn relatief kleine thuismarkt, is het van meer dan gemiddeld belang dat internationale standaarden tot stand komen en worden ondersteund door de industrie. De Programmacommissie zal dan ook met nadruk eisen stellen aan de toepassing van bestaande standaarden en het meewerken aan de ontwikkeling van nieuwe standaarden.

## 5.6 Relatie met andere programma's

In sectie 2.4 is reeds een opsomming gegeven van direct gerelateerde programma's. De Programmacommissie zal zorgdragen voor afstemming met de activiteiten in die programma's. Om dat zo goed mogelijk te bewerkstelligen is ervoor gezorgd dat in de organisatie van dit TST-

meerjarenprogramma vertegenwoordigers uit het veld die relaties hebben met gerelateerde programma's zijn opgenomen (zie sectie 7.2). De Programmacommissie zal er op toezien dat waar mogelijk samenwerking opgezet wordt voor wat betreft kennisoverdracht, netwerkvorming en verankering (zie hoofdstuk 6).

Daarnaast is er ook een aantal aanverwante projecten die uitgevoerd zullen gaan worden in het kader van de Bsik-programma's MultimediaN en ICIS. In MultimediaN wordt aandacht besteed aan *Information Extraction* uit multimedia documenten zoals registraties van TV Journaals en andere omroep producties. Bij de ontwikkeling van *Broadcast News Corpora* in het onderhavige programma zal de Programmacommissie er op toezien dat naar maximale synergie gerealiseerd wordt met de activiteiten in MultimediaN. In het Bsik programma ICIS wordt onderzoek gedaan naar multimodale mens-systeem interactie en dialoog management. Ook hier zal de Programmacommissie zorgdragen voor een goede afstemming.

## 5.7 Draagvlak: kennisinstellingen en industrie

In de Technologieverkenning uitgevoerd door M&I/Partners in samenwerking met Montemore wordt geconcludeerd dat er veel draagvlak is voor een gezamenlijk Vlaams Nederlands programma. Er vindt een continue discussie plaats tussen marktpartijen en publieke kennisinstellingen over de juiste prioriteiten. Het goed regelen van de intellectuele eigendomsrechten op data, tools en / of modules is cruciaal om een goede samenwerking en een stevig draagvlak te behouden.

## 5.8 Beoordelingscriteria

Een projectvoorstel zal aan de in de oproep geformuleerde aanvraaginstructies moeten voldoen en dient met name in het Engels opgesteld te zijn om toetsing door onafhankelijke internationale experts mogelijk te maken. De beoordelingscriteria zijn als volgt.

### Kwaliteit en innovatiegehalte van het projectvoorstel:

- Helderheid van de probleemstelling en oorspronkelijkheid van het project.
- Geschiktheid van de methode en -opzet. Hierbij noemen we met name dat het project een zo groot mogelijke impact moet hebben op een zo breed mogelijke laag van toepassingen of moet starten vanuit een concrete, voor het bedrijfsleven relevante toepassing. Daarnaast dient het voorstel een expliciete component van evaluatie te bevatten, en dient voor de aanmaak van BaTaVo-data een validatieplan onderdeel te zijn van het project.
- Competentie deelnemende (onderzoeks)groepen (inclusief *past performance*).
- Haalbaarheid van de doelstellingen: zijn de doelstellingen realiseerbaar binnen de gevraagde tijd en met de gevraagde middelen. Dit houdt niet in dat met name onderzoeksvoorstellen niet risicovol mogen zijn (onderzoeksvoorstellen op het gebied van strategisch onderzoek zullen dat zeer waarschijnlijk juist zijn), maar eerder dat voor alle projecten, inclusief risicovolle, de te investeren middelen adequaat moeten zijn voor de beoogde doelstellingen zelfs al is er voor risicovolle projecten geen garantie op succes in de huidige stand van zaken van technologie en onderzoek.
- De mate van integratie van de verschillende lagen binnen het projectvoorstel. Welke lagen in de ketenbenadering spelen een rol en hoe zijn zij geïntegreerd?
- Evenwichtigheid van de samenwerking en taakverdeling binnen het project.
- Beschikbaarheid van benodigde infrastructuur.

### Economische aspecten van het projectvoorstel:

- Samenwerking met of steun van bedrijven;
- Zicht op spin-offs en/of andere nieuwe ontwikkelingen;
- Toepassingskansen van de resultaten in de industrie en/of in de maatschappij.

### Bijdrage aan het programma:

- Conformiteit aan de focus van het programma en aansluiting bij de daarin gestelde prioriteiten. Het project dient zich daarnaast te richten op het Nederlands, en een bijdrage te leveren aan de verbetering of tenminste het behoud van de positie van het Nederlands in de moderne informatie- en communicatiemaatschappij.

- Perspectief op kennisoverdracht en netwerkvorming. Met name noemen we hierbij dat het in het voordeel spreekt van een voorstel wanneer de expertise van Nederlandse en Vlaamse groepen of bedrijven gecombineerd wordt in een projectvoorstel, wanneer bedrijfsleven en kennisinstellingen samen een voorstel indienen, of wanneer een voorstel betrekking heeft op zowel spraak- als taaltechnologie.

#### IPR en voorkoming duplicatie:

- Het voorstel dient een duidelijk plan te bevatten voor de adequate afhandeling van intellectuele eigendomsrechten, zowel voor de door derden in te brengen basistaalvoorzieningen als voor de resultaten van het project. Uitgangspunt hierbij is dat de uitkomsten van het project op een niet-discriminatieve wijze toegankelijk gemaakt worden via de TST-centrale.
- Het voorstel moet aantonen dat de aanvragers een nauwkeurig en up-to-date beeld hebben van wat reeds beschikbaar is aan basistaalvoorzieningen. Bij voorkeur bestaan de te ontwikkelen basisvoorzieningen nog niet. Indien bekend is of redelijkerwijs aangenomen mag worden dat de basisvoorzieningen bestaan maar niet beschikbaar zijn, dient het voorstel een plan te bevatten om verstoring van de markt te voorkomen.
- De aanvragers dienen zich in het onderzoeksvoorstel bereid te verklaren de resulterende basistaalvoorzieningen ter beschikking te stellen aan de TST-centrale, hier onderhandelingen over te voeren met de TST-centrale, en te schetsen onder welke voorwaarden dit kan geschieden. Afsluiting van een contract over de IPR-regeling voor de aanvang van het project is een noodzakelijke voorwaarde voor honorering.
- Het voorstel dient aan te sluiten bij bestaande standaarden en die waar mogelijk toe te passen, of mee te werken aan de ontwikkeling van nieuwe standaarden, zodat een maximaal hergebruik van de ontwikkelde basistaalvoorzieningen gegarandeerd wordt.

## **5.9 Aanvraagprocedure**

STEVIN zal op zo'n manier opgezet worden dat maximale flexibiliteit in de financiering gewaarborgd wordt. Aangezien de prioriteiten zowel betrekking hebben op omvangrijke en geïntegreerde projecten als op kleinere types van projecten, zal de maximale omvang en het type van de projecten afhankelijk worden gemaakt van de inhoudelijke en organisatorische eisen van de prioriteit waar ze betrekking op hebben. Wat omvang betreft zullen zowel grote programma's (bijv. tot 12 fte voor 4 jaar) als kleine programma's (bijv. tot 6 fte voor 4 jaar) en individuele projecten (1 fte tot 4 jaar) mogelijk zijn. De oproep tot projectvoorstellen kan zowel aanbod-gedreven ("call for proposals", open invulling van gevraagde prioriteiten) als vraag-gedreven ("call for tender", specifieke onderzoeks- of ontwikkelingsvraag) zijn. Aan iedere oproep zal een *brokerage* worden voorafgegaan om één-op-één ideeënvorming en samenwerking te stimuleren van onderzoeksinstituten en bedrijven.

Voor de eerste oproep zullen voorstellen kunnen worden ingediend met een beperkte omvang. Ook kunnen voorstellen ten behoeve van de financiering van voorbereidende trajecten voor grotere consortiumvoorstellen worden ingediend. Het budget voor deze oproep is maximaal 2.000.000 Euro. De eerste oproep zal in september 2004 worden gepubliceerd en volledige voorstellen dienen medio november 2004 te worden ingediend. Deze zullen door een internationaal samengestelde expertcommissie worden beoordeeld. In december 2004 zal vervolgens de Programmacommissie de prioritering opstellen op basis waarvan het Programmabestuur nog voor het eind van het jaar een besluit zal nemen over honorering.

In de tweede oproep kunnen ook grotere consortiumvoorstellen worden ingediend. Deze oproep zal in het voorjaar van 2005 worden gepubliceerd en een omvang hebben van circa 4.200.000 Euro. Deze oproep zal in principe in twee fases worden opgedeeld: 1) uitnodiging tot indiening van vooraanmeldingen, één tot twee A4-tjes met daarin de essentie van het beoogde project. Uit de ingediende vooraanmeldingen selecteert de Programmacommissie de indieners die worden uitgenodigd een volledig voorstel uit te werken. 2) indiening van volledige voorstellen, externe beoordeling, prioritering door Programmacommissie, honoreringsbesluit door Programmabestuur.

Over de invulling van de 3e oproep zal in de loop van 2006 door Programmacommissie besloten worden op basis van de uitkomsten van de 1e en 2e oproep.

Algemene aanvraaginstructies en criteria voor de evaluatie van projectvoorstellen zullen samen met de oproepen tot het indienen van voorstellen tijdig bekend gemaakt worden. De Nederlandse Taalunie is samen met SenterNovem en NWO reeds begonnen met de uitwerking van de

## **5.10 Evaluatie (nulmeting)**

Om de voortgang en de impact van STEVIN te kunnen meten is het van belang het programma en de verschillende projecten die daarbinnen gefinancierd worden te monitoren en evalueren. Daarvoor zullen bij aanvang door de Programmacommissie succescriteria vastgesteld worden. Op basis van deze criteria zal bij de start van het programma een nulmeting uitgevoerd worden en dezelfde criteria zullen worden ingezet bij de voortgangscontroles en eindevaluatie.

Op het niveau van concrete projecten dient evaluatie een integraal onderdeel te vormen van het project, en de Programmacommissie zal de individuele projecten of kleine programma's specifiek op dit criterium beoordelen. Indien nog niet beschikbaar kan het ontwikkelen van *benchmarks*, *test-suites* and testdata opgenomen te worden als onderdeel van het projectvoorstel. De resulterende *test-suites*, *workbenches*, etc., dienen publiekelijk beschikbaar gemaakt worden via de TST-centrale.

Op het programmaniveau zal geëvalueerd worden wat de impact van STEVIN is op de taal- en spraaktechnologie in Nederland en Vlaanderen. Om dit te kunnen doen zal de Programmacommissie door een *trusted 3rd party* een initiële meting (nulmeting) aan het begin van het programma laten uitvoeren op vooraf geformuleerde succescriteria, en dezelfde meting tegen het eind van het programma herhalen. Voor het opstellen van de succes criteria zal uitgegaan worden van het *format* dat reeds is opgesteld door SenterNovem en NWO ten behoeve van de BSIK<sup>25</sup>-projecten. Voor de invulling van de nulmeting zal deels gebruik gemaakt worden van de cijfers die verzameld zijn in het kader van de technologieverkenning uitgevoerd door M&I/Partners en de EUROMAP *benchmark study*. Een aantal van deze cijfers is vooral kwalitatief beschreven. Een aantal aspecten van deze meting dient nog meer kwantitatief gerepresenteerd te worden om vergelijkingen met een meting aan het einde van STEVIN te vergemakkelijken (hoewel natuurlijk gewaakt moet worden voor het gevaar van "oversimplificatie" en de vertekende weergave die dit met zich mee kan brengen). Daarnaast zal de Programmacommissie laten inventariseren tegen welke voorwaarden het Vlaamse en Nederlandse bedrijfsleven resources die hun eigendom zijn - eventueel na bewerking of anonimisering - ter beschikking wil stellen.

In ieder geval zullen in de nulmeting de volgende elementen aan de orde komen: 1) criteria voor wetenschappelijke output, b) criteria voor economische output, 3) criteria voor maatschappelijke output, 4) criteria innovatietraject, incl. kennistransfer, netwerkvorming en verankering.

De Programmacommissie zal dezelfde meting (laten) uitvoeren tegen het eind van het programma, en een vergelijking met de nulmeting maken om de impact van STEVIN vast te stellen. Natuurlijk kunnen er factoren opduiken die geheel onafhankelijk zijn van het hier voorgestelde programma en buiten de controle van de Programmacommissie die een grote invloed kunnen hebben op de ontwikkeling van taal- en spraaktechnologie in Nederland en Vlaanderen (algemene economische ontwikkeling, beslissingen van specifieke (grote) bedrijven om hun activiteiten te intensiveren of juist af te bouwen, etc.). Deze kunnen de resultaten van de meting vertekenen, maar waar dat het geval is dient dat zo goed mogelijk beschreven en als relevante factor gerechtvaardigd te worden in de finale meting.

De Programmacommissie zal in november 2004 een meer gedetailleerd voorstel presenteren ten aanzien van deze meting. Hierbij zal ook gebruik gemaakt worden van de ervaring en resultaten van dergelijke metingen zoals die zijn uitgevoerd door het KP5 EUROMAP-consortium. Eind 2004 zal de uitvoering van de meting uitbesteed worden aan de reeds vermelde *trusted 3rd party*, zodat de nodige onafhankelijkheid gegarandeerd wordt. Het gedetailleerde voorstel dient een optimum te zijn tussen wat wenselijk is en wat financieel haalbaar is gegeven het budget dat uitgetrokken is voor deze activiteiten.

---

<sup>25</sup> Bsik: Besluit subsidies investering kennisinfrastructuur (voorheen ICES/KIS)

## 6. Kennisoverdracht, netwerkvorming en verankering

### 6.1 Inleiding

Een van de belangrijke doelstellingen van STEVIN is - naast kennisontwikkeling - het versterken van de samenwerking tussen de kennisinfrastructuur en het bedrijfsleven in zowel Nederland als Vlaanderen. Daarbij dient onderscheid gemaakt te worden tussen project specifieke kennisoverdracht en kennisoverdracht op programmaniveau. De Programmacommissie zal specifieke activiteiten ontwikkelen met betrekking tot a) netwerkvorming en kennisuitwisseling, b) kennisoverdracht en c) zwaartepuntvorming en verankering. In mei 2005 zal de Programmacommissie een specifieke uitwerking en planning van deze activiteiten ter goedkeuring voorleggen aan het Programmabestuur.

Waar mogelijk zal de Programmacommissie er voor zorgen dat de activiteiten samen met andere partijen die eveneens dit soort activiteiten uitvoeren (bijvoorbeeld in gerelateerde programma's) om een gecoördineerde inspanning op dit gebied te realiseren. Dit geldt ook en vooral voor de activiteiten die de Nederlandse Taalunie in het kader van de in haar beleidsplan opgenomen 'makel- en schakelactiviteiten' ten aanzien van de verbetering van de positie van het Nederlands in taal- en spraaktechnologie reeds van plan is uit te voeren.

Een aantal van deze activiteiten heeft een directe relatie met de eveneens via STEVIN beoogde vraagstimulering (zie sectie 5.3.3). Bij de invulling van deze activiteiten moet rekening gehouden worden met het feit dat voor TST verschillende groepen gebruikers geïdentificeerd kunnen worden: a) Academische gebruikers, b) Producenten van taal- en spraaktechnologie, c) Integrators en d) Eindgebruikers uit diverse marktsegmenten. Al deze gebruikersgroepen hebben hun eigen behoeftes en activiteiten dienen dus ook specifiek op de betreffende groepen toegesneden te worden. Er zal een Begeleidingscommissie geformeerd worden waarin vertegenwoordigers van de verschillende typen worden opgenomen (zie sectie 7.7).

Het is noodzakelijk per gebruikerscategorie de behoeften nader te definiëren, waarbij specifieke acties op verschillende doelgroepen gericht kunnen zijn.

- Product: elke doelgroep heeft behoefte aan zijn eigen producten (data , tools en corpora) en diensten (consultancy en validatie).
- Prijs: hier komen de volgende vragen aan de orde: wel of niet *open source*; mag/moet winst gemaakt worden; welke kostprijs wordt berekend.
- Promotie: nationaal of internationaal (workshops, seminars, beurzen, congressen, *roadmap* commissies, markteducatie/master classes, vakbladen, kranten.
- Plaats: welke kanalen kunnen we benutten: via bijeenkomsten, publicaties, websites
- People: wie is het best in staat om bepaalde doelgroepen te bereiken.: Taalunie, SenterNovem, IWT-Vlaanderen, NWO, MVG-AWI, NOTaS, CLIF, FENIT, groepen bedrijven voor een bepaald marktsegment, individuele bedrijven.

In de volgende secties wordt een aantal mogelijke activiteiten opgesomd en nader uitgewerkt.

### 6.2 Geplande activiteiten

Doelstelling	Activiteit	Doel / Resultaat	Doelgroep
Netwerk- vorming	Distributie interactieve "wie-is-wie"-databank i.s.m. Taalunie e.a. gerelateerde netwerken	Creëren van een virtuele TST "landkaart"	Alle belangstellenden in TSTonderzoek/ontwikkelingen
	Organisatie jaarlijkse TST-dag, i.s.m. Taalunie e.a. gerelateerde netwerken	Bijeenbrengen TST wetenschappers en bedrijfsleven.	Bedrijven en kennisinstellingen
	Organisatie TST brokerages	Tot stand brengen nieuwe samenwerkingsverbanden tussen kennisinstellingen en bedrijfsleven	Bedrijven, kennisinstellingen

Doelstelling	Activiteit	Doel / Resultaat	Doelgroep
	Internationalisering, o.a. opzetten ERA-NET LANGNET, i.s.m. Taalunie e.a. gerelateerde netwerken	Internationalisering en standaardisering TST	Buitenlandse bedrijven en kennisinfrastructuur
<b>Kennis-overdracht</b>	Internetsite	Algemene en project-specifieke informatie	Bedrijven en kennisinstellingen
	Artikelen in (wetenschappelijke) tijdschriften.	Publiek maken van resultaten van het onderzoek	Alle TST-experts in kennisinstellingen en bedrijven
	Gebruikersbijeenkomsten	Kennisuitwisseling en voortgangstoetsing	Bedrijven en kennisinstellingen betrokken bij het programma
	Participatie aan beurzen en congressen, bv TST stand op IST congres 2004 en evt. organisatie LANGTECH	Bekendheid TST	Bedrijven en kennisinstellingen
	Tijdelijke plaatsing van onderzoekers vanuit bedrijven in kennisinstellingen	Directe en praktische kennisuitwisseling	Bedrijven en kennisinstellingen
	Stages van universitaire onderzoekers binnen bedrijven	Directe en praktische kennisuitwisseling	Bedrijven en kennisinstellingen
	AIO trainingen presentatie-vaardigheden, octrooien, ondernemerschap	Opleiden TST-specialisten	Bedrijven en kennisinstellingen
	Uitdragen TST-'boodschap' mbv aantrekkelijke demonstrators	Potentiële toepassers overtuigen van mogelijkheden en nut van TST	Bedrijven
	Markteducatie, b.v. master classes over metadata tagging	Gezamenlijke actie om bedrijven beter op de hoogte te brengen van TST-mogelijkheden	Bedrijven
	Persplan, i.s.m. Taalunie e.a. gerelateerde netwerken	Vergroting bekendheid bij media	Landelijke pers/vakbladen
<b>Verankering</b>	Overleg en samenwerking met IMIX, MMI en andere gerelateerde programma's	Afstemming en kruisbestuiving met andere strategische onderzoeksprogramma's	Afvaardiging in commissies gerelateerde programma's
	Relatie zoeken met TST-expertise bij Syntens en Technostarters en de Vlaamse tegenhanger daarvan	Relatie met MKB verstevigen	MKB-starters
	Relaties met TST Platform	Beleidsafstemming	Beleid

In mei 2005 zal de Programmacommissie met een specifieke uitwerking en planning gereed hebben van bovengenoemde activiteiten. Er zal gebruik gemaakt worden van het handboek dat binnen SenterNovem ontwikkeld is<sup>26</sup>.

<sup>26</sup> Handboek beschikbaar op <http://www.senter.nl/asp/page.asp?id=i001291&alias=iop>

### 6.3 Netwerkvorming en kennisuitwisseling

De bevordering van het ontstaan van netwerken tussen instellingen en bedrijfsleven en het daarbij tot stand brengen van aansluiting bij internationale programma's en netwerken is een van de doelstellingen van STEVIN. Netwerkvorming is een belangrijke basis om kennis uit te kunnen wisselen en de benutting van de ontwikkelde kennis te stimuleren. In alle gevallen dient afstemming plaats te vinden met de activiteiten van de Nederlandse Taalunie die in het kader van zijn eigen beleidsplannen reeds activiteiten op dit gebied initieert (makel- en schakelfunctie (zie hoofdstuk 2) zie o.a. <http://www.taaluniversum.org/technologie>). Ook dient samengewerkt te worden met andere lopende gerelateerde programma's (zie sectie 2.4). De Programmacommissie zal specifieke acties ondernemen om kennisuitwisseling binnen het programma te stimuleren (zie 6.4). Voor het tot stand brengen van nieuwe samenwerkingsverbanden tussen kennisinstellingen en bedrijven zullen de door SenterNovem in het kader van de IOP-programma's beproefde *brokerages* worden ingezet. Op deze bijeenkomsten zal *commitment* van bedrijven aan projecten en inzicht van universiteiten in de behoeften van het bedrijfsleven, idealiter resulterend in gezamenlijke projectvoorstellen, gestimuleerd worden.

Uit de M&I technologieverkenning is gebleken dat er qua netwerkrelaties al een goede basis aanwezig is, vooral in Vlaanderen, waar op voortgebouwd kan worden. Ook kan worden voortgebouwd op de Vlaams-Nederlandse relaties die tot stand gekomen zijn in het project *Corpus Gesproken Nederlands*. In Vlaanderen participeren alle kennisinstellingen in CLIF (Computational Linguistics in Flanders). Veel van de Nederlandse kennisinstellingen participeren in NOTaS. Deze stichting is opgericht door een aantal MKB's uit de TST-sector en heeft tot taak de belangen te behartigen van bedrijven en kennisinstellingen die actief zijn op het terrein van taal- en spraaktechnologie en de sector een eigen gezicht te geven in binnen- en buitenland. De stichting NOTaS is bezig te bezien of het mogelijk is het netwerk van momenteel 20 bedrijven en kennisinstellingen in Nederland uit te breiden naar Vlaanderen. Vele Vlaamse en Nederlandse kennisinstellingen en bedrijven zijn betrokken in internationale netwerken (o.a. ELSNET, ELRA/ELDA, ISCA etc). In de technologieverkenning werd een eerste inventarisatie gemaakt van deelname van bedrijven en kennisinstellingen aan verschillende netwerken.

In sectie 6.2 wordt reeds een aantal acties genoemd die de Programmacommissie zal ondersteunen ter bevordering van de netwerkvorming. Nadere uitwerking en planning van de specifieke activiteiten zal de Programmacommissie formuleren in het werkplan dat in mei 2005 ter goedkeuring wordt voorgelegd aan het Programmabestuur. Via de geplande evaluaties kan gevolgd worden hoe deze situatie zich tijdens de uitvoering van STEVIN ontwikkelt. In ieder geval zal de Programmacommissie bij de aankondiging van de eerste subsidieronde een *kick-off meeting* met *brokerage* organiseren.

### 6.4 Kennisoverdracht

Onderscheid moet gemaakt worden tussen projectspecifieke kennisoverdracht en kennisoverdracht op programmaniveau (flankerende activiteiten). Kennisoverdracht dient een wezenlijk onderdeel uit te maken van zowel het gehele programma als van elk van de deelprojecten. Zoals eerder opgemerkt zal ook op dit punt actief samengewerkt worden met andere gerelateerde programma's alsook met de activiteiten die de Nederlandse Taalunie uitvoert in het kader van de 'makel- en schakelfunctie' (zie hoofdstuk 2).

Binnen de Taalunie-website is een speciaal deel ingericht specifiek over taal- en spraaktechnologie (<http://taalunieversum.org/taal/technologie/>); deze TST-infodesk zou verder moeten worden uitgebreid met *best practices* en andere verwijzingen naar o.a. standaarden en validatiemethodes.

Voor de interne kennisoverdracht zal de Programmacommissie op regelmatige basis bijeenkomsten van projectmedewerkers en speciale gebruikersbijeenkomsten organiseren. In sectie 6.2 wordt een aantal andere mogelijke activiteiten genoemd, waaronder het jaarlijks inrichten van de TST-stand op het Nederlandse ICT-Kenniscongres en het verkennen van de mogelijkheid om een internationale congres op het gebied van taal- en spraaktechnologie, b.v. LANGTECH (<http://www.lang-tech.org/>) of Interspeech/Eurospeech, naar Nederland of Vlaanderen te halen.

Voor het faciliteren van de kennisoverdracht tussen bedrijfsleven en kennisinfrastructuur zullen stages van onderzoekers binnen bedrijven en ook tijdelijke plaatsing van bedrijfsonderzoekers binnen kennisinstellingen actief gestimuleerd worden. Met name moet het programma ook bevorderen dat een nieuwe generatie TST-ers, die hard nodig is in het bedrijfsleven, opgeleid wordt. Een belangrijke taak die dit programma kan vervullen is het uitdragen en verduidelijken van de van de bijdrage die taal- en spraaktechnologie kan leveren aan ICT-applicaties en -diensten.

Aantrekkelijke demonstrators voor specifieke doelgroepen zullen ontwikkeld worden om de zichtbaarheid van de technologie en haar toepassingsmogelijkheden te ondersteunen. Deze kunnen gedemonstreerd worden op het jaarlijkse ICT-Kenniscongres. Een andere mogelijkheid om het bedrijfsleven duidelijk te maken wat er precies mogelijk is met behulp van taal- en spraaktechnologie (markteducatie) en hoe ze dat kunnen realiseren is het organiseren van master classes voor bedrijfsgroepen over specifieke TST-onderwerpen.

Een speciale Werkgroep Flankerend Beleid (zie sectie 7.6) zal worden ingesteld en deze zal in mei 2005 een gespecificeerd werkplan met tijdschema opleveren. Ook hier geldt dat via de geplande evaluaties gevolgd zal worden hoe deze activiteiten zich tijdens dit programma ontwikkelen. De Programmacommissie zal - vooruitlopend op de nadere uitwerking van de plannen - er voor zorgdragen dat samen met andere betrokken partners een TST-stand ingericht wordt op het IST Event 2004 dat in november in Den Haag georganiseerd wordt.

## 6.5 Zwaartepuntvorming en verankering

Door zwaartepuntvorming, taakverdeling en samenwerking tussen de onderzoeksgroepen is het mogelijk de kennisinfrastructuur te versterken. In de technologieverkenning van M&I/Partners staat een overzicht van de taal- en spraaktechnologische terreinen waarop de verschillende kennisinstellingen zich momenteel richten. Door bij de keuze van projecten de samenwerking tussen de onderzoeksgroepen te bevorderen zullen zwaartepuntvorming en taakverdeling de kennisinfrastructuur versterken. Ook kunnen op die manier de beschikbare middelen effectief worden ingezet. Via de geplande evaluaties zal gevolgd worden hoe de zwaartepuntvorming zich tijdens het programma ontwikkelt.

Zoals de meeste onderzoeksprogramma's heeft ook dit TST-meerjarenprogramma een tijdelijk karakter. Het is dus van groot belang te streven naar duurzaamheid van het gerealiseerde ofwel verankering van verworvenheden, zoals gebruik van resultaten, continuering en verdere uitbreiding van netwerken, en voortzetting van de kennisontwikkeling. Het realiseren van een centrale mogelijkheid voor beheer van de resultaten (zie sectie 6.6) vormt hiervoor een zeer goede basis, waar de Programmacommissie actief aan wil bijdragen. Ook de centrale activiteiten van de Nederlandse Taalunie in het kader van de make- en schakelfunctie op het gebied van Nederlandse taal- en spraaktechnologie zullen een positieve bijdrage leveren. Verder zal aan onderzoekers die projectvoorstellen indienen gevraagd worden in te gaan op de inbedding en verankering van hun project.

Naast dit TST-meerjarenprogramma is er een aantal gerelateerde lopende onderzoeksprogramma's (zie sectie 2.4). Afstemming van activiteiten - als ook met internationale onderzoeksprogramma's - is van groot belang. Om dat zo goed mogelijk te bewerkstelligen is ervoor gezorgd dat in de organisatie van dit programma vertegenwoordigers uit het veld met directe relaties met andere gerelateerde programma's zijn opgenomen (zie sectie 7.2). Afstemming op beleidsniveau vindt plaats via het Programmabestuur. Hoewel er geen specifieke actie gepland is om TST-technostarters te stimuleren of te ondersteunen, zal via de website wel informatie gegeven worden over de bestaande overheidsorganisaties die zich hier reeds op richten.

## 6.6. Onderhoud, beheer en exploitatie van resultaten

Tussen 1999 en 2002 werd door het Instituut voor Nederlandse Lexicologie (INL) in opdracht van het TST-platform reeds een uitvoerige blauwdruk <sup>27</sup> uitgewerkt voor een doeltreffend beheer, onderhoud, distributie en terbeschikkingstelling van met overheidsmiddelen ontwikkelde digitale materialen voor het Nederlands. De beleidsaanbevelingen (hoofdstuk 9 van de blauwdruk) worden als *bijlage 1* aan onderhavige nota toegevoegd. In deze aanbevelingen wordt de noodzaak beklemtoond om met overheidsmiddelen ontwikkelde Nederlandstalige digitale materialen door een centraal orgaan (met rechtspersoonlijkheid) te laten beheren en onderhouden en uiteraard ook beschikbaar te laten stellen - daarbij rekening houdend met alle mogelijke randvoorwaarden van bijvoorbeeld economische of juridische aard - voor onderwijs- en onderzoeksdoeleinden en voor de ontwikkeling van taal- en spraaktechnologische toepassingen.

De Nederlandse Taalunie heeft beheer, onderhoud, distributie en beschikbaarstelling van Nederlandstalige digitale materialen als een specifieke doelstelling opgenomen in haar

---

<sup>27</sup> De volledige *Blauwdruk voor onderhoud, beheer en distributie van door de overheid gefinancierde digitale materialen* (INL, m.m.v. SPEX) kan onder meer worden gevonden op de volgende URL: [http://www.inl.nl/pub/blauwdruk\\_volledig.pdf](http://www.inl.nl/pub/blauwdruk_volledig.pdf).

meerjarenbeleidsplan 2003 - 2007<sup>28</sup>. Voor deze opdracht wordt een jaarlijks budget van 500.000 euro gereserveerd. Gevolg gevend aan de aanbevelingen geformuleerd in de blauwdruk is de Taalunie inmiddels overgegaan tot de oprichting van een Vlaams-Nederlandse "TST-centrale", van waaruit alle met overheidsmiddelen ontwikkelde digitale materialen voor het Nederlands kunnen worden beheerd, onderhouden en beschikbaar gesteld. De Taalunie, die eigenaar wordt van de bij de TST-centrale ingediende materialen, staat daarbij in voor de organisatorische en regelgevende omkadering, terwijl het INL (met vestigingen in Leiden en Antwerpen), sinds 1 januari 2004 zorgt voor het technische en inhoudelijke beheer en onderhoud van de materialen. Gedurende het eerste werkingsjaar wordt het INL daarbij geruggensteund door het Max Planck Instituut voor Psycholinguïstiek. Voor elk van de digitale materialen die voor beheer, onderhoud, distributie en beschikbaarstelling aan de TST-centrale worden overgedragen (onder meer het zopas voltooide Corpus Gesproken Nederlands) wordt een adviescommissie opgericht die nauwlettend zal toezien op het beheer, onderhoud en terbeschikkingstelling van de betreffende materialen. Om versnippering te voorkomen zullen waar mogelijk adviescommissies voor bepaalde groepen materialen worden aangewezen. De voorzitters van deze commissies vormen samen de globale adviescommissie, die ook als raadgever zal zetelen in het Programmabestuur dat de uitvoering van STEVIN zal superviseren (zie verder hoofdstuk 7).

## 6.7. Internationalisering

Internationale afstemming van het onderzoek is van groot belang. De Programmacommissie zal activiteiten op dit gebied actief ondersteunen en aanmoedigen, o.a. door ingediende voorstellen op dit punt te toetsten. Vlaamse en Nederlandse experts zijn reeds zeer goed ingebed in het internationale taal- en spraaktechnologie veld. Dat blijkt uit de bovengemiddelde deelname van Nederlandse en Vlaamse experts aan conferenties en in Europese samenwerkingsprojecten als ook uit de positie van Nederland en Vlaanderen op de Europese HLT-scorecard (zie sectie 3.2). Al een groot aantal jaren wordt vanuit Nederland ELSNET (European Network of Excellence in Speech and Language Technology) gecoördineerd. En momenteel wordt door Nederlandse en Vlaamse experts meegewerkt aan het opzetten van een ERA-NET voorstel, LANGNET genoemd, dat in oktober 2004 zal worden ingediend. Het voorstel wordt getrokken door vertegenwoordigers van de Franse en Duitse ministeries voor onderzoek en technologie. Naast Duitsland, Frankrijk, Vlaanderen en Nederland zijn ook Denemarken, Finland, Italië, Noorwegen, Spanje, Tsjechië en Zweden betrokken.

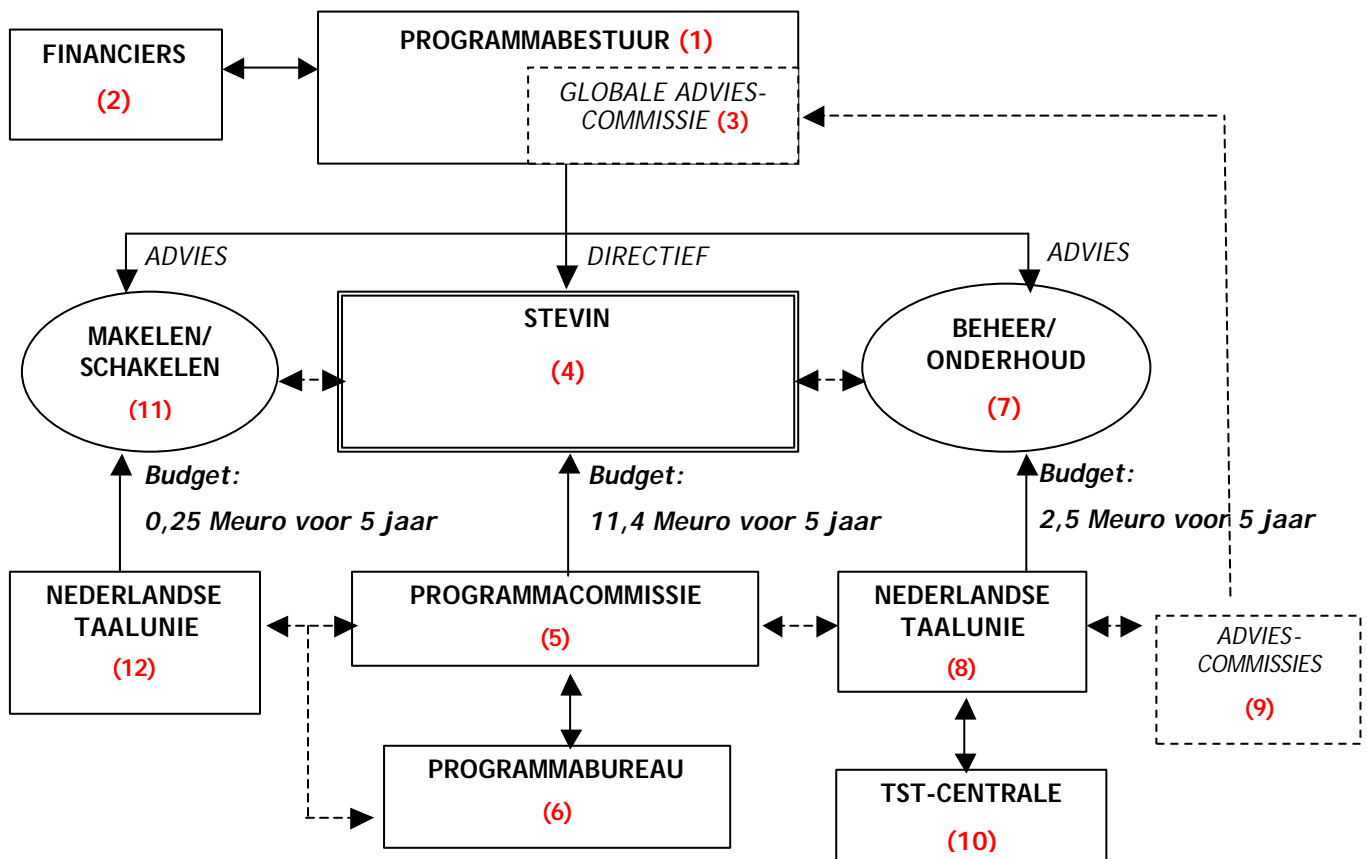
## 7. Organisatie van het programma

### 7.1. Inleiding en organigram

In feite heeft de beleidskeuze van het Nederlandse Ministerie van Economische Zaken, namelijk om te laten onderzoeken of een Vlaams-Nederlands TST-meerjarenprogramma eventueel kan bijdragen tot de verdere uitbouw van de Nederlandstalige digitale taalinfrastructuur, aanleiding gegeven tot een conceptueel kader waarin de verschillende actielijnen van het eerder genoemde TST-platform verder kunnen worden gerealiseerd in onderlinge verbondenheid, en waarin de verschillende aandachtspunten voor Nederlandstalige taal- en spraaktechnologie (i.e. ontwikkeling van basisvoorzieningen, strategisch onderzoek, vraagstimulering en intellectuele eigendomsrechten), verspreid over de diverse lagen van het innovatiesysteem, zo goed mogelijk aan bod kunnen komen. Dit kader wordt geconcretiseerd in het onderstaande organisatieschema. In de volgende paragrafen worden de taken en verantwoordelijkheden van de verschillende gremia nader uitgewerkt. Een samenvatting is opgenomen in *bijlage 5*.

---

<sup>28</sup> Het meerjarenbeleidsplan 2003 - 2007 van de Nederlandse Taalunie is beschikbaar vanaf de volgende URL: [http://www.taalunieversum.org/taalunie/publicaties/meerjaren\\_2003-2007.pdf](http://www.taalunieversum.org/taalunie/publicaties/meerjaren_2003-2007.pdf).



## 7.2 De Programmacommissie, het Programmabestuur en het Programmabureau

Het *Programmabestuur* (1) komt grotendeels overeen met het huidige TST-platform, en wordt in eerste instantie als volgt samengesteld:

- de Nederlandse Taalunie (als financieel coördinator);
- de vertegenwoordigers van de beoogde *financiers* (2). Dit zijn:
  - voor *Vlaanderen*: het Ministerie van de Vlaamse Gemeenschap (vertegenwoordigd door de administratie Wetenschap en Innovatie), het IWT-Vlaanderen en het FWO-Vlaanderen;
  - voor *Nederland*: het Ministerie van Economische Zaken (EZ), het Ministerie van Onderwijs, Cultuur en Wetenschap (OCW), de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) en SenterNovem;
- de "globale adviescommissie" (3). Hierin zitten de voorzitters van de adviescommissies die worden ingesteld voor elk digitaal bestand dat voor beheer, onderhoud, distributie en beschikbaarstelling in de TST-centrale wordt ondergebracht (zie hieronder). Eventueel wordt deze globale adviescommissie, die een raadgevende rol vervult in het Programmabestuur, nog aangevuld met enkele experts uit het veld.

Het Programmabestuur staat in voor de globale supervisie (directief) van STEVIN. Het Programmabestuur heeft tot taak de goede voortgang van dit onderzoeksprogramma te bewaken, toezicht te houden op de werkzaamheden van de andere actoren in de organisatiestructuur (d.w.z. de Programmacommissie, het Programmabureau en de Nederlandse Taalunie - zie hieronder), en als bemiddelaar op te treden bij eventuele geschillen. De leden van het Programmabestuur beoordelen STEVIN en de jaarwerkplannen zoals deze door de Programmacommissie worden opgemaakt (zie hieronder), selecteren de projecten die in het kader van elke oproep worden gefinancierd op basis van het advies geformuleerd door de Programmacommissie (zie hieronder), en dienen al deze stukken voor definitieve goedkeuring in bij de financierende instanties die zij vertegenwoordigen. Wat Vlaanderen betreft betekent dat concreet dat de vermelde stukken ter goedkeuring worden voorgelegd aan de Vlaamse minister bevoegd voor wetenschappen en technologische innovatie. In Nederland dienen de financierende instanties regelmatig op de hoogte gehouden te worden van de

beslissingen van het Programmabestuur. Indien in een latere fase nog bijkomende financiers (bijvoorbeeld vanuit de industrie) in STEVIN stappen, zullen ook zij worden vertegenwoordigd worden in het Programmabestuur en mee beslissen over het programmaverloop.

Een belangrijke schakel in de voorgestelde organisatiestructuur is de *Programmacommissie* (5). Deze commissie is verantwoordelijk voor de inhoudelijke uitwerking van STEVIN voor Nederlandstalige taal- en spraaktechnologie. De Programmacommissie wordt formeel ingesteld door het Programmabestuur. Concreet betekent dit dat de Programmacommissie instaat voor:

- de transformatie van de BaTaVo-prioriteitenlijst in een "geamendeerde BaTaVo", dus het onderhavige hybride *meerjarenprogramma* waarin de concrete onderzoeklijnen, de criteria waaraan de onderzoeksresultaten moeten voldoen en de instrumenten die daarbij zullen worden gehanteerd worden uiteengezet;
- het vertalen van STEVIN in concreet uitvoerbare *jaarwerkplannen*;
- het uitschrijven van *oproepen* tot het indienen van projectvoorstellen en het formuleren van een *advies* aangaande de in het kader van elke oproep te financieren projecten.

De samenstelling van deze Programmacommissie is als volgt:

NAAM	VL	NL	TT	ST	KI	I	ORGANISATIE
Jan Odijk (voorzitter)	X	X	X	X	X	X	Scansoft Belgium (80%) Universiteit Utrecht (20%)
vacature			X				vertegenwoordiger van een technologie provider
Jean-Pierre Martens	X			X	X		Universiteit Gent (ELIS) CLIF
Frank van Eynde	X		X		X		K.U.Leuven (CCL) CLIF
Walter Daelemans	X	X	X		X		Universiteit Antwerpen (CNTS) (80%) Universiteit Tilburg (20%) CLIF
Debbie Kenyon- Jackson		X	X			X	Polderland Language & Speech Technology BV Stichting NOTaS
Piek Vossen		X	X			X	Irion Technologies IMIX Programmacommissie
Arjan van Hessen		X		X	X	X	TeleCats BV Universiteit Twente (o.a. Parlevink Language Engineering Group) Stichting NOTaS
Lou Boves		X		X	X		K.U.Nijmegen (Onderzoeksgroep A2RT) IMIX (NWO) MMI (IOP)
Jeanine Beeken	X	X	X	X	X		Instituut voor Nederlandse Lexicologie (INL) TST-centrale
vacature	X		X			X	vertegenwoordiger klanten van TST-applicaties en/of -diensten
vacature		X					vertegenwoordiger klanten van TST-applicaties en/of diensten

Zoals uit bovenstaande tabel mag blijken, wordt bij de samenstelling van de Programmacommissie een evenwicht nagestreefd tussen Vlaanderen (VL) en Nederland (NL), taaltechnologie (TT) en

spraaktechnologie (**ST**), en kennisinstellingen (**KI**) en industrie (**I**). Er zijn nog drie vacatures binnen de Programmacommissie. Twee daarvan zijn specifiek bedoeld voor vertegenwoordigers van klanten van TST applicaties en/of diensten. Hierbij valt te denken aan een vertegenwoordiger uit een financiële instelling of bijvoorbeeld een omroeporganisatie.

Om belangenvermenging te vermijden staat de Programmacommissie dus enkel in voor de inhoudelijke uitwerking en opvolging van STEVIN. Alle voorstellen die door de Programmacommissie worden geformuleerd, worden door het Programmabestuur (dat waakt over neutraliteit en een objectieve besteding van de middelen) beoordeeld en vervolgens voor definitieve goedkeuring voorgelegd aan de financierende instanties in Vlaanderen (i.e. de Vlaamse regering, vertegenwoordigd door de Vlaamse minister bevoegd voor wetenschappen en technologische innovatie) en in Nederland (i.e. NWO en de ministeries van EZ en OCW).

Het *Programmabureau* (**6**) staat in voor het praktische projectbeheer (d.w.z. het uitsturen van oproepen, het bewaken van de projectvoortgang, het inwachten van de projectresultaten, het coördineren van de verslaggeving, het voeren van de nodige publicitaire acties, ...). Ook zorgt het Programmabureau voor de administratieve, secretariële en organisatorische ondersteuning van de Programmacommissie en het Programmabestuur. Het is in opdracht van het Programmabestuur tevens belast met de financiële afhandeling en begeleiding van STEVIN. De kosten van het Programmabureau komen ten laste van het programmabudget. Voor de verdeling van de middelen zal gebruikt gemaakt worden van de formele Vlaams-Nederlandse status van de Nederlandse Taalunie. Voor de praktische organisatie is de capaciteit en expertise van andere organisaties noodzakelijk. Meer concreet zal het Programmabureau bestaan uit NWO en SenterNovem, die gezamenlijk instaan voor het praktische beheer van STEVIN. Vooral waar het gaat over samenwerking en kennisoverdracht tussen kennisinstellingen en bedrijven is samenwerking tussen NWO en SenterNovem geboden. De bestaande samenwerkingsovereenkomst tussen NWO en SenterNovem biedt hiervoor een raamwerk.

De *Nederlandse Taalunie* (**8**) en (**12**) treedt bij de uitvoering van TST-meerjarenprogramma op als financieel coördinator. Met het oog op een volwaardig "Vlaams-Nederlands" programma is dit laatste een erg belangrijk gegeven. Bedoeling is immers dat de verschillende Vlaamse en Nederlandse financiers de middelen beschikbaar stellen aan de Taalunie, die op haar beurt deze middelen kan besteden in Vlaanderen en Nederland op basis van de werkzaamheden van de Programmacommissie (inhoudelijk) en het Programmabureau (praktisch), en onder supervisie van het Programmabestuur, zonder dat er nog een strikte geografische opsplitsing tussen Vlaanderen en Nederland hoeft te bestaan bij de besteding van de middelen (weliswaar wordt het evenwicht 1/3 Vlaanderen en 2/3 Nederland zoveel mogelijk nagestreefd). STEVIN kan een voorbeeld vormen in het kader van de lopende ontwikkelingen om te komen tot meer *cross-border funding* initiatieven (zie *Intentieverklaring voor de versterking van de strategische samenwerking tussen Vlaanderen en Nederland op het vlak van innovatie in bijlage 2*).

Daarnaast staat de Nederlandse Taalunie ook in voor de verdere uitvoering van de *makel- en schakelfunctie* (**11**) en voor *beheer, onderhoud, distributie en beschikbaarstelling* (**7**) van met overheidsmiddelen ontwikkelde digitale materialen. Beide taaklijnen zijn opgenomen in het meerjarenbeleidsplan 2003 - 2007 van de Taalunie, m.i.v. jaarbudgetten a rato van resp. 50.000 en 500.000 euro. De met overheidsmiddelen ontwikkelde digitale taalbestanden worden ondergebracht in de *TST-centrale* (**10**). De Taalunie, die eigenaar wordt van de ingediende materialen, staat daarbij in voor de organisatorische en regelgevende omkadering, terwijl het Instituut voor Nederlandse Lexicologie (INL) het technische en inhoudelijke beheer en onderhoud van de materialen op zich neemt. Voor elk digitaal bestand dat aan de TST-centrale wordt overgedragen wordt een *adviescommissie* (**9**) geïnstalleerd die toeziet op het beheer en onderhoud en de distributie en beschikbaarstelling van de materialen. De voorzitters van deze commissies vormen samen de *globale adviescommissie* (**3**), en die zal dus ook zetelen als raadgever in het Programmabestuur van STEVIN. Ook de basistaalvoorzieningen die in het kader van dit onderzoeksprogramma worden ontwikkeld, zullen na voltooiing in de TST-centrale worden ondergebracht, zodat beheer, onderhoud, distributie en beschikbaarstelling van deze materialen kunnen worden gegarandeerd.

Het onderhavige TST-meerjarenprogramma moet worden gezien als een "*geamendeerde BaTaVo*" (**4**). Dit wil zeggen dat de BaTaVo-prioriteitenlijst nog steeds blijft gelden als uitgangspunt, maar waar nodig geactualiseerd is en ingebed in het meer hybride karakter van het meerjarenprogramma. Daarbij zijn de nodige accenten gelegd zodat het globale programma is gericht op het realiseren van een adequate digitale taalinfrastructuur voor het Nederlands, maar daarnaast ook voldoende oog heeft voor de overige aandachtspunten die een hybride programma moet omvatten (d.w.z. strategisch onderzoek, vraagstimulering en intellectuele eigendomsrechten), en zodoende ook kan beantwoorden aan de specifieke eisen die uitgaan van de betrokken financiers, met name NWO en het Ministerie van EZ. Bovendien is - zoals werd aanbevolen in de *Technologieverkenning*

*Nederlandstalige Taal- en Spraaktechnologie* - een onderscheid gemaakt tussen "BaTaVo-data" en "BaTaVo-tools". Dit onderscheid werd intrinsiek reeds gemaakt in de BaTaVo-prioriteitenlijst, maar omdat beide componenten zich in feite op verschillende lagen van het innovatiesysteem (zie sectie 5.1) bevinden, en ook om de doelstellingen bij de concrete uitwerking van STEVIN duidelijk te kunnen afbakenen, verdient het aanbeveling de BaTaVo-prioriteitenlijst in "data" en "tools" op te splitsen. Hoewel de grens tussen data en tools over het algemeen helder getrokken kan worden, is het meestal wel het meest vruchtbaar om deze in samenhang te ontwikkelen, om redenen die eerder (zie 5.3.1) geschetst zijn.

Uitgaande van het hierboven beschreven organisatorische kader is het hybride TST-meerjarenprogramma voor Nederlandstalige taal- en spraaktechnologie opgezet en gerealiseerd door optimaal gebruik te maken van reeds bestaande structuren die voortvloeien uit de realisatie van het "*Actieplan voor het Nederlands in taal- en spraaktechnologie*" (zie hoofdstuk 2 hierboven). Bij het uittekenen van STEVIN zelf is voldoende rekening gehouden met de verschillende aandachtspunten voor Nederlandstalige taal- en spraaktechnologie. Daarbij staat de realisatie van een adequate digitale taalinfrastructuur voor het Nederlands nog steeds centraal, maar er wordt ook afdoende gereflecteerd over strategisch onderzoek, vraagstimulering en intellectuele eigendomsrechten, zodat de verschillende lagen van het innovatiesysteem kunnen worden afgebakend. De twee laatstgenoemde aspecten maken daarnaast ook expliciet deel uit van het takenpakket dat in het kader van respectievelijk de make- en schakelfunctie en het beheer en onderhoud en de distributie en beschikbaarstelling van digitale materialen verder zal worden uitgevoerd.

De Nederlandse Taalunie is met de TST-centrale bezig een plan uit te werken waarin beschreven staat welke onderhouds- en beheerstaken de TST-centrale hier specifiek voor levert. Na overleg met de Nederlandse Taalunie is besloten om 3% van het totaalbudget van STEVIN te reserveren voor het door de TST-centrale uit te voeren onderhoud en beheer. De Programmacommissie zal in eerste instantie zelf optreden als de adviescommissie, zoals bedoeld in (9) voor wat betreft de digitale materialen en andere onderzoeksresultaten die in het kader van dit meerjarenprogramma worden ontwikkeld en zullen worden ondergebracht in de TST-centrale.

### **7.3 Projectleiders en onderzoekers**

De projectleiders en onderzoekers zijn verantwoordelijk voor de uitvoering van de projecten. Zij zijn verantwoording schuldig aan de Programmacommissie, die zich kan laten adviseren door de Begeleidingscommissies. Ieder half jaar zal er een schriftelijke verslaglegging worden ingediend en een mondelinge toelichting gegeven tijdens een gezamenlijke bijeenkomst van onderzoekers, Programmacommissieleden en leden van de Begeleidingscommissies.

Naast het doen van onderzoek zijn projectleiders en onderzoekers medeverantwoordelijk voor de kennisoverdracht. Ter ondersteuning van de kennisoverdracht zullen verschillende activiteiten worden georganiseerd en middelen worden ingezet gedurende de looptijd van het onderzoeksprogramma. Projectleiders en onderzoekers worden geacht hieraan een bijdrage te leveren. Deze verplichting zal uitdrukkelijk vermeld worden in de oproepen voor het indienen van projectvoorstellen.

### **7.4. Werkgroep Flankerend Beleid**

De Werkgroep Flankerend Beleid zal meteen na de start van het programma door de Programmacommissie worden ingesteld en verantwoordelijk zijn voor de ontwikkeling en implementatie van het instrumentarium ter bevordering van de kennisoverdracht, netwerkvorming en verankering. In mei 2005 zal deze werkgroep komen met een specifieke uitwerking en planning van de activiteiten die in dit kader gepland zijn (zie sectie 6.2).

### **7.5. Begeleidingscommissie**

De Programmacommissie wil nog nader bepalen of er één gezamenlijke Begeleidingscommissie zal worden ingesteld of dat er voor elk project, of voor een cluster van nauw samenhangende projecten, een Begeleidingscommissie gevormd zal worden. Een Begeleidingscommissie bestaat uit vertegenwoordigers uit de industrie en de kennisinfrastructuur. In deze begeleidingscommissie dienen vertegenwoordigers van de verschillende typen gebruikers (zie sectie 6.1) vertegenwoordigd te zijn. De voorzitter is een lid van de Programmacommissie. Het Programmabureau zal de ondersteuning verzorgen. Een Begeleidingscommissie komt twee maal per jaar bijeen met de projectleiders en onderzoekers.

Een Begeleidingscommissie heeft de volgende taken en verantwoordelijkheden:

- Zij begeleidt een project actief, doet bijvoorbeeld suggesties voor verder onderzoek en helpt bij het maken van keuzes. Ze kan wetenschappelijke input geven en/of kan uitleg geven over de industriële situatie en de situatie in het bedrijfsleven in het bijzonder.
- Zij bewaakt de doelen van het project, en de besteding van tijd en (financiële) middelen. Hierbij wordt gelet op de voortgang van het project, de eventuele noodzaak van wijzigingen ten opzichte van het projectplan en onderbouwing van eventuele wijzigingen.
- Zij beoordeelt of er aspecten van het project octrooieerbaar zijn.
- Zij helpt mee resultaten te verspreiden onder een breder publiek.
- Zij stimuleert dat vervolgonderzoek zal plaatsvinden.
- Zij is mede verantwoordelijk voor het succes van een project en draagt bij aan het succes van STEVIN.

## 7.6 Belangenverstrengeling

Voor alle overheidsinstanties die subsidieregelingen uitvoeren speelt de problematiek rond persoonlijke betrokkenheid van leden van adviescolleges bij aanvragen een belangrijke rol. Zowel NWO als SenterNovem hebben naar aanleiding van de "Aorta-uitspraak" een uitgebreide richtlijn opgesteld. Daarin wordt een handleiding gegeven die de zorgvuldigheid van de procedure moet helpen garanderen. Dat geldt met name voor de richtlijnen voor het bepalen van en omgaan met betrokkenheid van leden van adviescommissie bij aanvragen dan wel aanvragers. Het algemene uitgangspunt is dat elke betrokkenheid uitgesloten moet worden. Echter zoals ook wordt geconstateerd in het advies van de directeur Algemene Beleidscoördinatie van EZ aan SenterNovem (zie *bijlage 3*) is dit bij op innovatie gerichte onderzoeksprogramma's niet altijd realistisch aangezien deze zich afspelen op overzichtelijke onderzoeksvelden, met beperkte aantallen spelers. Hierdoor zal het in de praktijk niet mogelijk zijn elke betrokkenheid van met name de leden van de Programmacommissie bij aanvragen en aanvragers uit te sluiten.

Binnen STEVIN zal er wel naar worden gestreefd om persoonlijke betrokkenheid van leden van de Programmacommissie geen rol te laten spelen bij de beoordeling van projectvoorstellen. Bij de beoordeling van voorstellen zal gebruik gemaakt worden van externe referenten of een internationale expertcommissie. Daarnaast nemen de leden van de Programmacommissie de volgende richtlijn in acht: "Indien de betreffende persoon betrokken is bij een projectaanvraag, dan onthoudt hij/zij zich van beoordeling en prioritering van het betreffende voorstel. Hij/zij kan wel deelnemen aan discussies van de voorstellen waarbij hij/zij niet betrokken is." In de verslaglegging van de beoordelingsvergaderingen zal aandacht besteed worden aan het aspect betrokkenheid en hoe hier mee is omgegaan.

## 8. Financiën van het programma

### 8.1 Middelen: MVG-AWI, EZ, NWO, OCW

In dit hoofdstuk wordt de financiële onderbouwing van STEVIN weergegeven. STEVIN wordt gefinancierd door een aantal Vlaamse en Nederlandse partijen.

Het Vlaamse gedeelte van de financiële middelen wordt gedragen door het ministerie van de Vlaamse Gemeenschap (MVG). Hiertoe wordt in de algemene uitgavenbegroting van de Vlaamse Gemeenschap een aparte kredietlijn ("Dotatie aan de Nederlandse Taalunie voor de uitvoering van het programma "Basistaalvoorzieningen voor het Nederlands") voorzien op het begrotingsprogramma "Strategisch en beleidsgericht onderzoek" van de administratie Wetenschap en Innovatie (AWI). Vanaf het begrotingsjaar 2004 wordt op deze kredietlijn jaarlijks een bedrag van 760.000 euro ingeschreven, en dit tot en met het begrotingsjaar 2008.

In Nederland zijn er drie financierende partijen: het Ministerie van Economische Zaken (EZ), het Ministerie van Onderwijs, Cultuur en Wetenschap (OCW) en de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). EZ zal zijn bijdrage (2.500.000 euro) voor de helft ten laste brengen van het CIC-fonds en voor de andere helft van het IOP-fonds. De bijdrage van NWO wordt bijeengebracht door de NWO gebieden Geesteswetenschappen en Exacte Wetenschappen en het Algemeen Bestuur. Een deel van deze bijdrage zal bestaan uit resultaten uit het reeds lopende NWO

IMIX-programma. Daarnaast lijkt er een mogelijkheid gevonden te zijn om een restantbudget uit een andere post over te hevelen naar deze bestemming. OCW tenslotte heeft laten weten zeer te hechten aan de beschikbaarheid van een adequate digitale taalinfrastructuur voor het Nederlands en de plannen om daar samen met Vlaanderen iets aan te doen. OCW gaat momenteel met NWO na via welke kanalen de totale bijdrage van deze beide organisaties in de verdere uitbouw van de Nederlandstalige digitale taalinfrastructuur op (afgerond) 5 miljoen euro voor 5 jaar kan worden gebracht. Het financieringskanaal dat daarvoor het meest in aanmerking komt is het Nederlandse Innovatieplatform<sup>29</sup>, een initiatief van de Nederlandse overheid om de innovatiekracht van Nederland te versterken, zodat het in 2010 opnieuw een koploper kan zijn in de Europese kenniseconomie. De OCW-bijdrage moet echter nog gerealiseerd worden.

## 8.2 Kennisontwikkeling

Voor kennisontwikkeling wordt 8.550.000 van het totaal vrij beschikbare budget begroot. Dit bedrag zal in minimaal in drie subsidierondes (oproepen) worden uitgezet. De oproep tot projectvoorstellen kan zowel aanbod-gedreven (*call for proposals*, open invulling van gevraagde prioriteiten) als vraag-gedreven (*call for tender*, specifieke onderzoeks- of ontwikkelingsvraag) zijn. Aan iedere oproep zal een *brokerage* worden voorafgegaan om één-op-één ideeënvorming en samenwerking te stimuleren van onderzoeksinstituten en bedrijven. Enige flexibiliteit is ingebouwd teneinde de mogelijkheid te creëren om op een bepaald moment een specifieke *call for tender* open te stellen om door de Programmacommissie geconstateerde ernstige lacunes in te vullen.

De eerste oproep met een omvang van maximaal 2.000.000 Euro zal eind september 2004 worden gepubliceerd en is gericht op kortlopende specifiek gerichte projecten. Begin 2005 zal de tweede oproep met een omvang van maximaal 4.250.000 euro worden gepubliceerd. In deze ronde kunnen ook grotere projecten, uit te voeren door consortia bestaande uit verschillende partners, worden aangevraagd.

Deze subsidieronde zal in principe in twee stappen worden opgedeeld:

1. uitnodiging tot indiening van vooraanmeldingen, één tot twee A4-tjes met daarop de essentie van het beoogde onderzoeksproject. Uit de ingediende vooraanmeldingen selecteert de Programmacommissie de voorstellen waarvan de indieners worden uitgenodigd een volledig voorstel uit te werken;
2. indiening van volledige voorstellen, externe beoordeling, prioritering door de Programmacommissie, honoreringsbesluit door het Programmabestuur.

Over de invulling van de derde oproep zal in de loop van 2006 door Programmacommissie en Programmabestuur besloten worden op basis van de evaluatie van de eerste twee oproepen.

Er zal ook naar gestreefd worden extra financiële bijdragen van bedrijven te verwerven. Belangrijker echter is dat gestreefd wordt naar het beschikbaar maken van dan wel voortbouwen op resources die het bedrijfsleven reeds gerealiseerd heeft. Dit laatste is ook tot uitdrukking gebracht in de beoordelingscriteria.

## 8.3 Flankerende activiteiten

Voor flankerende activiteiten, bedoeld om kennisoverdracht, netwerkvorming, zwaartepuntvorming, verankering, vraagstimulering en evaluatie (inclusief nulmeting) te bewerkstelligen wordt 1.947.000 begroot. Gezien het belang van de bovengenoemde activiteiten, die ook vraagstimuleringsactiviteiten omvatten, is besloten 17% van het beschikbare totaal budget hiervoor te reserveren. Dit percentage ligt hoger dan in menig ander onderzoeksprogramma maar dat is gerechtvaardigd vanwege de activiteiten die de Programmacommissie met name ten aanzien van vraagstimulering (zie 5.4) wil gaan uitvoeren. De Programmacommissie zal voor deze activiteiten op basis van input van o.a. de Werkgroep Flankerend Beleid (zie 7.4) en in overleg met andere partijen die reeds activiteiten op dit gebied uitvoeren een specifiek werkplan met tijdfasering opstellen. Voor de evaluatie (incl. nulmeting) zal het plan reeds in de zomer van 2004 worden uitgewerkt en voor het eind van het jaar worden uitbesteed aan een *trusted 3rd party*. Het werkplan voor de andere flankerende activiteiten, o.a. de organisatie van workshops, congressen, master classes etc (zie sectie 6.2) zal uiterlijk in mei 2005 ter goedkeuring worden voorgelegd aan het Programmabestuur.

---

<sup>29</sup> Zie ook <http://www.innovatieplatform.nl>.

## 8.4 Programmamanagement en beheer

Voor het programmamanagement wordt 641.000 euro gereserveerd. Dit is circa 5,5% van het beschikbare totaalbudget. Voor het beheer van de resultaten door de TST-centrale is -- na overleg met de Nederlandse Taalunie -- een bedrag gereserveerd van 300.000 euro, iets meer dan 3% van het budget dat voor kennisontwikkeling wordt uitgetrokken.

## 8.5 Planning middelen in de tijd

De onderstaande tabel geeft de verdeling van de middelen over de jaren weer (bedragen in keuro).

<b>Inkomsten</b>	<b>2004</b>	<b>2005</b>	<b>2006</b>	<b>2007</b>	<b>2008</b>	<b>2009</b>	<b>2010</b>	<b>Totaal</b>
MVG-AWI	760	760	760	760	760			<b>3.800</b>
EZ (IOP en CIC)		600	600	500	500	300		<b>2.500</b>
NWO-GW, NWO-EW en NWO-AB	220	370	120	120	120	50		<b>1.000</b>
NWO IMIX	50	50	50	50	50			<b>250</b>
OCW		500	500	600	450	450		<b>2.500</b>
te verwerven OCW/NWO/STW			250	250	400	450		<b>1.350</b>
<b>Totaal inkomsten</b>	<b>1.030</b>	<b>2.280</b>	<b>2.280</b>	<b>2.280</b>	<b>2.280</b>	<b>1.250</b>		<b>11.400</b>
<b>Committering</b>	<b>2004</b>	<b>2005</b>	<b>2006</b>	<b>2007</b>	<b>2008</b>	<b>2009</b>	<b>2010</b>	<b>Totaal</b>
Committering 1e oproep	1.000	1.000						<b>2.000</b>
Committering 2e oproep		1.000	1.000	1.150	1.100			<b>4.250</b>
Committering 3e oproep			500	1.000	800			<b>2.300</b>
<b>tot. committering nieuwe VL-NL TST projecten</b>	<b>1.000</b>	<b>2.000</b>	<b>1.500</b>	<b>2.150</b>	<b>1.900</b>			<b>8.550</b>
<b>Uitgaven</b>	<b>2004</b>	<b>2005</b>	<b>2006</b>	<b>2007</b>	<b>2008</b>	<b>2009</b>	<b>2010</b>	<b>Totaal</b>
reeds toegekend NWO IMIX	50	50	50	50	50			<b>250</b>
betalingen nieuwe VL-NL TST projecten		1.000	2.075	1.075	2.200	2.200		<b>8.550</b>
uitgaven flankerende activiteiten	121	606	363	363	242	242	9	<b>1.947</b>
uitgaven onderhoud en beheer				83	83	83	50	<b>300</b>
uitgaven Programmabureau	57	114	114	114	114	114	14	<b>641</b>
<b>totaal uitgaven</b>	<b>228</b>	<b>1.770</b>	<b>2.602</b>	<b>1.686</b>	<b>2.690</b>	<b>2.640</b>	<b>73</b>	<b>11.688</b>
<b>rente</b>	<b>10</b>	<b>66</b>	<b>53</b>	<b>86</b>	<b>69</b>	<b>3</b>		<b>288</b>

De kennisopbouw zal dus worden ingevuld door het organiseren van minimaal drie subsidierondes, waarbij in 2004, 2005 en 2007 subsidiebedragen zullen worden toegezegd aan onderzoekers. De met deze subsidierondes verbonden verplichtingen bedragen in totaal 8.550.000 euro. De betalingen vinden plaats tussen 2005 en 2009. De in de begroting opgenomen bijdrage uit het NWO IMIX-programma is reeds gecommiteerd. Het IMIX-programma heeft een totale omvang van 2.000.000 euro. Met de Vlaamse financier is overeengekomen dat 250.000 van dat bedrag als *matching* voor de Vlaamse bijdrage mag gelden. Rechten op tools en data voortkomend uit NWO IMIX worden overgedragen aan de Nederlandse Taalunie zodat deze tools en data ter beschikking gesteld kunnen worden.

De rente-inkomsten, begroot op 5% van de reserve, uit de Vlaamse en Nederlandse bijdragen die binnen de Nederlandse Taalunie beheerd worden zullen terugvloeien in het onderzoeksprogramma.

De flankerende activiteiten, inclusief het vervaardigen van demonstratieprojecten, zullen meteen vanaf de start van STEVIN een belangrijke rol spelen bij het verder uitbouwen van het netwerk tussen de bedrijven en de kennisinstellingen. In de latere fase zal de nadruk gaan verschuiven naar de verankering van de verworvenheden van het project. Ook de evaluatie en nulmeting zullen vanuit dit budget bekostigd worden.

Een ander belangrijk aspect is het beheer, onderhoud en beschikbaarstelling van de resultaten door de TST-centrale. Daarvoor wordt 3% van het totaalbudget gereserveerd dat in de tweede helft van STEVIN als de eerste resultaten beschikbaar komen, is begroot. De beheerskosten bedragen 5% van het totaalbudget en zijn gelijkelijk over de jaren verdeeld.

## 9. Activiteitenplan 2004 - 2005

<b>2004</b>	
begin juli	goedkeuring TST-meerjarenprogramma door financiers
zomer	formulering subsidieregeling en IPR regeling door bureau formulering 1 <sup>e</sup> oproep door Programmacommissie i.s.m. Programmabureau voorbereiding nulmeting door Programmacommissie i.s.m. Programmabureau
september	nadere omschrijving invulling Begeleidingscommissies door Programmacommissie <i>kick-off meeting</i> plus <i>brokerage</i> door Programmacommissie i.s.m. Programmabureau openstellen 1 <sup>e</sup> oproep
november	indienen volledige voorstellen kortlopende projecten beoordeling door internationale adviescommissie aanbesteding nulmeting
december	prioritering door Programmacommissie honoreringsbesluit door Programmabestuur

<b>2005</b>	
januari	1 <sup>e</sup> projecten opstarten PC formulering 2 <sup>e</sup> oproep door Programmacommissie i.s.m. Programmabureau
februari	organisatie <i>brokerage</i> 2 door Programmabureau
maart	openstellen 2 <sup>e</sup> oproep vooraanmeldingen
april	plan kennisoverdracht door Programmacommissie voorleggen aan Programmabestuur plan vraagstimulering door Programmacommissie voorleggen aan Programmabestuur
mei	selectie vooraanmeldingen door Programmacommissie opstellen concept activiteitenplanning 2006 - 2007 door Programmacommissie
september	indienen volledige voorstellen beoordeling door internationale adviescommissie
november	prioritering door Programmacommissie honoreringsbesluit door Programmabestuur formulering activiteitenplanning 2006 - 2007 door Programmacommissie en voorleggen aan Programmabestuur

## **Bijlage 1: Beleidsaanbevelingen (uit *Blauwdruk voor het beheer en onderhoud van met overheidsmiddelen gefinancierde digitale materialen*)**

### **9.1 Inleiding**

Het rapport *De positie van het Nederlands in Taal- en Spraaktechnologie* (Bouma en Schuurman 1998) houdt een hartstochtelijk pleidooi voor de versterking van het TST-onderzoek van het Nederlands, en dat zowel op het niveau van de immateriële als de materiële infrastructuur. Wat het eerste niveau betreft, is in de tussentijd al een overlegorgaan op het vlak van het beleid ingesteld, m.n. het TST-platform, dat tot in 2004 een bepaald takenpakket heeft uit te voeren. De materiële infrastructuur zal worden versterkt door het stimuleren van de ontwikkeling van allerlei hulpmiddelen en door het verbeteren van de beschikbaarheid van TST-materialen van het Nederlands. Welke basismaterialen bovenaan de prioriteitenlijst staan van zowel ontwikkelaars als gebruikers, blijkt uit het rapport van de actielijnen B en C van het TST-actieplan (Daelemans en Strik 2002). Onder TST-materialen verstaan wij corpora van geschreven en gesproken taal en spraakcorpora, software en trainingsmateriaal voor de diverse soorten verrijking van Nederlands taalmateriaal, alsmede elektronische woordenboeken en computationele lexica (zie 1.2.). Verwerving, onderhoud, beheer en voorwaarden en wijze van beschikbaarstelling van die materialen zijn even belangrijk. Alle hulpmiddelen die kunnen worden ingezet bij de ontwikkeling van TST-producten en bij wetenschappelijk onderzoek moeten beschikbaar zijn en moeten permanent worden onderhouden. Daarom pleiten we er hier voor dat de locatie van de acties die nodig zijn voor het verwerven, onderhouden en distribueren van de materiële infrastructuur wordt gecentraliseerd in de vorm van een consortium van gespecialiseerde instellingen, een TST-centrale met andere woorden. Tegen deze achtergrond is in deze *Blauwdruk* beschreven welke verschillende aspecten van verwerving, bewerking, administratie, verrijking, beheer, onderhoud en distributie van TST-materialen van vitaal belang zijn om tot een solide basis te komen waarop productontwikkeling en onderzoek kunnen plaatsvinden. Die beschrijving voert ons tevens tot de volgende aanbevelingen.

### **9.2 Aanbevelingen**

#### Aanbeveling 1. Een TST-centrale is noodzaak

Om te voorkomen dat basis-TST-materialen die met overheidsgelden buiten een permanente infrastructuur gemaakt zijn, niet voor hergebruik geschikt blijken of niet voortdurend worden onderhouden, is een rechtspersoon (TST-centrale) noodzakelijk. Daarbij kan gedacht worden aan instellingen die reeds gespecialiseerd zijn in TST-materialen en waarvan het onderzoeksprogramma in belangrijke mate gesubsidieerd wordt met overheidsgelden.

#### Aanbeveling 2. Betreft vorm van de TST-centrale en de rol Taalunie

De permanente infrastructuur kan de vorm aannemen van een binationaal consortium van instellingen met een statutaire opdracht. De centrale die wij aanbevelen dient niet lokaal, dat wil zeggen gebonden aan een universiteit of hogeschool, te zijn, maar landelijk of internationaal. Dat laatste vanwege de Nederlands-Vlaamse belangen op het gebied van de Nederlandse taal. Het gaat bij sommige TST-materialen bovendien om sterk specialistische bestanden zodat zowel bij de selectie en de verwerving als bij bewerking, verrijking, onderhoud en beheer vakspecialisten nodig zijn, die lang niet altijd in één land beschikbaar zijn. De coördinatie tussen de leden van dat consortium dient optimaal gewaarborgd te zijn. Daartoe zou een coördinator kunnen worden aangesteld die in dienst is van de Nederlandse Taalunie. Waarom de Nederlandse Taalunie? De Nederlandse Taalunie zal onder meer als drijvende motor achter het TST-platform, vaak de financiering en vooral de beschikbaarstelling van TST-materialen aan alle belanghebbenden of belangstellenden stimuleren. Zij zorgt dat de belangen en wensen van het TST-veld als geheel behartigd worden; zij zorgt voor de toepassing van algemeen aanvaarde standaards en voor sluitende juridische voorzieningen, maar zij bepleit ook de productie van hulpmiddelen bij financierende instanties. Zij fungeert als een soort makelaar. Ook dient de Nederlandse Taalunie te bevorderen dat onderzoeksfinanciers als universiteiten en de nationale en internationale onderzoekscoepels als subsidievoorwaarden stellen dat TST-materialen die met hun middelen tot stand gebracht zijn, voor onderhoud en beheer beschikbaar gesteld dienen te worden aan de TST-centrale. Concreet betekent dit dat bij projectaanvragen standaard middelen gereserveerd dienen te worden voor onderhoud en beheer.

### Aanbeveling 3. Betreft taken van de TST-centrale met prioritering

Een specificatie van de hoofd- en neventaken van de TST-centrale is gebaseerd op de volgende uitgangspunten: (a) TST-data en TST-software voortkomend uit tijdelijke, door de overheid gesubsidieerde projecten waarvoor geen permanente infrastructuur beschikbaar is, gaan per definitie naar de TST-centrale onder de beperking van aanbeveling 6 en (b) de distributie van TST-materialen dient te geschieden door daarin gespecialiseerde instanties als ELRA en LDC.

#### *Hoofdtaken*

##### Taak 1. Beheer

Onder beheer wordt verstaan het nemen van die maatregelen die bewerkstelligen dat data en software niet verloren gaan respectievelijk onbruikbaar worden. Onder beheer verstaan wij technisch beheer van TST-data, TST-software, systeemsoftware en apparatuur, inclusief documentatie. Voor uitgebreide informatie zie men hoofdstuk 5.

##### Taak 2. Toegankelijkheid van de data en software

Onder toegankelijk maken en houden verstaan wij het hergebruik van TST-materialen mogelijk maken. Daartoe behoort de technische, juridische, administratieve afhandeling van het traject dat loopt van ontwikkelaar, via TST-centrale naar distribuerende instantie of gebruiker (bij on-line toegang). Relevante hoofdstukken ter zake zijn hoofdstuk 7 en 2.

##### Taak 3. Onderhoud

Onder onderhoud verstaan we het nemen van die maatregelen die ervoor zorgen dat hergebruik van data en software op langere termijn mogelijk blijft. Hieronder vallen: (1) Het technisch onderhoud van: formaten van TST-data, TST-software, systeem- en applicatiesoftware, apparatuur en media. (2) Juridisch onderhoud van alle contracten. (3) Inhoudelijk onderhoud van: de TST-data inclusief annotaties, TST-software.

##### Taak 4. Gebruikersondersteuning

Onder gebruikersondersteuning verstaan wij de dienstverlening aan de gebruikers van de TST-data en TST-software die onder de verantwoordelijkheid vallen van de TST-centrale. Tot genoemde ondersteuning behoren het onderhouden van de website, de mailinglijst en helpdesk; het leveren van TST-data en TST-software op maat, softwareservice en advisering.

#### *Neventaak*

##### Taak 5. Verwerving

Onder verwerving wordt verstaan het actief verwerven of accepteren van TST-data en TST-software ontwikkeld door bedrijfsleven of door gevestigde onderzoeksinstituten. Onder verwerving valt de acquisitie van TST-data en TST-software waaraan een brede behoefte binnen het TST-veld is.

TST-materialen die geschikt zijn voor hergebruik zijn lang niet altijd met overheidsgelden ontwikkeld. Ondernemingen in Nederland en Vlaanderen hebben vaak al vele jaren geïnvesteerd in de ontwikkeling en productie van software en datasets. Het is nauwelijks denkbaar dat genoemde ondernemingen hun producten vrij beschikbaar zullen stellen. Dat immers conflicteert met het begrip 'commerciële exploitatie'. Toch dient vermeden te worden dat ten gevolge daarvan met overheidsgelden software en/of datasets die reeds in het bedrijfsleven bestaan, opnieuw ontwikkeld worden. Ook bij gesprekken en overleg hierover zou de Nederlandse Taalunie het voortouw dienen te nemen. Zo kunnen bijvoorbeeld bepaalde betaalde opdrachten aan bedrijven worden uitbesteed voor de ontwikkeling van nieuwe producten gebaseerd op de bestaande software en datasets. Die nieuwe producten kunnen dan gezamenlijk door het bedrijf en de TST-centrale worden geëxploiteerd. Tevens kan de Taalunie bedrijven stimuleren en uitdagen om aan te geven wanneer het voor hen aantrekkelijk is bij te dragen aan de ontwikkeling van TST-materialen en aan de ontwikkeling van een TST-infrastructuur.

### Aanbeveling 4. Kosten te dragen door overheid

Het takenpakket van de TST-centrale is te omvangrijk om als neventaak uitgevoerd te worden naast het onderzoeksprogramma van de leden van het consortium. Extra mankracht is daartoe nodig. Daarnaast is het redelijk te veronderstellen dat de verzameling TST-data en TST-software dermate groot is of groot wordt dat de materiële infrastructuur van de leden van het consortium niet toereikend is. Dat betekent dat er ook extra apparatuur nodig is. De kosten van extra personeel en apparatuur kunnen niet volledig worden gedekt door de gebruikers van de TST-centrale (vergelijk aanbeveling 5). Dit kan uitsluitend tot de conclusie leiden dat extra overheidsinvesteringen nodig zijn.

#### Aanbeveling 5. Kosten te dragen door gebruikers van de TST-centrale

Afhankelijk van het type gebruik en gebruiker dienen er algemene voorzieningen getroffen te worden om tot verschillende billijke tarieven te komen. Men kan daar uitvoerig over lezen in hoofdstuk 2. Indien er sprake is van bijzondere voorzieningen, d.w.z. van het op maat maken van TST-materialen, dan dient daarvoor een bedrag gefactureerd te worden dat minstens kostendekkend moet zijn.

#### Aanbeveling 6. Acceptatie van TST-data en TST-software door de TST-centrale

De TST-centrale kan TST-data en TST-software weigeren voor beheer indien ze niet aan bepaalde kwaliteitseisen (ook met betrekking tot documentatie) voldoen of indien ze niet essentieel zijn voor een ruim scala van toepassingen. Dit geldt ongeacht of die TST-data en TST-software ontwikkeld zijn door bedrijfsleven, gevestigde onderzoeksinstituten of op projectbasis buiten een permanente infrastructuur. De TST-centrale draagt zorg voor de opstelling van de acceptatie-eisen.

#### Aanbeveling 7. Internationale participatie

Om de positie van het Nederlands veilig te stellen in meertalig TST-onderzoek en -product-ontwikkeling dient de TST-centrale met steun van de beleidsmakers in staat gesteld te worden te participeren in Europese en/of mondiale projecten die gerelateerd zijn aan haar taken. Zowel op internationaal als op nationaal niveau dient de TST-centrale op basis van zijn praktijkervaring bij te dragen aan de vorming van standaarden en aan methoden voor het evalueren en valideren van TST-taalmaterialen.

#### Aanbeveling 8. Ontwikkeling en behoud TST-expertise

Gezien de schaarste aan spraak- en taaltechnologen dient de overheid zorg te dragen voor een beleid waarin TST-expertise ontwikkeld wordt en behouden blijft.

## **Bijlage 2: Intentieverklaring voor de versterking van de strategische samenwerking tussen Vlaanderen en Nederland op het vlak van innovatie**

De Nederlandse Minister van Economische Zaken en de Vlaamse Minister van Financiën en Begroting, Ruimtelijke Ordening, Wetenschappen en Technologische Innovatie,

Overwegende het belang en het potentieel van de kennisdelta Eindhoven - Leuven - Aken;

Overwegende de voorbeeldfunctie die uitgaat van de as Eindhoven - Leuven in het kader van de versterking van de bilaterale strategische samenwerking tussen Vlaanderen en Nederland op het vlak van innovatie;

Overwegende het cruciale belang om ook standpunten uit te wisselen ten aanzien van Europese en andere multilaterale initiatieven waarbij de overheid strategische keuzes moet maken, daarbij inspeland op de 3% doelstelling van de Lissabon-strategie waarmee de Europese Unie zichzelf ten doel stelt tegen 2010 de meest concurrerende en dynamische kenniseconomie van de wereld te worden, gericht op duurzame economische groei met meer en betere banen en een hechtere sociale samenhang;

Hebben de gezamenlijke intentie om voor zover dit kan binnen de hen gegeven bevoegdheden:

1. stimulerend en faciliterend op te treden, door het uitbreiden van de gezamenlijke Vlaams-Nederlandse coördinatie, afstemming en regie op het vlak van onderwerpen van strategisch belang vanuit overheidsperspectief, met het oog
2. de grensoverschrijdende samenwerking te intensiveren, mede in het kader van programma's die in het bijzonder door de Europese Unie worden ontwikkeld en die kunnen worden ingezet bij het streven naar het realiseren van de doelstellingen geformuleerd in het kader van de genoemde Lissabon-strategie.
3. bij te dragen tot het scheppen van gunstige voorwaarden voor het leggen van rechtstreekse contacten en voor de activiteiten van bedrijven en andere rechtspersonen, voor het aanmoedigen van investeringen en het bevorderen van de uitwisseling van innovatiegerelateerde informatie.

Teneinde de omschreven doelen te bereiken zullen de ministers ernaar streven:

1. ter versterking van regionale innovatie over te gaan tot het in gang zetten van een structurele strategische dialoog, waarbij de contacten tussen de betrokken bewindslieden en administraties zullen worden geïntensiveerd door middel van het uitwisselen van informatie en documentatie en door regelmatig overleg.
2. de rechtstreekse samenwerking tussen hun regionale en lokale overheden - alsmede organisaties, instellingen, bedrijven en personen - te bevorderen, teneinde meer synergie en een meer eenduidig gericht beleid tot stand te brengen.
3. door waar opportuun gezamenlijke standpunten in te nemen die de toegang tot internationale partnerschapsnetwerken bevorderen, aansluitend bij multilaterale of supranationale programma's, in het bijzonder deze welke door de Europese Unie worden geïnitieerd.
4. concrete grensoverschrijdende activiteiten op te zetten, waaronder:

- op het vlak van taal- en spraaktechnologie, het opstarten van een Vlaams-Nederlands samenwerkingsverband voor de verdere stimulering van Nederlandstalige taal- en spraaktechnologie;
- op het vlak van *automotive industry* te streven naar een Euregionaal *competence centre* voor assembleurs, toeleveranciers en de toptechnologische bedrijven. Gestart wordt vanuit het ATC in Nederland en Flanders' Drive in Vlaanderen. De samenwerking zal gezocht worden met CAR in Nord Rhein Westfalen;
- het nagaan van de mogelijkheden voor het ontwikkelen van een gezamenlijke visie op gemeenschappelijke onderwerpen binnen ICT onderzoek en innovatie;
- het opzetten van een structureel forum ter bevordering van innovatie in de regio en daarbij gebruik makend van Europese en andere initiatieven;
- het uitwerken van experimenten met de aanwending van bij voorkeur EU-middelen ten behoeve van onder meer starters en kennisvouchers, en het laten groeien van initiatieven met gezamenlijke, grensoverschrijdende inzet van middelen, zoals het reeds vermelde samenwerkingsverband ten behoeve van Nederlandstalige taal- en spraaktechnologie.

De ministers zijn voornemens elkaar te ontmoeten om hun bilaterale samenwerking en concrete onderwerpen van gemeenschappelijk belang te bespreken.

Deze intentieverklaring werd ondertekend in Leuven op 7 april 2004 en gaat van kracht op de datum van ondertekening. De intentieverklaring werd opgemaakt in twee authentieke versies waarvan elke partner verklaart er één ontvangen te hebben.

De Nederlandse Minister van  
Economische Zaken

De Vlaamse minister van Financiën  
en Begroting, Ruimtelijke Ordening,  
Wetenschappen en  
Technologische Innovatie

L.J. Brinkhorst

D. Van Mechelen

### Bijlage 3: Aanbevelingen technologieverkenning M&I Partners/Montemore

1. TST kan een aanzienlijke bijdrage leveren tot duurzame economische groei. Het is daarbij niet zozeer de vraag of TST een economische impact heeft. Dit lijkt eerder triviaal. De relevante vraag is of het nationale innovatiesysteem de output kan genereren om die impact op de economie daadwerkelijk te genereren, of dat Vlaanderen en Nederland een integrator worden van technologie die grotendeels elders is ontwikkeld, zelfs voor de moedertaal.
2. Het model van het dynamisch innovatiesysteem (DIS) is goed bruikbaar voor het formuleren van een programmatische aanpak. In dit geval kan de aanbodzijde het beste in drie lagen worden opgesplitst, namelijk TST-basisvoorzieningen (eerste laag), TST-onderzoek en ontwikkeling in enge zin (tweede laag) en integratie/*embedding* van TST-componenten (derde laag). Het DIS levert met name het inzicht op dat stimulering op de derde laag (integratie en *embedding*) niet mogelijk is. Dit levert een extra argument op om prioriteit te geven aan vraagstimulering, naast de aanmaak van basistaalvoorzieningen en strategisch onderzoek.
3. Een programmatische aanpak is de aangewezen manier om een extra injectie aan taal- en spraaktechnologie te geven. Die programmatische aanpak moet een hybride karakter hebben en mag niet beperkt blijven tot de financiering van basistaalvoorzieningen in de enge zin van data en componenten.  
Er is veel draagvlak voor een dergelijke programmatische aanpak. Er is een continue discussie tussen marktpartijen en publieke kennisinstellingen over de juiste prioriteiten. Het goed regelen van het intellectueel eigendom van data, tools en/of modules is cruciaal om een goede samenwerking en draagvlak te behouden.
4. Gelet op de ambitieuze doelstelling om achterstanden voor de eigen Nederlandse taal in te halen en om op wereldschaal kopposities in deelterreinen in te nemen (of te behouden) in het gebruik van taal en spraak, wordt geadviseerd om over te gaan tot het opzetten van een hybride onderzoeksprogramma voor Nederlandstalige taal- en spraaktechnologie dat de verschillende lagen van het innovatiesysteem omvat. Dit betekent concreet dat de ontwikkeling van de benodigde Nederlandstalige basisvoorzieningen op basis van de BaTaVo prioriteitenlijst integraal deel uitmaakt van dit programma, maar dat daarnaast ook de nodige aandacht wordt besteed aan strategisch onderzoek, vraagstimulering en intellectuele eigendomsrechten.

## Bijlage 4: Relatie met andere lopende (internationale) programma's en projecten

Het onderzoeksprogramma is gerelateerd aan een aantal nationale en internationale lopende onderzoeksprogramma's. In Nederland lopen reeds een aantal grote programma's die zich deels op hetzelfde onderzoeksgebied richten dan wel op aanverwante onderzoeksgebieden. Sterk gerelateerde programma zijn het NWO *IMIX* onderzoeksprogramma, het IOP<sup>30</sup> *MMI* programma, het BTS<sup>31</sup>-project *Waterland* en het CIC<sup>32</sup> programma *Pidgin*. Maar ook in de Bsik<sup>33</sup> programma's *MultimediaN* en *ICIS* worden een aantal aanverwante projecten uitgevoerd (zie sectie 5.6). IWT-Vlaanderen<sup>34</sup> heeft in het kader van de SBO-regeling (strategisch basisonderzoek) twee gerelateerde projecten gefinancierd *FlaVoR* en *AtraNoS*. Daarnaast is er reeds een lopend project waarin Vlaanderen en Nederland samenwerken dat wordt gefinancierd door FWO-Vlaanderen<sup>35</sup> - NWO via het VNC<sup>36</sup>-programma: *PROSIT*.

### NWO<sup>37</sup> onderzoeksprogramma Interactieve Multimodale Informatie Extractie (IMIX)

Het NWO programma IMIX stelt zich tot doel om kennis en technologie te ontwikkelen die nodig zijn om specifieke antwoorden op specifieke vragen in Nederlandstalige documenten te kunnen zoeken.

Zoekmachines als Google zijn een eerste stap op weg naar het automatisch vinden van relevante informatie in het World Wide Web. Maar als Google terugkomt met duizenden links die waarschijnlijk relevant zijn, blijft het een tijdrovende taak om de specifieke informatie te vinden.

De vragen waar mensen mee zitten zijn niet altijd even specifiek. Soms worden ze dat pas in een dialoog met een specialist, door nieuwe vragen te stellen die eerdere antwoorden kunnen verduidelijken. Veel mensen geven voor zo'n dialoog de voorkeur aan spreken boven typen. Bij het presenteren van antwoorden kan een tabel, figuur of grafiek soms veel duidelijker zijn dan een lange omschrijving. En informatiebestanden krijgen zelf steeds meer een multimedia karakter, door interactieve afbeeldingen in teksten en de combinatie van beeld en spraak in video-archieven.

Binnen het programma zijn vier specifieke onderzoeksthema's als prioriteit aangewezen: a) automatische spraakherkenning in de context van multimodale interactie; b) dialoogmanagement en redenering; c) taal-centrische presentatie van informatie in multimodale systemen; d) informatie extractie (*question-answer*ing).

De resultaten van het onderzoek in het IMIX programma zullen geïntegreerd worden in een praktisch werkende demonstrator. Met behulp van die demonstrator kunnen de mogelijkheden om de nieuwe kennis en technologie in te zetten voor concrete toepassingen over het voetlicht gebracht worden en wordt kennisuitwisseling bevorderd. De looptijd is vijf jaar: 2003 - 2007. Meer informatie over het programma is te vinden op <http://www.nwo.nl/imix>.

### SenterNovem Innovatiegericht Onderzoeksprogramma Mens-Machine Communicatie (IOP-MMI)

Het vakgebied "Mens-Machine Interactie" streeft ernaar om opties te genereren voor een gemakkelijke en plezierige interactie met systemen, producten en diensten. De strategie is hierbij het ontwerpen van intelligente systemen met natuurlijke, persoonlijke en adaptieve interfaces. Hierbij wordt gebruik gemaakt van vele geavanceerde technologieën, zoals *pen-based computing*,

---

<sup>30</sup> IOP: EZ subsidieinstrument Innovatiegerichte Onderzoeksprogramma's (<http://www.senter.nl/iop>)

<sup>31</sup> BTS: EZ subsidieinstrument voor Bedrijfsgerichte Technologische Samenwerkingsprojecten (<http://www.senter.nl/asp/page.asp?id=i000008&alias=technologischesamenwerking>)

<sup>32</sup> Het programma Concurrenieren met ICT Competenties (CIC) ontwikkeld door de ministeries van Economische Zaken en Onderwijs, Cultuur en Wetenschap heeft tot doel ICT doorbraken te stimuleren. Voor meer informatie zie <http://www.cic-online.nl>)

<sup>33</sup> Bsik: Besluit subsidies investeringen kennisinfrastructuur (vh. ICES/KIS) (<http://www.senter.nl/bsik>)

<sup>34</sup> IWT-Vlaanderen: Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen (<http://www.iwt.be>)

<sup>35</sup> FWO-Vlaanderen: Fonds voor Wetenschappelijk Onderzoek - Vlaanderen (<http://sun.fwo.be/>)

<sup>36</sup> VNC: Vlaams Nederlands Comité: Programma voor Vlaams Nederlandse samenwerking. (<http://www.nwo.nl/vnc>)

<sup>37</sup> NWO: Nederlandse Organisatie voor Wetenschappelijk Onderzoek (<http://www.nwo.nl>)

*computer vision*, spraaktechnologie, zelflerende algoritmes, agents en dergelijke. Het onderzoek kan ook het ontwikkelen van methoden en gereedschappen omvatten, die de bedoeling hebben om ontwerpers en gebruikers systematisch in het interface ontwerp te betrekken. In de eerste fase van het programma richtte men zich op het klassieke "user interface" (ergonomie, human factors) in de zin van "interactie met het apparaat. In de tweede fase zal het IOP-MMI zich gaan richten op de interactie met het achterliggende systeem, met de applicaties. Het systeem moet de gebruiker werk uit handen nemen; niet passief zijn instructies uitvoeren maar pro-actief zijn intenties ondersteunen. Het dominerende thema van het nieuwe onderzoeksprogramma is het genereren van "(systeem)intelligentie" en het faciliteren van de gebruikersinteractie hiermee. Centraal staat de vraag: "Welke kennis van elkaar moeten systeem en gebruiker op welke wijze verwerven en toepassen om samen de gebruiker optimaal te ondersteunen bij het bereiken van zijn doel?" De looptijd is acht jaar: 2000 - 2008. Meer informatie over het programma is te vinden op <http://www.senter.nl/iop-mmi>.

### **CIC PidGin - Zelflerend Cross Lingual Interface**

Onder de naam PidGin heeft het Delftse bedrijf Irion Technologies B.V. - voortgekomen uit TNO - de aanzet gegeven tot een chatsysteem dat in de communicatie op het internet, taaluitingen van twee kanten kan vertalen. PidGin - dat mede tot stand komt door de financiering vanuit het Programmabureau CIC - is zelflerend; het achterliggende systeem leert uit de chatsessies hoe uitingen het beste kunnen worden vertaald. Daarbij draait het niet om perfectie. Het gaat er vooral om dat duidelijk wordt wat de bedoelingen van de gesprekspartners zijn.

De betrokken partijen zijn: Irion Technologies B.V. (Delft), Carp Technologies B.V. (Enschede), Universiteit Twente (Enschede), New Law Facilities B.V. (Leiden) en Intex B.V. (Geldrop). De looptijd is twee jaar: 2002 - 2004. Meer informatie over dit project is te vinden op: <http://www.pidgin.nl/index2.htm>

### **BTS Workflow with semi-Automatic meTadata Extraction for fully digital media pRoduction in the NetherLANDs (BTS- Waterland)**

Doel van dit project is de digitalisering van het productieproces te stimuleren. Waterland richt zich met name op het generieke content productieproces en de *packaging* en *broadcast* van televisie en radio content. In dit project wordt samengewerkt door de NOB Cross Media Facilities, de Nederlandse Omroep Stichting (NOS), de Nederlandse Organisatie voor Toegepast Onderzoek - TNO, de Universiteit Twente en het CWI - Centrum voor Wiskunde en Informatica in Amsterdam. De looptijd is 4 jaar: 2002 - 2005. Meer informatie is te vinden op: <http://www.innovatie.nob.nl/waterland/>

### **IWT Flexibel Large Vocabulary Recognition: Incorporating Linguistic Knowledge Sources Through a Modular Recogniser Architecture (IWT- FlaVoR)**

De voorbije decennia is er grote vooruitgang geboekt in de performantie van spraakherkennings-systemen met grote vocabularia, door het steeds maar uitbreiden en verfijnen van de in de jaren '70 en '80 ontwikkelde HMM-gebaseerde herkenningsstrategie. De belangrijkste vooruitgang werd geboekt door de verbetering van de statistische akoestisch-fonetische modellering. Ondanks alle vooruitgang is spraakherkenning echter nog steeds verre van perfect en vereisen commerciële systemen steeds opnieuw het ontwikkelen van "subtalen" die betrekking hebben op een duidelijk afgelijnd onderwerp. Bovendien moeten gebruikers zich aanpassen aan de beperkingen van de systemen. De noodzaak om een herkenner aan te passen en te optimaliseren voor elke nieuwe taak is één van de basishindernissen die het verder verspreiden van de technologie in de weg staan. Een bijkomend probleem is het feit dat taakspecifieke (semantische) informatie niet geïsoleerd kan worden, maar steeds volledig verweven is met syntactische informatie. Verder moet voor zowel de semantische als voor de syntactische informatie teruggevallen worden op zeer eenvoudige formalismen. Dit komt door de alles-in-één strategie die in de huidige spraakherkenners gevolgd wordt. Mede omdat herkenning op basis van akoestische informatie alleen nog steeds als onvoldoende goed ingeschat wordt, wordt het als cruciaal aanvaard het taalmodel zo vroeg en zoveel mogelijk te gebruiken in het zoekproces. Dit kan het best met zo'n alles-in-één strategie. Dit uitgangspunt heeft echter een enorme beperkende invloed gehad op de mogelijkheden voor de taaltechnologische component in een spraakherkenner. De linguïstische beperkingen die vrijwel alle herkenners aan banden leggen zijn (i) het gebruik van een volledig geëxpandeerd lexicon waarin alle (gewenste) mogelijke woordvormen van dezelfde stam expliciet dienen opgenomen te zijn, en (ii) het gebruik van een statistisch N-gram taalmodel. Binnen dit project wordt de klassieke architectuur

met een volledige geïntegreerde zoekstrategie volledig herbekeken en vervangen door een gelaagde structuur. De eerste laag is generisch voor een taal en genereert als output metadata waarop de volgende laag verder kan werken. Deze metadata bevat meerdere stromen informatie, enerzijds de akoestisch-fonetische (een foneemnetwerk), anderzijds intonatie-gerelateerde parameters en informatie betreffende de spreker. De tweede laag is een zoekproces dat start van de metadata en gestuurd wordt door zowel generische als domeinspecifieke linguïstische informatiebronnen (morfofonologie, morfo-syntax, zinssyntax, etc.). Hierbij moet worden opgemerkt dat de metadata een dicht netwerk met waarschijnlijke hypothesen is en niet een enkelvoudige opsomming van alternatieven.

In dit project wordt samengewerkt door de KU Leuven en de Universiteit Antwerpen. De looptijd is 4 jaar: 2002 - 2006. Meer informatie is te vinden op:  
<http://www.esat.kuleuven.ac.be/~spch/projects/FLaVoR/>

### **IWT Automatic Transcription and Normalisation of Speech (IWT- AtraNOS)**

De voorbije jaren is reeds veel vooruitgang geboekt wat betreft de ontwikkeling en het gebruik van automatische spraakherkenningssystemen voor bvb. dicteerapparatuur en computergestuurde telefoondiensten. Er kan echter verwacht worden dat de markt voor automatische spraakherkenningssystemen slechts spectaculair zal groeien, wanneer niet enkel voorgelezen spraak kan herkend worden, maar ook spontane spraak, ongeacht de opnameomstandigheden. Om deze overgang te maken moeten echter nog een reeks van problemen worden opgelost. Ten eerste gaat het bij dicteersystemen telkens om één spreker, waar de herkenner zich aan kan aanpassen. Bij continue spraak is het nodig verschillende sprekers te onderscheiden. Ten tweede bestaat de invoer bij dicteersystemen per definitie uit goed voorbereide en grammaticaal correcte zinnen, wat bij continue spraak niet het geval is. Al deze factoren zorgen ervoor dat het foutpercentage voor de transcriptie van spontane spraak door de huidige herkenningssystemen nog steeds te hoog is voor praktische toepassingen. De algemene doelstelling van het project is de verbetering van de technologie voor de automatische transcriptie van spontane spraak en voor de conversie van deze transcripties in een vorm die beter aangepast is aan de noden van de gebruikers. Een toepassing die zal bestudeerd worden als gevalstudie is het genereren van ondertiteling bij televisie-uitzendingen, dit ten behoeve van gehoorstoorden. Meer specifiek kunnen de beoogde resultaten van het project als volgt samengevat worden:

- Een methode om betrouwbaar een continue audio-stroom te segmenteren in homogene segmenten (met slechts één type spraak) en om die segmenten te beschrijven aan de hand van een aantal kenmerken.
- Methodes om tegen te gaan dat automatische transcripties moeilijker leesbaar worden door de aanwezigheid in de spraak van woorden die het herkenningssysteem niet kent (zogenaamde *out-of-vocabulary* woorden)
- Methodes om de negatieve invloed die aarzelingen, herhalingen, afgebroken woorden, etc. - overvloedig aanwezig in spontane spraak - op de correctheid van automatische transcripties hebben, tegen te gaan.
- Methodes, zowel statistische als kennisgebaseerde, om woordelijke, automatische transcripties van spraak om te zetten in ondertiteling.

In dit project wordt samengewerkt door de K.U. Leuven, de Universiteit Gent en de Universiteit Antwerpen. De looptijd is 4 jaar: 2000 - 2004. Meer informatie is te vinden op:  
<http://atranos.esat.kuleuven.ac.be/>

### **PROSIT (Text Analysis and Machine Learning for Prosody)**

Doel van dit project is empirisch te onderzoeken of een natuurlijk klinkende prosodie kan worden gegenereerd op basis van twee methodes die recent succesvol zijn gebleken in andere taalverwerkingsdomeinen: (a) robuuste analyse van tekst met behulp van technieken uit *information retrieval* en *information extraction*, en (b) geavanceerde zelflerende en meta-lerende systemen.

In dit project wordt samengewerkt door de Universiteit Antwerpen (CNTS) en de Universiteit van Tilburg (ILK). De looptijd is 4 jaar: 2001 - 2004. Meer informatie is te vinden op:  
<http://cnts.uia.ac.be/cnts/projects/2001prosody.html>.

## EU-projecten

Ook zijn Vlaamse en Nederlandse onderzoekers als partner betrokken (geweest) bij een groot aantal internationale onderzoeksprogramma's. De belangrijkste nog lopende projecten daarvan zijn:

- FP5-IST MUMIS (Multimedia and Searching Environment), 2000-2002. Partners Universiteit Twente (CTIT), Max Planck Instituut voor Psycholinguïstiek, Katholieke Universiteit Nijmegen, DFKI, Universiteit Sheffield, ESTEAM. <http://parlevink.cs.utwente.nl/projects/mumis.html>
- FP5-IST SMADA (Speech-driven Multimodal Automatic Directory Assistance), 2000-2003. Partners KPN Research (tot medio 2001), France Télécom R&D, Csel/Loquendo, Alcatel SEL, Université d'Avignon, Politecnico di Torino, Katholieke Universiteit Nijmegen.
- FP5-IST COMIC (Conversational Multimodal Interaction with Computers), 2002-2004. Partners Max Planck Instituut voor Psycholinguïstiek, Max Planck Institut für Biologische Kybernetik, Katholieke Universiteit Nijmegen, DFKI, Universiteit Sheffield, Universiteit Edinburgh, ViSoft GmbH. <http://www.hcrc.ed.ac.uk/comic/>
- FP5-IST (Multilingual Subtitling of Multimedia Content), 2002- 2005. Partners: ILSP, Greece (coordinator), CNTS, BBC, Systran, Lumiere Cosmos, ESAT/PSI. <http://sifnos.ilsp.gr/musa>
- FP5-IST M4 (MultiModal Meeting Manager). 2002-2005. Partners: U Sheffield, U Edinburgh, EPF Lausanne, TU München, IDIAP, TNO TPD, U Geneva, U Twente, Brno UT, The International Computer Science Institute, USA. <http://www.dcs.shef.ac.uk/spandh/projects/m4/>
- FP5-QoL BIOMINT (Biological Text Mining), 2003-2005. Partners University of Manchester (coordinator), CNTS, PharmaDM, OFAI, SIB, University of Geneva. <http://www.biomint.org/>
- FP6-IST 2002-METIS (Statistical Machine Translation using Monolingual Corpora). 2002-2003. Partners: ILSP, Greece (coordinator), KU Leuven. <http://www.ilsp.gr/metis/>
- FP6-IST Network of Excellence, PASCAL: (Pattern Analysis, Statistical Modelling and Computational Learning), 2003-2007. <http://www.pascal-network.org/>
- FP6-IST Integrated Project SAFIR (Speech Automatic Friendly Interface Research). 2003 - 2007. Partners: BASF IT Services, Voice Insight, Panasonic European Laboratories >li> Panasonic Speech Technology Laboratories, GFI Informatique, ALIS Europe S.A., Geodan Mobile Solutions, MDS Marathon Data Systems, Region Wallonn, Joint Research Centre ISPRA, JPASS International, BCCI, NBT AD, VRATSA Region, Ulearn2B, Brussels Region Informatics Centre, THALES Communications, U Twente. <http://www.amiproject.org/>
- FP6 IST Integrated Project AMI (Augmented Multi-party Interaction), 2004-2007. Partners: IDIAP, DFKI, ICSI, TNO,BUT, TUM, UEDIN, USFD, UT, FastCom, Novauris, Philips, RealVNC, Spiderphone S.A. <http://www.amiproject.org>.
- COST action COST278 (Spoken Language Interaction in Telecommunication), 2003-2006. Partners: institutions from 19 European countries (e.g. KUN, TUE, Ugent, Multitel, etc.) <http://www.cost278.org>.

## Bijlage 5: Overzicht omvang Kennisinstructuur:

In 2003 is door M&I/Partners in samenwerking met Montemore een *Technologieverkenning Nederlandstalige Taal- en Spraaktechnologie* uitgevoerd. Onderdeel daarvan was een overzicht van de omvang van het taal- en spraaktechnologische onderzoeksveld. De belangrijkste gegevens zijn in de drie hieronder staande tabellen met enige kleine rekentechnische correcties herhaald.

*Omvang en financiering onderzoeksgroepen: In onderstaande tabel is de omvang van het onderzoeksveld TST in aantal fte (fulltime equivalenten onderzoekscapaciteit) in kaart gebracht.*

Fte in Taal en Spraak	Vlaanderen			Nederland		
	taal	spraak	som	taal	spraak	som
hoogleraar	3,8	3,7	7,5	6,1	1,8	7,9
wetenschappelijk personeel	27,0	10,8	37,8	36,5	23,2	59,7
AIO/OIO	0	4,0	4,0	36,0	7,0	43,0
technisch medewerker	0	0,5	0,5	6,9	2,7	9,6
studenten	4,0	3,0	7,0	36,0	7,0	43,0
Totaal (excl. studenten)	30,8	19,0	49,8	85,5	34,7	120,2

*Of anders geordend:*

Fte in Taal en Spraak	Taal			Spraak			totaal
	Vlaanderen	Nederland	som	Vlaanderen	Nederland	som	
hoogleraar	3,8	6,1	9,9	3,7	1,8	5,5	15,4
wetenschappelijk personeel	27,0	36,5	63,5	10,8	23,2	34,0	97,5
AIO/OIO	0	36,0	36,0	4,0	7,0	11,0	47,0
technisch medewerker	0	6,9	6,9	0,5	2,7	3,2	10,1
studenten	4,0	36,0	40,0	3,0	7,0	10,0	50,0
Totaal (excl. studenten)	30,8	85,5	116,3	19,0	34,7	53,7	170,0
	<b>18%</b>	<b>50%</b>	<b>68%</b>	<b>11%</b>	<b>20%</b>	<b>32%</b>	<b>100%</b>

Het door M&I Partners genoemd aantal studenten lijkt erg laag te zijn. Het aantal is echter bepaald op basis van het aantal studenten dat een voltijds TST-opleiding volgt inclusief een stage en een afstudeeropdracht. Uit een snelle rondvraag blijkt dat:

1. de aantallen aan het stijgen zijn: In Vlaanderen is het aantal ruim vervijfvoudigd ten opzichte van de cijfers in de Technologieverkenning. In Antwerpen is namelijk dit jaar het aantal studenten Computertaalkunde 10 en in Leuven zijn 6 studenten ingeschreven voor de specialisatie Computerlinguïstiek in de Master taalkunde. In Leuven zijn verder dit jaar 16 studenten ingeschreven in het Masterprogramma *Artificial Intelligence*.
2. er is een grotere groep studenten die één of meerdere TST-vakken volgt. Deze volgen de colleges in het kader van zeer verschillende hoofdstudies, bijvoorbeeld Taalkunde, Fonetiek, Informatica, Natuurkunde, Wiskundige Logica, Bedrijfskunde etc. In bijvoorbeeld Twente is vorig jaar een Mastercollege gegeven met 64 studenten en in het jaar daarvoor werd het college Taal bijgewoond door 60 studenten en het college Spraak door 84 studenten. Maar ook zijn er bijvoorbeeld in Gent jaarlijks een 30-tal studenten die het vak Spraakverwerking volgen. en die op die wijze kennis opdoen over taal- en spraaktechnologie.

Daarnaast is het zo dat momenteel steeds meer TST-onderzoek wordt uitgevoerd door elders in de wereld opgeleide onderzoekers, wat in het kader van de internationale inbedding van het onderzoek een pluspunt is. Stimulering van het onderzoek zal overigens ongetwijfeld als resultaat hebben dat nog meer Vlaamse en Nederlandse studenten geïnteresseerd raken.

In de Technologieverkenning werd berekend dat de totale omvang van het onderzoek net boven € 10 miljoen per jaar ligt, waarvan € 3,4 miljoen in Vlaanderen en € 6,7 in Nederland. Deze bedragen zijn gebaseerd op kostprijs per jaar, zoals dit door NWO wordt gehanteerd. De omvang van het taalonderzoek bedraagt € 6,8 miljoen en die van spraak € 3,3 miljoen per jaar. Indien TNO buiten beschouwing gelaten wordt (andere wijze van financiering), dan worden de onderzoekers voor 35% gefinancierd uit de 1<sup>e</sup> geldstroom, voor 45% uit de 2<sup>e</sup> geldstroom en voor 20% uit de 3<sup>e</sup> geldstroom. Wat betreft de 2<sup>e</sup> geldstroom is NWO de belangrijkste financier. Daarnaast worden SenterNovem (IOP-instrument) en KNAW genoemd.

*Het financiële volume<sup>38</sup> (in euro per jaar) dat hier naar schatting mee is gemoeid, is als volgt:*

<b>Financiële omvang TST</b> (in miljoen euro/jaar)	Vlaanderen taal	Nederland taal	som taal	Vlaanderen spraak	Nederland spraak	som spraak
hoogleraar	0,5	0,8	1,3	0,5	0,2	0,7
wetenschappelijk personeel	1,6	2,1	3,7	0,6	1,4	2,0
AIO/OIO	0,0	1,4	1,4	0,2	0,3	0,4
technisch medewerker	0,0	0,4	0,4	0,03	0,1	0,2
<b>Totaal (excl. studenten)</b>	<b>2,1</b>	<b>4,7</b>	<b>6,8</b>	<b>1,3</b>	<b>2,0</b>	<b>3,3</b>

In de Technologieverkenning werd ook gekeken naar de omvang van de verschillende taal- en spraaktechnologische groepen in Nederland en Vlaanderen. Hun gegevens zijn in de onderstaande tabel weergegeven. Nadere beschouwing leert dat in onderstaande getallen ook puur taalkundige activiteiten zoals Nederlands taalkunde, kindertaalverwerving, fonetiek e.d. lijken te zijn meegeteld.


*TST kennisinstellingen en hun omvang in fte<sup>39</sup> (overgenomen uit Rapport M&I/Partners). Een nadere specificatie van de specialiteiten van deze instellingen is te vinden in het rapport.*

	<b>Taal</b>	<b>Spraak</b>	<b>Totaal</b>
Universiteit Antwerpen (CNTS)	23,0		23,0
Universiteit van Amsterdam (UvA, ILLC)	25,0		25,0
K.U. Leuven (ESAT)		13,6	13,6
K.U. Leuven (Letteren)	7,8	0,2	8,0
KU Nijmegen (KUN, Letteren)	7,0	16,0	23,0
RU Groningen (Letteren)	14,0		14,0
Universiteit Gent (ELIS)		5,1	5,1
Universiteit van Tilburg (UvT, Letteren)	19,4	2,5	21,9
TNO Telecom	1,0	6,1	7,1
TNO Technische Menskunde		5,5	5,5
TNO TPD	4,4		4,4
Universiteit Twente	9,8	1,8	11,6
Universiteit Utrecht	4,9	2,8	7,7
<b>Totaal</b>	<b>116,3</b>	<b>53,6</b>	<b>169,9</b>

<sup>38</sup> Cijfers overgenomen uit de Technologieverkenning Nederlandstalige Taal- en Spraaktechnologie (M&I/Partners BV en Montemore NV). Het financieel volume is niet exact, maar afgeleid op basis van aantal en type fte's.

<sup>39</sup> Cijfers eveneens overgenomen uit de genoemde Technologieverkenning: een aantal kleine groepen ontbreekt, maar zij vormen naar inschatting minder dan 10% van het onderzoeksvolume.

## Bijlage 6: Persoonlijke betrokkenheid bij aanvragen van leden van adviescolleges

<b>bijlage 1</b>				Ministerie van Economische Zaken	
Aan					
Drs H.F.G. Geijzers					
Directeur Senter					
Postbus 30732					
2500 GS DEN HAAG					
Datum					
<b>21 DEC. 1998</b>		Uw kenmerk		Ons kenmerk	
		DDB85026.MDR		I&D/ABC/R&O9885526	
		DDB85027.MDR		0	
Bijlage(n)					
Onderwerp					
Persoonlijke betrokkenheid van leden adviescolleges bij aanvragen					
Geachte heer Geijzers,					
<p>De beantwoording van uw brieven van 13 maart jl. (kenmerk DDB85026.MDR en DDB85027.MDR) heeft helaas enige vertraging opgelopen. Daarvoor mijn excuses. In deze brieven geeft u de reacties weer van de verschillende adviescollege's op de uitleg van de uitspraak van de Raad van State in de zaak Aorta/Novem, zoals verwoord in de brief van de heer Idenburg van 21 oktober 1997 (kenmerk ID/ABC/RO/97051147) en in de brief van de heer Vijlbrief van 20 november 1997 (kenmerk ES/ATB/KI/97072115). Daarnaast vraagt u mijn akkoord om adviescommissies te instrueren over de consequenties van de genoemde uitspraak van de Raad van State voor hun werkwijze.</p>					
<b>Instructie</b>					
Wat betreft uw voorgenomen instructie aan de adviescommissies merk ik allereerst op dat de bestuurlijke verhoudingen strikt genomen niet zó liggen, dat u een instructie kunt geven aan deze adviescommissies. De adviescommissies geven de minister een advies over besluiten op subsidie-aanvragen. Ingevolge artikel 2:4, tweede lid, van de Algemene wet bestuursrecht waakt het bestuursorgaan ertegen dat tot het bestuursorgaan behorende personen die een persoonlijk belang bij een besluit hebben, de besluitvorming beïnvloeden. Ook adviescommissies zijn bestuursorganen. Zij stellen hun eigen werkwijze vast.					
U hebt mandaat om namens de minister de besluiten te nemen waarover de adviescommissies adviseren. Daarbij rust op u de taak om te beoordelen of u een advies zult opvolgen of niet. In dat kader dient u zich ervan te vergewissen of het advies op zorgvuldige wijze tot stand is gekomen. Het niet-beïnvloed zijn van de besluitvorming					
Bezoekadres		Doorkiesnummer		Telefax	
Bezuidenhoutseweg 20		(070) 379 6160		(070) 379 6529	
Hoofdkantoor		Telefoon (070) 379 89 11		X.400 adres S = EZPOST/C = NLJA = 400NET/JP = MIN EZ	
Bezuidenhoutseweg 30		Telefax (070) 347 40 81		Internetadres ezpost@minvz.nl	
Postbus 20101		Telex 31099 ecza nl			
3500 EP DEN HAAG		X.400			



van een adviescommissie door leden die een persoonlijk belang bij een besluit hebben is een belangrijk aspect van die zorgvuldigheid.

#### **Normen**

In dat kader kan ik mij voorstellen dat u de adviescommissies ervan op de hoogte stelt welke normen u op dit gebied hanteert. Het is echter de vraag of deze materie zich leent voor een gedetailleerde nadere normstelling. In de brief van dr Idenburg van 21 oktober 1997 is opgemerkt dat "...belang bij het te nemen besluit onder omstandigheden bijvoorbeeld kan worden afgeleid uit een betrokkenheid bij de aanvraag of bij een aanvrager...". Daarbij is gesteld dat ons inziens nauwe betrokkenheid bij een aanvraag enger is dan nauwe betrokkenheid bij een aanvrager, maar dat het de voorkeur verdient ook de nodige zorgvuldigheid in acht te nemen indien sprake is van nauwe betrokkenheid bij een aanvrager. In feite geven wij daarmee nadere invulling aan een situatie die niet ter sprake is geweest in de zaak Aorta/Novem, doch die evenzeer gebaat is bij de gewenste zorgvuldigheid bij de advisering.

Genoemde brief had niet de pretentie een alomvattend stelsel van nadere normen te geven. Dat is niet mogelijk. De problematiek van betrokkenheid bij een besluit is namelijk niet beperkt tot betrokkenheid bij een aanvraag of bij een aanvrager. Een lid van een adviescollege kan ook getrouwd zijn met iemand die betrokken is bij een aanvrager of bevriend zijn met de opsteller van een aanvraag. Waar hier de grenzen liggen is niet in zijn algemeenheid te zeggen. Uiteindelijk komt het steeds neer op de vraag of er bij degene die invloed heeft op een besluit van een bestuursorgaan belangen een rol kunnen spelen die niet behoren tot de belangen dat het bestuursorgaan uit hoofde van de hem opgedragen taak behoort te vervullen.

#### **Adviescolleges op de hoogte stellen**

Uit het voorgaande volgt dat u de adviescommissies geen instructie kunt geven. Ik acht het echter wel zinvol om de adviescommissies op de hoogte te stellen van de normen die u hanteert bij de beoordeling of een advies op zorgvuldige wijze tot stand is gekomen. U dient daarbij aan te geven dat het niet alleen gaat om activiteiten van een commissielid, maar ook om andere feiten, omstandigheden en hoedanigheden die resulteren in belangen, en ook niet alleen om betrokkenheid bij een aanvraag of een aanvrager, maar om betrokkenheid bij een besluit.

Het lijkt mij zinvol om het algemene uitgangspunt, namelijk dat u bepaalt of een advies op zorgvuldige wijze tot stand is gekomen zonder (de schijn van) belangen verstrengeling, iets verder te concretiseren. Maar daarbij dient nadrukkelijk te worden aangegeven dat dit geen alomvattend normenstelsel is.

De lijn die u hanteert in de begeleidende brief bij de door u voorgestelde instructie zou echter tot een striktere interpretatie van de door u gebruikte aanduiding 'te nauwe betrokkenheid' kunnen leiden dan gezien de uitspraak van de Raad van State



noodzakelijk is. U wenst voor adviescommissies bij tenderregelingen dezelfde consequenties te verbinden aan betrokkenheid bij een aanvraag als bij een aanvrager, namelijk categorische uitsluiting van het commissielid bij discussie, beoordeling en rangschikking van alle aanvragen in de gehele tender. Zoals reeds eerder gezegd kan de mate van betrokkenheid bij een aanvrager vele gradaties aannemen, maar zoals gebleken is uit de reacties, zou een strikte interpretatie van betrokkenheid bij sommige adviescommissies tot onwerkbaar situaties kunnen leiden omdat bij die regelingen veelal experts worden benaderd die vaak zelf in enigerlei relatie tot de subsidie-aanvrager staan. In de brief van de heer Idenburg is aangegeven, welke typen relaties in het bijzonder tot zorgvuldigheid nopen.

Algemeen geldt dat, indien naar uw oordeel een ongewenste betrokkenheid van een lid van een adviescommissie bij een aanvraag bestaat, u kunt besluiten dat het betrokken commissielid niet deelneemt aan de discussie, beoordeling en rangschikking van de betreffende aanvraag, dan wel van alle aanvragen in de betreffende tender. Eén en ander wordt vastgelegd in het verslag van de commissievergadering

Mijns inziens zou u de adviescommissies uit hoofde van deze bevoegdheid kunnen verzoeken in ieder geval onderstaande werkwijze te hanteren teneinde het Senter mogelijk te maken om te beoordelen of het advies op zorgvuldige wijze tot stand is gekomen.

#### *Verslaglegging betrokkenheid bij alle regelingen*

Bij elke vergadering van een adviescommissie wordt de mogelijke persoonlijke betrokkenheid op de agenda geplaatst, wordt deze zo nodig besproken en van het besprokene wordt verslag gedaan.

#### *Niet-tenderregelingen*

Een commissielid dat te nauw betrokken is bij een aanvraag of bij een aanvrager in een niet-tenderregeling neemt geen deel aan de discussie over en de beoordeling van de desbetreffende aanvraag.

#### *Tenderregelingen*

- Een commissielid dat te nauw betrokken is bij een *aanvraag* in een tenderregeling neemt geen deel aan de discussie over, de beoordeling en rangschikking van alle aanvragen in de gehele tender.
- Een commissielid dat te nauw betrokken is bij een *aanvrager* in een tenderregeling neemt geen deel aan de discussie over, beoordeling en rangschikking van de desbetreffende aanvraag. Het commissielid wordt daarmee dus in principe niet uitgesloten van beoordeling van de andere aanvragen in de tender tenzij naar het eigen oordeel van het commissielid en/of naar het oordeel van de overige commissieleden sprake is van een zodanig betrokkenheid bij de aanvrager dat het



betreffende commissielid ook uitgesloten dient te worden van beoordeling van de andere aanvragen in de tender.

*De IOP-regeling*

De IOP-regeling is in deze een geval apart. De Stuurgroep IOP, welke wordt voorgezeten door de directeur Algemeen Technologiebeleid van EZ, adviseert op grond van artikel 8 van de Subsidieregeling IOP de minister over subsidieaanvragen. Zij laat zich daarbij zelf echter weer adviseren door de commissies die de uitvoering van de onderscheiden onderzoeksprogramma's begeleiden: de Programmacommissies. Deze Programmacommissies bestaan uit vertegenwoordigers van kennisinstellingen en bedrijfsleven. De innovatie gerichte onderzoeksprogramma's spelen zich in het algemeen op voor Nederland overzichtelijke onderzoeksvelden af, met beperkte aantallen spelers zowel aan vraag- als aanbodzijde. Hierdoor zal het in praktijk niet mogelijk zijn elke betrokkenheid van met name de leden van Programmacommissies bij aanvragen en aanvragers uit te sluiten. In de IOP-praktijk wordt wel algemeen de regel gevolgd dat leden van de Programmacommissie niet mogen oordelen over aanvragen waarbij zij zelf een direct belang hebben.

Door een strikte hantering van bovengenoemde norm voor tenderregelingen door Senter zou de IOP-regeling echter in feite onuitvoerbaar worden.

De Algemene wet bestuursrecht voorziet dat het niet in alle gevallen mogelijk zal zijn vermenging van belangen te voorkomen. In die gevallen moet naar mogelijkheden worden gezocht om persoonlijke invloed te beperken.

Van groot belang is in deze de controleerbaarheid van de gevolgde beoordelingsprocedure.

Aan de Programmacommissies zou u daarom kunnen verzoeken bij discussie over en beoordeling en ranking van projectvoorstellen c.q. subsidieaanvragen zelf af te wegen of de leden een te nauwe betrokkenheid hebben bij aanvragen of aanvragers om tot een objectieve beoordeling te kunnen komen en hiervan expliciet verslag te doen.

De Stuurgroep IOP zal zich op deze basis een oordeel kunnen vormen over de afwegingen van de Programmacommissies. Het verdient aanbeveling in het verslag van de Stuurgroep-vergadering aandacht te besteden aan de betrokkenheid van de programmacommissies en wat de stuurgroep daaraan heeft gedaan. Vervolgens behoort deze afwegingsregel ook op de Stuurgroep IOP zelf van toepassing te zijn. Ook hieraan dient in het verslag aandacht te worden besteed.

Aangezien bij de uitvoering van de IOP-regeling Senter aanwezig is bij zowel de Stuurgroepvergaderingen als bij de vergaderingen van de Programmacommissies, zal het voor Senter steeds goed mogelijk zijn zich een oordeel te vormen over de gevolgde procedures.



**Tot slot**

Ik adviseer u met enige regelmaat binnen uw organisatie overleg te voeren over deze problematiek ten einde een eenduidige lijn van beoordeling van de gegeven adviezen te kunnen vast te houden.

Ik ga er, samen met mijn collega's van het directoraat-generaal voor Economische Structuur en het directoraat-generaal voor Energie, vanuit dat de adviescommissies de redelijkheid van de hierboven geadviseerde werkwijze zullen inzien.

J. van Sinderen  
Directeur Algemene Beleidscoördinatie

## **Bijlage 7: Overzicht taken en verantwoordelijkheden Programmabestuur, Programmacommissie, begeleidingscommissie en programmabureau**

### **Programmabestuur**

- het uitvoeren van de globale supervisie van het onderzoeksprogramma, het bewaken van de voortgang, het toezicht houden op de werkzaamheden van de andere actoren in de organisatiestructuur (d.w.z. de Programmacommissie, het programmabureau en de Nederlandse Taalunie, en als bemiddelaar optreden bij eventuele geschillen;
- het beoordelen van het meerjarenprogramma en de jaarwerkplannen opgesteld door de Programmacommissie;
- het selecteren de projecten die in het kader van elke oproep worden gefinancierd op basis van het advies geformuleerd door de Programmacommissie en het voorleggen van deze stukken voor definitieve goedkeuring aan de financierende instanties die zij vertegenwoordigen. Wat Vlaanderen betreft betekent dat concreet dat de vermelde stukken ter goedkeuring worden voorgelegd aan de Vlaamse minister bevoegd voor wetenschappen en technologische innovatie. In Nederland dienen de financierende instanties regelmatig op de hoogte gehouden te worden van de beslissingen van het Programmabestuur;
- het controleren van de Programmacommissie op objectieve en zorgvuldige beoordeling en een goede uitvoering van het programma in het algemeen en in het bijzonder het bewaken van het programmatische karakter van het programma. In dat laatste kader past o.a. een goede verdeling van financiële middelen over de gehele looptijd van het programma, over de verschillende typen projecten en over de betrokken partijen.

### **Programmacommissie**

- het bij aanvang van het programma definiëren van succesfactoren op basis waarvan het programma kan worden geëvalueerd. Daarbij past ook een nulmeting, die de huidige stand aangeeft op deze succesfactoren;
- het vertalen van het meerjarenprogramma in jaarwerkplannen, waarin de concrete onderzoeklijnen en de in te zetten subsidie-instrumenten worden gespecificeerd;
- het uitschrijven van oproepen tot het indienen van projectvoorstellen met daarin de criteria voor de beoordeling van de projectvoorstellen (Call for proposals / Call for tender);
- het formuleren van een advies aangaande de beoordeling en prioritering van de projectvoorstellen;
- het formuleren van voorstellen voor allocatie van financieringsmiddelen;
- het bewaken van de relevantie van het in uitvoering genomen onderzoek en de controle op de uitvoering van toegewezen projecten;
- actieve betrokkenheid bij de opzet van plannen en acties voor kennisoverdracht;
- het opstellen van jaarrapportages en werkplannen en het uitvoeren van de nulmeting en de geplande evaluaties.

### **Begeleidingscommissie**

- het actief begeleiden van een of meer projecten, doet bijvoorbeeld suggesties voor verder onderzoek en helpt bij het maken van keuzes. Ze kan wetenschappelijke input geven en/of kan uitleg geven over de industriële situatie en de situatie in het bedrijf in het bijzonder;
- het bewaken van de doelen van een of meer projecten, en de besteding van tijd en (financiële) middelen. Hierbij wordt gelet op de voortgang van het project, de eventuele noodzaak van wijzigingen ten opzichte van het projectplan en onderbouwing van eventuele wijzigingen;
- het beoordelen of er aspecten octrooieerbaar zijn;
- heeft een belangrijke taak in de kennisoverdracht tussen bedrijven en kennisinstellingen;
- helpt mee resultaten te verspreiden onder een breder publiek;
- stimuleert dat vervolgonderzoek zal plaatsvinden.

## Programmabureau

- het organiseren en uitvoeren van subsidierondes;
- het uitvoeren van de voortgangscontrole;
- het zorgdragen voor een efficiënt en doelmatig beheer van de projecten en middelen;
- het in opdracht van het Programmabestuur financieel afhandelen en begeleiden van het programma;
- het zorgdragen voor de afstemming tussen en binnen projecten;
- het opstellen van jaarwerkplannen en bestedingsplannen, begrotingen, jaarrekeningen en jaarverslagen;
- het leggen en onderhouden van contacten met het veld van interne en externe (internationale) relaties;
- het organiseren van *brokerages*, workshops, symposia, congressen en andere publicitaire acties;
- het voeren van het secretariaat van de Programmacommissie en het Programmabestuur.

## Bijlage 8: Verklaring gebruikte afkortingen

ALVV	Adviescommissie Lexicografische Vertaalvoorzieningen, voorheen CLVV
AWI	Vlaamse administratie Wetenschap en Innovatie - <a href="http://innovatie.vlaanderen.be/">http://innovatie.vlaanderen.be/</a>
BaTaVo	BAisTAalVOorziening
BSIK	Besluit subsidies investeringen kennisinfrastructuur (voorheen ICES/KIS) - <a href="http://www.senter.nl/iceskis/">http://www.senter.nl/iceskis/</a>
BTS	Bedrijfsgerichte Technologische Samenwerkingsprojecten - <a href="http://www.senter.nl/">http://www.senter.nl/</a>
CALL	Computer Assisted Language Learning
CGN	Corpus Gesproken Nederlands, Vlaams-Nederland project - <a href="http://lands.let.kun.nl/cgn/">http://lands.let.kun.nl/cgn/</a>
CIC	Concurreren met ICT Competenties (EZ-OCW programma) - <a href="http://www.cic-online.nl/">http://www.cic-online.nl/</a>
CLIF	Computer Linguistics in Flanders - <a href="http://clif.uia.ac.be/">http://clif.uia.ac.be/</a>
DIS	Dynamisch Innovatiesysteem
EAGLES	Expert Advisory Group on Language Engineering Standards - <a href="http://www.hltcentral.org/projects/EAGLES/">http://www.hltcentral.org/projects/EAGLES/</a>
ELDA	Evaluations and Language resources Distribution Agency European - <a href="http://www.elda.fr/">http://www.elda.fr/</a>
ELRA	European Language Resources Association - <a href="http://www.elra.fr/">http://www.elra.fr/</a>
ELSNET	European Network of Excellence in Speech and Language Technology - <a href="http://www.elsnet.org/">http://www.elsnet.org/</a>
ENABLER	European National Activities for Basic Language Resources Network - <a href="http://www.enabler-network.org/">http://www.enabler-network.org/</a>
EUROMAP	KP5 IST-HLT project EUROMAP - <a href="http://www.hltcentral.org/htmlengine.shtml?id=56">http://www.hltcentral.org/htmlengine.shtml?id=56</a>
EZ	Ministerie van Economische Zaken - <a href="http://www.minez.nl/">http://www.minez.nl/</a>
FENIT	Federatie van Nederlandse ondernemingen in de Informatietechnologie - <a href="http://www.fenit.nl/">http://www.fenit.nl/</a>
FWO	Fonds voor Wetenschappelijk Onderzoek - <a href="http://sun.fwo.be/">http://sun.fwo.be/</a>
HLT	Human Language Technology
HMM	Hidden Markov Modelling
ICES/KIS	Zie Bsik
ICT	Informatie- & Communicatie Technologie
IMIX	Interactieve Multimodale Informatie eXtractie
INL	Instituut voor Nederlandse Lexicologie - <a href="http://www.inl.nl/">http://www.inl.nl/</a>
IOP	Innovatiegerichte Onderzoeksprogramma's - <a href="http://www.senter.nl/iop/">http://www.senter.nl/iop/</a>
IPR	Intellectual Property Rights (Intellectuele Eigendomsrechten)
ISCA	International Speech Communication Association - <a href="http://www.isca-speech.org/">http://www.isca-speech.org/</a>
ISLE	International Standards in Language Engineering
ISO	International Organization for Standardization - <a href="http://www.iso.org/">http://www.iso.org/</a>
IST	Information Society Technologies
IVR	Interactive Voice Response
IWT	Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen - <a href="http://www.iwt.be/">http://www.iwt.be/</a>
KMO	Kleine en Middelgrote Ondernemingen (VL, zie ook MKB)
KP	Europees KaderProgramma - <a href="http://fp6.cordis.lu/fp6/home.cfm">http://fp6.cordis.lu/fp6/home.cfm</a>
LANGNET	ERA-Net in Language Technologies
LANGTECH	European Forum for Speech and Language Technology - <a href="http://www.lang-tech.org/">http://www.lang-tech.org/</a>
LDC	Linguistic Data Consortium - <a href="http://www ldc.upenn.edu/">http://www ldc.upenn.edu/</a>
LISA	Localising Industry Standards Association - <a href="http://www.lisa.org/">http://www.lisa.org/</a>
MKB	Middelgrote en Kleine Bedrijven (NL, zie ook KMO)
MLIS	Multilingual Information Society
MMI	Mens-Machine Interactie
MVG	Ministerie van de Vlaamse Gemeenschap
NOTaS	De Nederlandse Organisatie voor Taal- en Spraaktechnologie - <a href="http://www.stichtingnotas.nl/">http://www.stichtingnotas.nl/</a>

NWO	Nederlandse Organisatie voor Wetenschappelijk Onderzoek - <a href="http://www.nwo.nl/">http://www.nwo.nl/</a>
OCW	Ministerie van Onderwijs, Cultuur & Wetenschap - <a href="http://www.minocw.nl/">http://www.minocw.nl/</a>
PDA	personal digital assistant
SBO	Strategisch Basis Onderzoek (subsidieinstrument IWT-Vlaanderen) - <a href="http://www.iwt.be/">http://www.iwt.be/</a>
STEVIN	Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands - <a href="http://taalunieversum.org/taal/technologie/stevin/">http://taalunieversum.org/taal/technologie/stevin/</a> .
TNO	Nederlandse Organisatie voor Toegepast Onderzoek - <a href="http://www.tno.nl/">http://www.tno.nl/</a>
TST	taal- en spraaktechnologie
VNC	NWO/FWO Programma voor Vlaams Nederlandse Samenwerking - <a href="http://www.nwo.nl/vnc/">http://www.nwo.nl/vnc/</a>