

ru | STEVIN

Dutch-Flemish research programme for
Dutch Language and Speech Technology

STEVIN final evaluation Fact File



February 2010

STEVIN Fact File

Contents

Title and duration programme	STEVIN: Flemish-Dutch research programme for Dutch Language and Speech Technology (STEVIN: Spraak- en taaltechnologische voorzieningen voor het Nederlands) – 2004 - 2011	
Summary	Summary of the objectives of the DUTCH-Flemish STEVIN Programme	page 3
STEVIN organisation	<ul style="list-style-type: none"> • HLT-Board • STEVIN Programme Committee • International Assessment Panel • STEVIN Programme bureau and some advisory groups 	page 4
STEVIN Budget and Funding Organisations	<p>STEVIN budget = 11.4 M€: Flanders 3.8 M€, the Netherlands 7.6 M€</p> <ul style="list-style-type: none"> • The Netherlands: Ministry of Education, Culture and Science, Netherlands Organisation for Scientific Research, Ministry for Economic Affairs • Flanders: EWI, Department of Economics, Science and Innovation of the Flemish Government 	page 5
STEVIN Calls	STEVIN Funding instruments – assessment procedures and statistics Distribution funding over Dutch-Flemish/academic-industrial recipients	page 6
STEVIN Funding Instruments (and their max. budget) 2003-2009	<ul style="list-style-type: none"> • 1st Call for Proposals for strategic research proposals and HLT resources (data & tools) (max. budget 2 M€) • 2nd Call for Proposals for strategic research proposals and HLT resources (data & tools) (max. budget 3,8 M€) • Three Calls for tender for specific HLT resources (max budget 1,6 M€) • Call for proposals for applied research (max. budget 2,3 M€) • Three Calls for proposals for demonstration projects (max. budget 1 M€) • Three Calls for educational project master classes (max. budget 110 k€) 	page 8
STEVIN R&D priorities	Overview STEVIN R&D priorities (as were already included in the original STEVIN programme description) and how they are covered by the STEVIN projects funded in the different funding schemes.	page 13
STEVIN assessment criteria	Standard STEVIN R&D assessment criteria (as were already included in the original STEVIN programme description) and code of conduct for the STEVIN IAP and STEVIN PC	page 15
STEVIN project details	Overview of STEVIN projects, including details of STEVIN priorities covered and consortium partners, budget, duration and project summary.	page 18
STEVIN project status	Overview status of STEVIN projects	Page 64
STEVIN IPR	STEVIN IPR and standards policy, schematic overview HLT actors, leaflet for data providers	page 66
Publication list	Scientific outputs of STEVIN programme in international literature	page 69
HLT activities	List of HLT activities organised by or financially supported by STEVIN	page 77



Introduction – Summary STEVIN Objectives

Dutch is ranked as the 40th most widely spoken language of the world's 6,000 languages. Most of the 22 million Dutch native speakers live in the Netherlands and the Flemish region of Belgium. Nevertheless the market for human language technology for Dutch (HLTD) is too limited to attract important investments by industry in HLTD. Therefore, cross border cooperation among governments, businesses and academia has been established, resulting in a Flemish/Dutch HLTD research programme. The programme is called STEVIN, which is a Dutch acronym for 'Essential Speech and Language Technology Resources for Dutch'.

The STEVIN programme for Dutch language and speech technology is a coordinated effort of:

- o the Dutch Language Union (NTU);
- o the Flemish Government department Economy, Science and Innovation (EWI);
- o the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT-Vlaanderen);
- o the Flemish National Fund for Scientific Research (FWO-Vlaanderen);
- o the Dutch Ministry of Education, Culture and Science (OCW);
- o the Dutch Ministry of Economic Affairs (EZ);
- o the Dutch Ministry for Economic Affairs agency for innovation and sustainable development (Agentschap NL, formerly SenterNovem);
- o the Netherlands Organisation for Scientific Research (NWO);
- o the NWO-division for the Physical Sciences (NWO-EW);
- o the NWO- division for the Humanities (NWO-GW).

This six-year programme aims to contribute to the further progress of HLTD in Flanders and the Netherlands and stimulate innovation in this sector. In addition, it will strengthen the economic and cultural position of the Dutch language in the modern ICT-based society.

The STEVIN programme was launched in 2005. It is jointly financed by the Flemish (Department of Economy, Science and Innovation) and Dutch governments (Ministry of Education, Culture and Science, Ministry of Economic Affairs and the Netherlands Organisation for Scientific Research).

STEVIN will run until 2011 with a total budget of 11.4 million euros. STEVIN is coordinated by the Dutch Language Union and supervised by a board of representatives of the funding bodies. A programme committee, including both academic and industrial representatives, is responsible for scientific and content related issues. An international assessment panel of highly-respected HLT-experts evaluates the submitted R&D proposals. A programme office, a joint collaboration of the Netherlands Organisation for Scientific Research and the Dutch innovation agency Agentschap NL (formerly SenterNovem), takes care of the operational matters.

The STEVIN goals are defined in the framework of a stratified innovation system. Each actor of the system needs to be addressed, resulting in a mix of funding instruments and programme activities.

- funding for collaboration between universities and between universities and industry to realise an adequate basic language resources kit
- funding strategic R&D on HLTD
- funding demonstration projects illustrating the feasibility and value of HLTD applications
- funding networking activities to stimulate HLTD knowledge transfer
- organising networking activities and HLTD promotional events

To enable the use and re-use of STEVIN results, a specific IPR-arrangement has been set up. The materials (software, data etc.) must be handed over to the Dutch Language Union so they can be made available to third parties through the Dutch HLT Agency ('TST Centrale' www.tst.inl.nl). The Dutch HLT Agency helps resolve IPR issues, is responsible for the management, maintenance and distribution of HLTD materials, and also acts as a service desk. The STEVIN IPR and standards policy is described in detail in one of the annexes to this fact file. More information on the Flemish/Dutch STEVIN programme and the STEVIN projects can be found on: www.stevin-tst.org.

STEVIN organisation – Tasks and Responsibilities

Programme Board (composition as of January 2007)

The Board of the STEVIN programme is responsible for the final funding decisions. The Board is also responsible for supervising two activities that are related to the STEVIN programme, i.e. the *TST-centrale* and the *Makel en Schakel* activities carried out by the *Nederlandse Taalunie*.

- o the general secretary of the Nederlandse Taalunie - chair;
- o representatives of the partners financing the STEVIN programme:
 - * the Flemish Government department Economy, Science and Innovation (EWI)
 - * the Institute for the Promotion of Innovation by Science and Technology in Flanders
 - * the National Fund for Scientific Research (Belgium) (FWO)
 - * the Dutch Ministry of Education, Culture and Science (OCW)
 - * the Dutch Ministry of Economic Affairs (EZ)
 - * the Netherlands Organisation for Scientific Research (NWO, NWO-GW, NWO-EW))
- o two senior language and speech technology experts: prof. Dirk Van Compernelle, Leuven University and prof. John Nerbonne, Groningen University.

Programme Committee (composition as of July 2009)

- o Prof. dr. Jan Odijk (Utrecht University (formerly also Nuance)) – chair
- o Prof. dr. Jean-Piere Martens (Gent University)
- o Prof. dr. Frank van Eynde (Leuven University)
- o Prof. dr. Walter Daelemans (Antwerpen University)
- o Dr. Arjan van Hessen (Telecats BV / Twente University)
- o Prof. dr. Louis Boves (Radboud University Nijmegen)
- o Drs. Remco van Veenendaal (INL / TST centrale)
- o Dhr. Jan van Sas (The LingWareHouse / Karel de Grote-hogeschool Antwerpen)
- o Dr. ir. Kris Van Bruwaene (VRT)
- o Dr. ir. Ruud Smeulders (Rabobank Groep ICT, IBA)
- o Dr. Leonoor van der Beek (Q-go Amsterdam).

STEVIN International Assessment Panel: assessment and ranking (composition per 1/1/2007)

- o Prof. dr. Hans Uszkoreit (DFKI - Germany)
- o Prof. dr. Gábor Prószéky (Morphologic - Hungary)
- o Prof. dr. Roger Moore (Sheffield University – UK)
- o Dhr. Paul Heisterkamp (DaimlerChrysler - Germany)
- o Dr. Gilles Adda (LIMSI - France);
- o Dr. Nicoletta Calzolari (ILC - Italy)
- o Dr. Stelios Piperidis (ILSP - Greece)
- o Prof. dr. Anne Abeillé (Université Paris 7 – France).

STEVIN Programme Office (Agentschap NL/NWO) and STEVIN Coordinating office (NTU)

Together the Dutch organisations NWO and Agentschap NL have been selected by the *Nederlandse Taalunie* to form the STEVIN Programme Bureau that coordinates the STEVIN activities, including the handling and selection process of applications from both Dutch and Flemish applicants. NTU was responsible for the overall coordination of the STEVIN programme and for coordinating its activities with related HLT activities carried out under the auspices of NTU (HLT Agency and HLT PR activities).

- o NWO Alice Dijkstra, Brigit van der Pas
- o Agentschap NL Dieneke Meijer
- o Nederlandse Taalunie dr. Peter Spyns, Elisabeth D'Halleweyn

Furthermore some advisory groups have been set up:

- o Working Group for STEVIN supporting activities (which includes representatives from non-STEVIN programmes and projects), to coordinate HLT supporting activities in the Netherlands and Flanders.
- o IPR Working Group (led by the Dutch Language Union, includes academic and industrial HLT experts on IPR and legal experts), to co-ordinate and optimize STEVIN IPR practices.

Summary Dutch-Flemish STEVIN Programme budget

Of the total STEVIN budget 1/3rd is funded by the Flemish government and 2/3rd is funded by a consortium of Dutch ministries and funding organisations.

Funding by Dutch and Flemish government and funding organisations		
Flanders	€ 3.800.000	*
the Netherlands	€ 7.600.100	**
interest 2.5%	€ 262.304	
	€ 11.662.404	
<p>* Dutch funding provided jointly by the Ministry of Education, Culture and Science, the Netherlands Organisation for Scientific Research (GW, EW, AB) and the Ministry for Economic Affairs</p> <p>** Flemish funding provided by the Department of Economics, Science and Innovation (EWI) of the Flemish Government</p>		
Budget STEVIN funding schemes, supporting activities and management		
R&D projects	€ 8.906.716	76,37%
Demonstration projects	€ 996.044	8,54%
Supporting activities	€ 688.380	5,90%
Dutch HLT Agency	€ 300.000	2,57%
STEVIN management	€ 771.264	6,61%
	€ 11.662.404	

As can be seen in the table above: 76,4 % of the budget is spent on R&D projects (creating HLT resources, carrying out basic and application-oriented research) that were funded in one of the three open calls or in one of the three calls for tender. About 8.5% was spent for demonstration projects which may stimulate demand for HLT technology. Furthermore, 5,9% of the budget was allocated for the creation of networks, the consolidation of language and speech technology activities, educate new HLT experts and promote discussion and transfer of HLT knowledge. For making sure STEVIN results is maintained and supported and become widely available via the Dutch HLT Agency 2.6% of the total budget, or 3% of the R&D budget is reserved. For the management of the STEVIN programme 6.6% of the programme budget will be spent.

STEVIN Funding instruments – assessment procedures and statistics

STEVIN handling agencies

NWO has acted as main handling agency for the three open calls for strategic research projects and projects aiming at realizing part of the Dutch basic language resources kit and the calls for tender for i) A speech recognition toolkit for Dutch, ii) A lexical resource for the semantic processing of Dutch and iii) An annotated written Dutch corpus.

Agentschap NL is responsible for the administrative and financial management of the STEVIN projects.

Agentschap NL has also acted as main handling agency for the three calls for demonstration projects and the calls for proposals for educational and for networking activities.

STEVIN subsidieregeling

The legal rules applying for the specific Dutch-Flemish granting schemes are laid down in the *Subsidieregeling van de Nederlandse Taalunie tot subsidieverstrekking in het kader van Nederlandse taal- en spraaktechnologie "STEVIN" (STEVIN-subsidieregeling)*, which are available from the STEVIN website.

Selecting the best proposals & managing conflicts of interest

The Nederlandse Taalunie has formally assigned the task of STEVIN Programme Bureau jointly to NWO and Agentschap NL. NWO's primary responsibility is to organise the assessment procedure of the open calls and the calls for tenders. For each open call NWO has invited the International Assessment Panel to come to either Amsterdam or Brussels to assess and prioritize all proposals. Proposals were evaluated on the basis of a set of assessment criteria already laid down in the formal STEVIN project description and repeated in the formal call publications. For the calls for tender tender-specific criteria were added. The assessment meetings of the IAP and PC were attended by two representatives from NWO and one from Agentschap NL who gave special attention to safeguarding the fairness of the assessment procedure.

One of the main concerns of the Programme Bureau was to deal with conflicts of interest as especially members of the PC would have personal involvement in one of more applications. Considering the size and connectedness of the language and speech technology community in The Netherlands and Flanders, it is like in any other innovation-oriented research area, not realistic to exclude all involvement.

However a number of actions was taken to secure that only the best proposals were selected. One of the main actions taken was to have an international panel of experts fly in for the selection process.

Furthermore, modelled on the code of conduct used by the European Commission for its Framework Programme, a STEVIN Code of Conduct was formulated for the IAP and PC. All members were required to sign a declaration of conflict of interest and confidentiality and to formally indicate in which – if any – of the submitted project they had any involvement. In doing so, the members committed themselves to strict confidentiality and impartiality concerning their tasks. If a member has a direct or indirect link with the project(s), or any other vested interest, or is in some way connected with the project(s), or has any other allegiance which impairs or threatens to impair his/her impartiality with respect to the project(s), the STEVIN Programme Bureau has ensured that those members did not participate in the review and ranking of the project(s) concerned.

In a two-day meeting, all eligible applications were assessed and ranked by the IAP. The assessment reports were sent to the applicants for a response. Subsequently, on the basis of a) the IAP assessment and ranking, b) the applicants response and c) knowledge of the Dutch and Flemish HLT field, the PC added their remarks to the IAP assessment and also ranked the proposals. In doing so, it was possible to incorporate a Dutch-Flemish perspective in the assessment procedure, which could not be obtained from the international experts alone.

The final funding decision was made by the HLT Board -- consisting of representatives of the Flemish and Dutch government, along with two senior experts from the field -- on the basis of 1) the IAP assessment and ranking 2) the responses of the applicants, 3) the PC assessment and ranking including a description of way the assessments and ranking were arrived at in the meeting and an explanation for possible differences with the ranking given by the IAP.

STEVIN Funding instruments

A number of funding instruments (open calls and calls for tender) were implemented:

1. a 1st open call (September 2004) – maximum budget 2 M€ - for focussed short term (maximum duration is 2 years) strategic research projects and projects aiming at realizing part of the Dutch basic language resources kit a self-contained result;
2. a 2nd open call (in the spring of 2005) – maximum budget 3.4 M€ - for more complex strategic research projects and projects aiming at realizing part of the Dutch basic language resources kit with a longer time frame;
3. a 3rd open call (in 2007) – maximum budget 2.3 M€ - for application-oriented research projects;
4. three calls for tender:
 - A) a call for tender (in 2005) – maximum budget 800 k€ - for i) A speech recognition toolkit for Dutch and ii) A lexical resource for the semantic processing of Dutch
 - B) a call for tender (in 2007) – maximum budget 836 k€ - for An annotated written Dutch corpus;
5. three calls for demonstration projects (in 2005, 2006 and 2007) – maximum total budget for the three calls: 1 M€ - for small SME supporting projects stimulating HLT demand;
6. call for proposals for educational projects (2007-2009), maximum total budget 110 k€.
7. continuous call for networking activities, maximum total budget 50 k€.

More details about these calls are given in the next sections.

Summary statistics STEVIN funding R&D and demonstration projects (in k€)

In the table below an overview is presented of how STEVIN funding awarded to R&D projects and demonstration projects was distributed over:

- a) Dutch and Flemish partners: 64% - 36%
- b) Academic and industrial partners: 83% - 17%

<i>Distribution STEVIN R&D funding over Dutch-Flemish recipients</i>			
Netherlands			
Universities	k€ 4.970		
Industry	k€ 1.236		
Total Netherlands		k€ 6.205	63,7%
Flanders			
Universities	k€ 3.248		
Industry	k€ 450		
Total Flanders		k€ 3.698	37,3%
Total STEVIN R&D funding		k€ 9.903	

<i>Distribution STEVIN R&D funding over academic and industrial recipients</i>			
Universities	k€ 8.218		83,0%
Industry	k€ 1.685		17,0%
Total STEVIN R&D funding		k€ 9.903	

From the figures above it can be concluded that the realizations meet the target percentages set by the Dutch and Flemish funding organisations.

1. 1st Open Call for Proposals for strategic research proposals and HLT resources (data & tools) for focussed short term projects with a self-contained result (maximum duration is 2 years) (max. budget 2 M€)

Objectives STEVIN 1st open call for proposals

Proposals had to relate to basic linguistic resources (tools and data), fundamental strategic research and applications in the areas of language and speech technology, all of which had to contribute to an appropriate digital language infrastructure for Dutch. Proposals could be submitted both in the area of language technology, and in the area of speech technology, and were preferably relevant to both areas. For cross-border consortiums the standard bench fee was increased by 50%.

Evaluation procedure full proposals submitted in the 1st open call

All applications were presented to a panel of international experts in language and speech technology. This international panel evaluated the applications based on the assessment criteria mentioned in the call and formulated a set of recommendations for the STEVIN Programme Committee. Due to time constraints, the procedure for this call did not allow the applicants to respond to the panel's recommendations. Based on the applications and the panel's recommendations, the Programme Committee also assessed the applications and determined the order of priority of the eligible proposals. On the basis of both the IAP advice and the Programme Committee's advice, the Board of the STEVIN programme finally determined which projects were funded.

Time frame STEVIN 1st call – 2004 – length call procedure 3 months

- * September 15: opening call and brokerage event in Tilburg
- * November 2: closing date call: 19 proposals were submitted
- * November 25 and 26: assessment and ranking of all proposals by STEVIN IAP
- * December 3: assessment and ranking of all proposals by STEVIN PC
- * December 15: determining short list by Board of the STEVIN programme

Statistics: Number and percentages submitted and funded proposals, listed by type

	speech	language	speech/language combined	total
Submitted proposals	6 (30%)	9 (50%)	4 (20%)	19 (100%)
Funded proposals	2 (40%)	3 (60%)	0 (0%)	5 (100%)

2. 2nd Open Call for Proposals for strategic research proposals and HLT resources (data & tools) (max. budget 3,4 M€)

Objectives STEVIN 2nd open call for proposals

Proposals had to relate to basic linguistic resources (tools and data), fundamental strategic research and applications in the areas of language and speech technology, all of which had to contribute to an appropriate digital language infrastructure for Dutch. Proposals could be submitted both in the area of language technology, and in the area of speech technology, and were preferably relevant to both areas. For cross-border consortiums the standard bench fee was increased by 50%.

Evaluation procedure pre-proposals/full proposals submitted in the 2nd open call

The selection of the pre-proposals was carried out by the Programme Committee. The Programme Committee specifically assessed the expected contribution to the STEVIN aims. Applicants of 18 promising pre-proposals received a recommendation to submit a full proposal according to a specified format. The project leaders of selected pre-proposals were invited by the STEVIN Programme Committee for a short session during which the PC advised them as to the way they might extend their pre-proposal into a full proposal.

All eligible full proposals submitted in the open call were presented to a panel of international experts in language and speech technology. For this call the same experts were asked to serve in the International

Assessment Panel as the ones that acted as such in the first STEVIN Call be it that - to lower the workload - for this call two extra experts were asked to serve in this panel. The composition of the International Assessment Panel was given on the STEVIN website. This international panel evaluated and ranked the eligible applications based on the assessment criteria mentioned in the call and formulated a set of recommendations for the STEVIN Programme Committee.

The International Assessment Panel's assessment was sent to the applicants for comments. Based on the applications, the panel's recommendations and the applicant's response to the panel assessment, the Programme Committee again assessed the applications and determined the order of priority of the eligible proposals. On the basis of the International Assessment Panel's advice and the Programme Committee's advice, the Board of the STEVIN programme determined which projects were funded.

Time frame STEVIN 2nd call – 2005 – length call procedure 8 months

- * March 30: opening call and brokerage event in Antwerpen
- * April 26: closing date call: 34 pre-proposals were submitted
- * May 23: assessment pre-proposals by STEVIN PC – 18 pre-proposals selected
- * June 13: interviews with applicants selected pre-proposals
- * September 2: closing date call: 18 full proposals submitted
- * October 6 and 7: assessment and ranking of all proposals by STEVIN IAP
- * October: applicants formulated response to IAP assessment
- * November 14 and 15: assessment and ranking of all proposals by STEVIN PC
- * November 29: determining short list by TST Board

Statistics: Number and percentages submitted pre-proposals, full proposals and funded proposals, listed by type

	speech	language	total
Submitted pre-proposals	16 (48%)	18 (52%)	34 (100%)
Submitted proposals	6 (33%)	12 (66%)	18 (100%)
Funded proposals	3 (50%)	3 (50%)	6 (100%)

3. 3rd Open Call for proposals for application-oriented research (max. budget 2,3 M€)

Objectives STEVIN 3rd open call for proposals

The STEVIN programme aims to have a balanced programme covering all layers (resources, research and development, technology integration in applications, end users). In the preceding calls, only a few projects focused on technology integration in applications. For this reason in the 3rd call especially proposals for *application-oriented research projects* were invited.

Evaluation procedure full proposals submitted in the 3rd open call

All eligible full proposals submitted in the open call were presented to a panel of international experts in language and speech technology. For this call the same experts were asked to serve in the International Assessment Panel as the ones that acted as such in the second STEVIN Call. The composition of the International Assessment Panel was available on the STEVIN website. This international panel evaluated and ranked the eligible applications based on the assessment criteria mentioned in the call and formulated a set of recommendations for the STEVIN Programme Committee.

The panel's assessment was presented to the applicants for comments. Based on the applications, the panel's recommendations and the applicant's response to the panel assessment, the Programme Committee also assessed the applications and determined the order of priority of the eligible proposals. On the basis of the International Assessment Panel's advice and the Programme Committee's advice, the Board of the STEVIN programme determined which projects were funded.

Time frame STEVIN 3rd call – 2007 – length call procedure 4 months

- * April 15: opening call
- * May 22: closing date call: 15 full proposals were submitted
- * July 5 and July 6: assessment and ranking of all proposals by STEVIN IAP
- * July: applicants formulated response to IAP assessment
- * August 7-8: assessment and ranking of all proposals by STEVIN PC
- * August 21: determining short list by Board of the STEVIN programme

Statistics: Number and percentages submitted and funded proposals, listed by type

	speech	language	total
Submitted proposals	5 (33%)	10 (66%)	15 (100%)
Funded proposals	2 (40%)	3 (60%)	5 (100%)

4. Three Calls for tender for specific HLT resources (max budget 1,6 M€)

Objectives STEVIN calls for tender

The realisation of a number of specific top priorities for Dutch HLT

1. A speech recognition toolkit for Dutch
2. A lexical resource for the semantic processing of Dutch
3. An annotated written Dutch corpus

Evaluation procedure full proposals submitted in the calls for tender

The assessment and ranking of full proposals targeting the specific priorities was carried out by the STEVIN Programme Committee. The proposals and the Programme Committee assessment were subsequently forwarded to the International Assessment Panel for commenting. On the basis of the Programme Committee's advice and the comments of the International Assessment Panel, the Board of the STEVIN programme finally determined which projects were funded.

Time frame STEVIN Call for tender 1 and 2 – 2005 – length call procedure 8 months

- * March 30: opening call for tender 1 and 2
- * May 9: closing date call: 4 proposals were submitted (1x tender 1; 3x tender 2)
- * May 23: assessment proposals by STEVIN PC – more info requested from consortia
- * June 30: second discussion extended proposals by STEVIN PC
- * August: written assessment and ranking of all tender proposals by STEVIN IAP
- * November 29: funding decision by Board of the STEVIN programme

Time frame STEVIN Call for tender 3 – 2007 – length call procedure 12 months

- * December 1: opening call for tender 3
- * February 28: closing date call: 1 proposal was submitted
- * March/April: assessment proposal by STEVIN IAP – more info requested from consortium
- * August 7: assessment tender proposal by STEVIN PC
- * October: revised version SoNaR proposal submitted
- * November 16 2008: funding decision SoNaR phase 1 by Board of the STEVIN programme
- * July 1 2009: funding decision SoNaR phase 2 by Board of the STEVIN programme

Statistics

The statistics for this call are rather straightforward: for tenders 1 and 3 the HLT field formed broad Dutch/Flemish consortia containing all essential actors in the field that submitted a joint proposal. For tender 2, three proposals were submitted, one of which was selected.

5. Three Calls for proposals for demonstration projects (max. budget 1 M€)

Objectives STEVIN calls for demonstration systems

The objective of the STEVIN calls for demonstration systems was to try and stimulate demand for HLT technology by funding short-term (maximum length 15 months) projects for building demonstration projects using proven HLT technologies. Demonstrators could target to open new markets or new domains. Project consortia had to be led by a Dutch or Flemish HLT SME and could consist of both industrial and academic partners. Maximum size of a demonstration project is € 100.000. Three calls were opened in respectively 2005, 2006 and 2006. The total budget for the three calls was € 1.000.000.

Evaluation procedure full proposals submitted in the STEVIN calls for demonstration projects

All applications were assessed by a committee consisting of the two senior managers of the STEVIN Programme bureau, the STEVIN coordinator at the Dutch Language Union, an ICT expert from the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) and an ICT expert from Agentschap NL, the Dutch Ministry for Economic Affairs agency for innovation and sustainable development. As of the 2nd call all proposals short listed by the assessment committee were sent for a sanity check to the STEVIN Programme Committee. On the basis of the advice from the assessment committee and the sanity judgment of the PC the HLT Board made the final funding decision.

Time frame STEVIN calls for demonstration projects 2005, 2006 and 2007 – length call procedure 2 months

- * July 2005: opening 1st call
- * October 15 2005: closing date 1st call: 8 proposals were submitted
- * November 2005: assessment and ranking of all proposals by assessment committee
- * December 2005: 3 proposals funded by Board of the STEVIN programme

- * July 2006: opening 2nd call
- * October 15 2006: closing date 2nd call: 19 proposals were submitted
- * November 2006: assessment and ranking of all proposals by assessment committee
- * December 2006: sanity check by STEVIN programme committee of short listed proposals
- * December 2006: 6 proposals funded by Board of the STEVIN programme

- * July 2007: opening 3rd call
- * October 15 2007: closing date 3rd call: 13 proposals were submitted
- * November 2007: assessment and ranking of all proposals by assessment committee
- * December 2007: sanity check by STEVIN programme committee of short listed proposals
- * December 2007: 5 proposals funded by Board of the STEVIN programme

Statistics: Number and percentages funded proposals, listed by type

	speech	language	speech/language combined	total
Funded proposals	4 (30%)	7 (50%)	3 (20%)	14 (100%)

6. Calls for Educational/Master class project proposals (max. budget 110 k€)

Educational Projects

Educational projects are aimed at making students between age 15-20 within educational settings (school, museums, etc) aware of the possibilities of language and speech technologies. The maximum budget for each call is € 55.000. Proposals submitted in the calls are assessed by a panel of Dutch and Flemish educational experts. Their advice is sent to the STEVIN Working Group for supporting activities and the STEVIN Programme Committee for a sanity check. On the basis of the advice of the assessment panel and the sanity judgment of the PC, the HLT Board makes the final funding decision.

Three calls for educational projects were opened in 2007, 2008 and 2009. In the 1st call only one eligible proposal was submitted and funded (TST op Kennislink, € 27.500). In the 2nd call again only one eligible proposal was submitted and funded (DiaDemo, € 32.113). In the 3rd call for educational projects three eligible proposals were submitted (total requested budget € 89.625) and one was funded (TST op Kennislink2, € 25.5000).

Masterclass Projects

Masterclass projects are aimed at increasing general awareness of HLT research and applications within government organisations and the industry. The maximum budget for each call is € 20.000. Proposals submitted in the calls are assessed by the Working Group for STEVIN supporting activities. Their advice is sent to the STEVIN Programme Committee for a sanity check. Based on the advice of the Working Group and the sanity judgment of the PC, the HLT Board makes the final funding decision.

The 1st Call for Masterclasses was opened in 2008. Two eligible proposals were submitted, of which one proposal was funded (ICT & Dyslexie, € 17.500). The 2nd Call was opened in 2009. One eligible proposal was submitted and funded (TST voor Nederlandse overheidsdiensten, € 29.000).

Time frame STEVIN calls for Educational/Masterclass projects 2007, 2008, 2009 – length call procedure 3 months

- * June 30 2007: opening 2007 call for Educational projects
- * September 30 2007: closing date 2007 call: 1 proposal submitted
- * November 2007: 1 proposal funded by HLT Board

- * February 15 2008: opening 2008 call for Educational projects
- * May 15 2008: closing date 2008 call: 1 proposal submitted
- * August 2008: 1 proposal funded by HLT Board

- * October 31 2008: opening 2008 call for Masterclass projects
- * January 31 2009: closing date 2008 call: 2 proposals submitted
- * May 2009: 1 proposal funded by HLT Board

- * June 15 2009: opening 2009 call for Educational projects
- * September 15 2009: closing date 2009 call: 3 proposals submitted
- * December 2009: final decision still to be made by HLT Board

- * June 15 2009: opening 2009 call for Masterclass projects
- * September 15 2009: closing date 2009 call: 1 proposal submitted
- * December 2009: final decision still to be made by HLT Board

The STEVIN Priorities (as formulated in the original STEVIN project description)

Proposals can relate to basic linguistic resources (tools and data), fundamental strategic research and applications in the areas of language and speech technology, all of which have to contribute to an appropriate digital language infrastructure for Dutch. Proposals can be submitted both in the area of language technology, and in the area of speech technology, and are preferably relevant to both areas. For cross-border consortiums the standard bench fee (see 'eligible costs' on page 4) will be increased by 50%. Examples of language and speech applications which can be targeted are presented after the priorities for speech technology resources and research and those for language technology resources and research.

For speech technology, the priorities are:

for resources:

- speech and multimodal corpora for:
 - applications such as CALL (Computer Assisted Language Learning);
 - applications in which names and addresses play an important role;
 - CCQA applications (questions and answers in call centres), educational applications;
- multimodal corpora for applications of broadcast news transcription or person identification;
- text corpora for the development of stochastic language models;
- tools and data for the development of:
 - robust speech recognition;
 - automatic annotation of corpora;
 - speech synthesis;

for research:

- robustness of speech recognition;
- output treatment (inverse text normalization);
- confidence measures;
- adaptation;
- lattices.

For language technology, the priorities are:

for resources:

- richly annotated monolingual Dutch corpora;
- electronic lexicons;
- aligned parallel corpora;

for research:

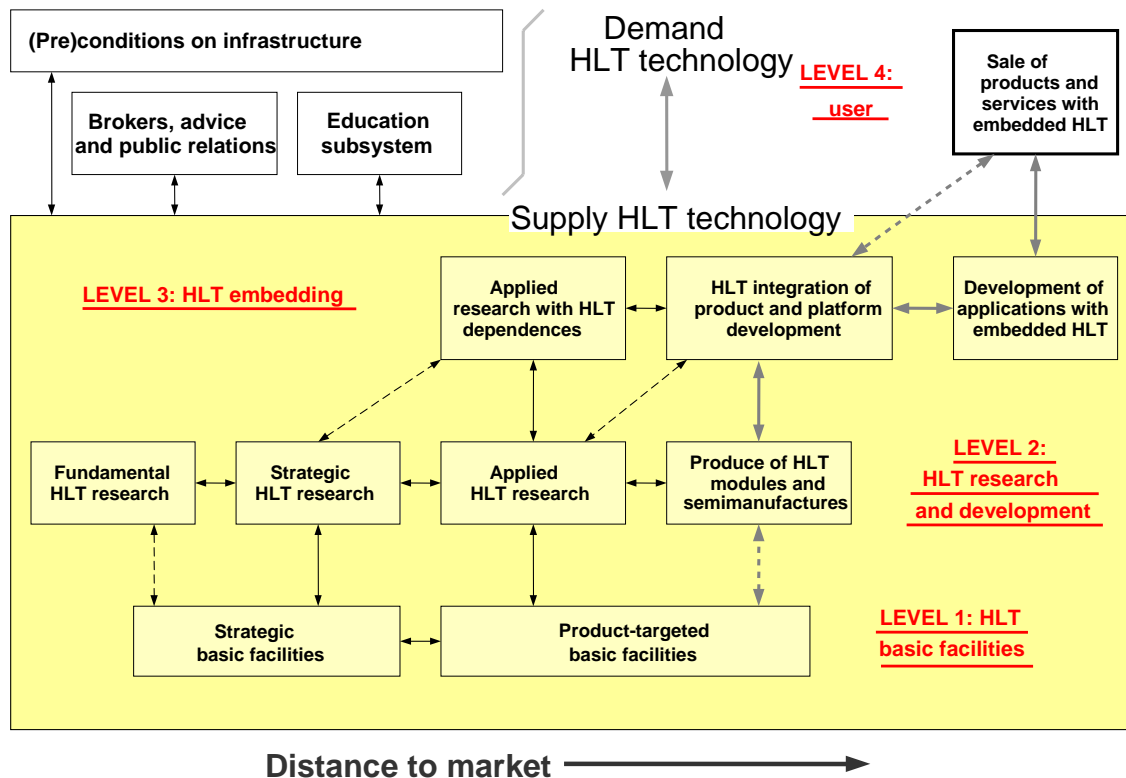
- semantic analysis, including semantic tagging and integrating morphological, syntactic and semantic modules;
- text pre-processing;
- morphological analysis;
- syntactic analysis (robust parsing).

In the area of applications (both for speech & language technology), examples to be targeted on are:

- information extraction from audio-transcripts created by speech recognizers;
- speaker accent and identity detection;
- monolingual or multilingual information extraction;
- semantic web;
- dialogue systems and Q&A solutions, especially in multimodal domains;
- automatic summarization and text generation applications;
- machine translation;
- educational systems.

Coverage of STEVIN Priorities by the projects funded within the STEVIN programme

STEVIN priorities together address different aspects of the stratified innovation system (as depicted below). STEVIN advocates an integrated approach: all layers in the stratified system are addressed, i.e. development of language and speech resources and tools, stimulating innovative fundamental and strategic research, stimulating application-oriented research, promote HLT embedding in existing applications and services, stimulate HLT demand via demonstrator projects and encourage cooperation and knowledge transfer between academia and industry.



In the table below an overview is given of how the projects funded in the various STEVIN funding schemes cover the STEVIN priorities and the layers of the innovation model.

Percentage of STEVIN funding per STEVIN priority		
• Speech technology resources	21,6%	
• Language technology resources	29,5%	
% STEVIN funding for basic resources		51,0%
• Speech technology research	14,3%	
• Language technology research	9,0%	
% STEVIN funding for basic research		23,3%
• Speech technology application-oriented research	7,3%	
• Language technology application-oriented research	8,1%	
% HLT Application-oriented research		15,4%
Speech technology demonstration projects	3,7%	
Language technology demonstration projects	6,5%	
% HLT Demonstration projects		10,2%
% STEVIN funding for speech technology		46,9%
% STEVIN funding for language technology		53,1%

Standard STEVIN Assessment criteria

Quality and innovative character of the proposal

- Clarity in problem definition and innovative power of the project.
- Suitability and effectiveness of the research design and methodology. In particular, an explicit component of evaluation, or in the case of linguistic resources, an explicit validation plan must be included in the proposal.
- Impact of the project on a wide range of applications and its importance to applications that are relevant to the industry.
- Competence of the participating groups (including past performance).
- Feasibility of the goals.
- The goal is to have a balanced programme covering all layers (basic resources, research and development, technology integration, end users) in the chain approach and properly integrating them. The contribution of individual projects to this overall programme goal will therefore be a criterion in their evaluation.
- Balanced cooperation and task division within the project.
- Availability of the required infrastructure.

Economic aspects of the project proposal

- Is there cooperation with or support by companies?
- What are the prospects for spin-offs and/or other new developments?
- Opportunities for applying the results in industry and/or society.

Contribution to the STEVIN-programme

- Conformity to the focus of the programme and fit in with the priorities set. The project must focus on the Dutch language and must contribute to improving or at least securing the position of the Dutch language in the modern information and communication society.
- Perspectives on knowledge transfer and network creation. In particular it is to the advantage of a project proposal if the expertise of Dutch and Flemish groups or companies are combined, if research institutes and companies jointly make a proposal, or if the proposal relates both to language and to speech technology.

IPR, avoiding duplication and standards

- The proposal must contain a clear plan for the proper treatment of intellectual property rights (IPR), both for the resources provided by third parties and for the results of the project. The working principle must be that the data, tools and other practical spin-offs resulting from the STEVIN-projects are made available in a non-discriminative way in the TST-centrale.
- The proposal must prove that the applicants have a precise and up-to-date picture of what is already available in terms of basic resources. Preferably the resources to be developed in the project do not exist yet. If it is known or can be presupposed that the resources exist but are not generally accessible, the proposal should contain a plan to avoid disturbance of the market, i.e. unfair competition must be avoided.
- The R&D Community must be able to access, use and exploit the basic resources resulting from the STEVIN-projects on non-discriminate terms. The applicants have to declare themselves willing to negotiate on this with the TST-centrale and to sketch the conditions that apply. Conclusion of a contract on the IPR-arrangements is a necessary condition for awarding of funding.
- The proposal must fit in with existing standards and apply these where possible, or cooperate on the development of new standards so that a maximum reuse of the basic resources developed is guaranteed.

Some specific criteria defining application-oriented proposals were added for the 3rd call where this type of proposals was specifically invited.

Code of Conduct for Independent Experts appointed in the international STEVIN Assessment Panel (IAP)

1. The task of an expert is to participate in a confidential, fair and equitable review of project(s) according to any programme-specific review documents. He/she must use his/her best endeavours to achieve this, follow any instructions given by the STEVIN Programme Bureau to this end and deliver a constant and high quality of work.
2. The reviewer works as an independent person. He/she is deemed to work in a personal capacity and, in performing the work, does not represent any organisation.
3. The independent expert must sign a declaration of conflict of interest and confidentiality before starting the work, by which he/she accepts the present Code of Conduct. Invited independent experts who do not sign the declaration will not be allowed to work as a reviewer.
4. In doing so, the independent expert commits him/herself to strict confidentiality and impartiality concerning his/her tasks. If a reviewer has a direct or indirect link with the project(s), or any other vested interest, or is in some way connected with the project(s), or has any other allegiance which impairs or threatens to impair his/her impartiality with respect to the project(s), he/she must declare such facts to the responsible STEVIN Programme Bureau official as soon as he/she becomes aware of this. The STEVIN Programme Bureau ensures that, where the nature of any link is such that it could threaten the impartiality of the reviewer, he/she does not participate in the review of the project(s) concerned.
5. Reviewers may not discuss any project details with others, including other reviewers or STEVIN Programme Bureau officials not directly involved in the review of the project, except during the formal review session moderated by or with the knowledge of the responsible STEVIN Programme Bureau official.
6. Where it has been decided that project details and/or project deliverables are to be posted or made available electronically to reviewers, who then work from their own or other suitable premises, the reviewer will be held personally responsible for maintaining the confidentiality of any documents or electronic files sent and returning or destroying all confidential documents or files upon completing the review as instructed. Reviewers may seek further information (for example through the internet, specialised databases, etc.) in order to allow them to complete their examination of the project details and/or deliverables, provided that the obtaining of such information respects the overall rules for confidentiality and impartiality. Reviewers may not show the contents of the deliverables or information on the project(s) to third parties (e.g. colleagues, students, etc.) without the express written approval of the STEVIN Programme Bureau. It is forbidden for reviewers to make direct contact with the project participants.
7. Reviewers are required at all times to comply strictly with any rules defined by the STEVIN Programme Bureau for ensuring the confidentiality of the review process and its outcomes. Failure to comply with these rules may result in exclusion from the immediate and future reviews, without prejudice to penalties that may derive from other applicable Regulations.

Code of Conduct for members of the STEVIN Programme Committee

1. The task of an expert is to participate in a confidential, fair and equitable review of project(s) according to any programme-specific review documents. He/she must use his/her best endeavours to achieve this, follow any instructions given by the STEVIN Programme Bureau to this end and deliver a constant and high quality of work.
2. The PC member works as an independent person. He/she is deemed to work in a personal capacity and, in performing the work, does not represent any organisation.
3. The PC member must sign a declaration of conflict of interest and confidentiality before starting the work, by which he/she accepts the present Code of Conduct. PC members who do not sign the declaration will not be allowed to be present at the assessment meeting of the PC.
4. In doing so, the PC member commits him/herself to strict confidentiality and impartiality concerning his/her tasks. If a reviewer has a direct or indirect link with the project(s), or any other vested interest, or is in some way connected with the project(s), or has any other allegiance which impairs or threatens to impair his/her impartiality with respect to the project(s), he/she must declare such facts to the responsible STEVIN Programme Bureau official as soon as he/she becomes aware of this. The STEVIN Programme Bureau ensures that, where the nature of any link is such that it could threaten the impartiality of the reviewer, he/she does not participate in the review of the project(s) concerned.
5. PC members may not discuss any project details with others, including other PC members or STEVIN Programme Bureau officials not directly involved in the review of the project, except during the formal assessment meeting moderated by or with the knowledge of the responsible STEVIN Programme Bureau official.
6. Where it has been decided that project details and/or project deliverables are to be posted or made available electronically to PC members, who then work from their own or other suitable premises, the PC members will be held personally responsible for maintaining the confidentiality of any documents or electronic files sent and returning or destroying all confidential documents or files upon completing the review as instructed. PC members may seek further information (for example through the internet, specialised databases, etc.) in order to allow them to complete their examination of the project details and/or deliverables, provided that the obtaining of such information respects the overall rules for confidentiality and impartiality. PC members may not show the contents of the deliverables or information on the project(s) to third parties (e.g. colleagues, students, etc.) without the express written approval of the STEVIN Programme Bureau. It is forbidden for PC members to make direct contact with the project participants.
7. PC members are required at all times to comply strictly with any rules defined by the STEVIN Programme Bureau for ensuring the confidentiality of the review process and its outcomes. Failure to comply with these rules may result in exclusion from the immediate and future reviews, without prejudice to penalties that may derive from other applicable Regulations.

Overview of STEVIN projects

1st Call for Proposals for strategic research proposals and HLT resources (data & tools) (max. budget 2 M€) - 2004

- Automata for deriving phoneme transcriptions of Dutch and Flemish names (AUTONOMATA)
- Coreference Resolution for Extracting Answers (COREA)
- Dutch Language Corpus Initiative (D-coi)
- Identification and Representation of Multi-word Expressions (IRME)
- Extension of CGN with speech of children, non-natives, elderly and human-machine interaction (JASMIN-CGN)

2nd Call for Proposals for strategic research proposals and HLT resources (data & tools) (max. budget 3,4 M€) - 2005

- Detecting and Exploiting Semantic Overlap (DAESO)
- Dutch Parallel Corpus (DPC)
- Large Scale Syntactic Annotation of written Dutch (Lassy)
- Missing Data Solutions (Midas)
- Northern and Southern Dutch Benchmark Evaluation of Speech recognition Technology (NBest)
- STEVIN can PRAAT

Call for proposals for applied research (max. budget 2,3 M€) - 2007

- Autonomata, Transfer of Output (Autonomata TOO)
- Dutch ILanguage Investigation of Summarization technology
- Development and Integration of Speech technology into COurseware for language learning (DISCO)
- Dutch Online Media Analysis (DuOMAn)
- Parse and Corpus based Machine Translation (PaCo-MT)

Three Calls for tender for specific HLT resources (max budget 1,6 M€) – 2005/2007

- Speech Processing, Recognition & Automatic Annotation Kit (Sprak)
- Combinatorial and Relational Network as Toolkit for Dutch Language Technology (Cornetto)
- Stevin Nederlandstalig Referentiecorpus (SoNaR)

Three Calls for proposals for demonstration projects (max. budget 1 M€) – 2005/2006/2007

- Rechtsorde
- GemeenteConnect!
- Spraakgestuurde Nummerbord Retrieval Tool
- Audiokrant
- Spelling- en grammaticacontrole voor dyslectische gebruikers
- Rechtspraakherkenning
- Klinkende Taal
- SpelSpiek
- Voice Assess
- Alfabetisering Anderstaligen Plan (AAP)
- Esay Info
- Hulp bij Auditieve Training na Cochleaire Implantatie (HATCI)
- Nederlandstalige Ondertiteling (Neon)
- Sprekende zelfcorrigerende woordvoorspeller voor dyslectische gebruikers (WooDy)

Three Calls for proposals for small educational projects/masterclasses – 2007/2008/2009

- Educa Project: Taal en spraaktechnologie op Kennislink
- Educa Project: Diademo
- Educa Project: Taal en spraaktechnologie op Kennislink2
- Masterclass: ICT en dyslexie
- Masterclass: TST voor Nederlandstalige overheidsdiensten

Overview proposals funded in the 1st Call for Proposals for strategic research proposals and HLT resources (data & tools)
(max. budget 2 M€)

<i>acronym</i>	<i>coordinating institute and other academic partners</i>	<i>industrial partners</i>	<i>VL/NL consortium nationality</i>	<i>STEVIN priorities addressed (subject)</i>	<i>planned duration</i>	<i>funding</i>
AUTONOMATA	Ghent University (Jean-Pierre Martens) Radboud Univ. Nijmegen Utrecht University	TeleAtlas Scansoft	VL/NL	Speech resources (speech synthesis)	24 mnths	€ 322.848
COREA	Groningen University (Gosse Bouma) Antwerpen University	Language and Computing	NL/VL	Language resources Language research (semantic annotation)	24 mnths	€ 353.875
D-coi	Radboud Univ. Nijmegen-CLST (Nelleke Oostdijk) Tilburg University Antwerpen University Twente University Utrecht University Groningen University Leuven University	Polderland	NL/VL	Language resources Speech resources (Corpus written Dutch protocols)	14 mnths	€ 566.531
IRME	Utrecht University (Jan Odijk) Groningen University	Van Dale Lexicografie	NL	Language resources Language research (semantic and syntactic annotation)	24 mnths	€ 389.500
JASMIN-CGN	Radboud Univ. Nijmegen- CLST (Catia Cucchiarini) Leuven University	TalkingHome	NL/VL	Speech resources (speech corpus)	24 mnths	€ 419.471



Automata for deriving phoneme transcriptions of Dutch and Flemish names (AUTONOMATA)

Project co-ordinator

Prof. dr. ir. J.-P. Martens
Gent University
ELIS Speech Lab
Sint-Pietersnieuwstraat 41
B-9000 Gent, België
Telephone: +32 9 264 33 95
E-mail: Jean-Pierre.Martens@elis.ugent.be
URL: www.elis.ugent.be

Project consortium

1. Prof. dr. ir. J.-P. Martens (Universiteit Gent, ELIS Speech Lab)
2. Dr. H. van den Heuvel (Radboud Universiteit Nijmegen, Centre for Language and Speech Technology - CLST)
3. Dr. ir. G. Bloothoof (Universiteit Utrecht, Utrecht institute of Linguistics - UiL-OTS)
4. Ir. L. Peirlinckx (TeleAtlas)
5. Dr. ir. J. Verhasselt (Nuance Communications International)

STEVIN funding: € 322.848

Duration: 01/06/2005 – 31/05/2007

Project summary

This project aims to build two resources: (1) a grapheme-to-phoneme (g2p) conversion tool set for creating good phonetic transcriptions for TTS (Text-to-Speech) and ASR (Automatic Speech Recognition) applications with a focus on phonetic transcriptions of names, and (2) a corpus of spoken name utterances for supporting more research towards better automatic name recognition.

Since all presently available g2p converters perform poorly on names, the project will create and make available to third parties, dedicated name g2p converters (for Dutch and Flemish) that will be designed to produce high quality canonical name transcriptions of person names and address items. The machine learning tools that will be used to design these converters will be made available to third parties as well. This way they can be applied to develop dedicated g2p converters for name categories that are not handled in this project.

It is acknowledged that the deployment of LST applications involving ASR of Dutch and Flemish could be raised significantly if (among other things) one would succeed in surpassing the present state-of-the-art in name recognition. This will first of all require tools for creating good canonical transcriptions of these names, as envisaged in this project, but on top of that it will also call for new methods for predicting the kind of variations of these pronunciations one is likely going to encounter in spoken name utterances of native and non-native speakers of Dutch and Flemish. For the development of such methods, one needs a substantial corpus of spoken name utterances. Such a corpus is presently not available for Dutch nor Flemish, and this project proposes to create one.

AUTONOMATA website: <http://speech.elis.ugent.be/autonomata/>

Coreference Resolution for Extracting Answers (COREA)

Project co-ordinator

Dr. G. Bouma
Rijksuniversiteit Groningen
Faculteit der Letteren, Informatiekunde Alfa-Informatica)
Oude Kijk in 't Jatstraat 26
Postbus 716
NL-9700 AS Groningen
Telephone: +31 50 363 59 37
E-mail: gosse@let.rug.nl
URL: www.rug.nl/let

Project consortium

1. Dr. G. Bouma (Rijksuniversiteit Groningen, Alfa-informatica)
2. Prof. dr. W. Daelemans (Universiteit Antwerpen, Centrum voor Nederlandse Taal and Spraak - CNTS, en Universiteit Tilburg, Induction of Linguistic Knowledge - ILK)
3. J.-L. Verschelde (Language and Computing NV)

STEVIN funding: € 353.875

Duration: 01/05/2005 – 30/04/2007

Project summary

Co reference resolution is a key ingredient for the automatic interpretation of text. It has been studied mainly from a linguistic perspective, with an emphasis on establishing potential antecedents for pronouns. Practical applications, such as Information Extraction (IE), summarization and Question Answering (QA), require accurate identification of co reference relations between noun phrases in general. Computational systems for assigning such relations automatically, require the availability of a sufficient amount of annotated data for training and testing. For Dutch, annotated data is scarce and co reference resolution systems are lacking.

In this project, we aim to develop a robust system for assigning such relations automatically, and we will investigate the effect of making co reference relations explicit on the accuracy of systems for IE and QA. We will annotate a limited amount of application-specific corpus material, which is required for the evaluation of the co reference resolution system in the context of IE and QA. The project contributes to the goals of STEVIN by providing a robust co reference resolution system which is applicable in a range of applications for Dutch, such as information extraction, question answering and summarization. In addition, general guidelines for co reference annotation will become available and a tool will be developed to support the annotation of co reference in text. Finally, a limited amount of data annotated with co referential information, including spoken language data, will be produced.

COREA website: <http://www.cnts.ua.ac.be/~hoste/corea.html>

Dutch Language Corpus Initiative (D-coi)

Project co-ordinator

Dr. N. Oostdijk
Radboud Universiteit Nijmegen
Faculteit der Letteren
Centre for Language and Speech Technology (CLST)
Postbus 9103
NL-6500 HD Nijmegen
Telephone: +31 24 361 27 65
E-mail: n.oostdijk@let.ru.nl
URL: www.let.ru.nl

Project consortium

1. Dr. N. Oostdijk (Radboud Universiteit Nijmegen, Centre for Language and Speech Technology - CLST)
2. Dr. A. van den Bosch (Universiteit Tilburg, Induction of Linguistic Knowledge - ILK)
3. Drs. Th. van den Heuvel (Polderland Language and Speech Technology BV)
4. Prof. dr. F. de Jong (Universiteit Twente, Human Media Interaction - HMI)
5. Dr. P. Monachesi (Universiteit Utrecht, Utrecht institute of Linguistics - UiL-OTS)
6. Dr. G. van Noord (Rijksuniversiteit Groningen, Alfa-informatica)
7. Prof. dr. F. Van Eynde (Katholieke Universiteit Leuven, Centrum voor Computerlinguïstiek - CCL)

STEVIN funding: € 566.531

Duration: 01/06/2005 – 31/12/2006

Project summary

The project proposed here can be characterized as a preparatory project and aims to produce a blueprint for the construction of a 500-million-word corpus of contemporary written Dutch. This will entail the design of the corpus and the development (or adaptation) of protocols, procedures and tools that are needed for sampling data, cleaning up, converting file formats, marking up, annotating, post editing, and validating the data. In order to support these developments, a 50-million-word pilot corpus will be compiled, parts of which will be enriched with linguistic annotations. The pilot corpus is intended to demonstrate the feasibility of the approach. It will provide the necessary testing ground on the basis of which feedback can be obtained about the adequacy and practicability of various annotation schemes and procedures, and the level of success with which tools can be applied. Moreover, it will serve to establish the usefulness of this type of resource and annotations for different types of HLT research and the development of applications. The Danish Center for Sprogteknologi (CST) will undertake the evaluation of the protocols and procedures. At the end of the project, the pilot corpus together with all other results obtained within the project will be handed over to the Dutch Language Union and be made available through the Flemish-Dutch HLT Agency (TST-centrale).

D-coi website: <http://lands.let.ru.nl/projects/d-coi/>

Identification and Representation of Multi-word Expressions (IRME)

Project co-ordinator

Prof. Dr. J. Odijk
Universiteit Utrecht
Faculteit der Letteren
Utrecht institute of Linguistics OTS (UiL-OTS)
Janskerkhof 13
NL-3512 JK Utrecht
Telephone: +31 30 253 60 76
E-mail: jan.odijk@let.uu.nl
URL: www-uilots.let.uu.nl

Project consortium

1. Prof. dr. J. Odijk (Universiteit Utrecht, Utrecht institute of Linguistics OTS - UiL-OTS)
2. Dr. G van Noord (Rijksuniversiteit Groningen, Alfa-Informatica)
3. Dr. G. Bouma (Rijksuniversiteit Groningen, Alfa-Informatica)
4. Dr. J. Zuidema (Van Dale Lexicografie BV)

STEVIN funding: € 389.500

Duration: 01/06/2005 – 31/08/2007

Project summary

The central problems that the project addresses are (i) the lack of large and rich formalized lexicons for multi-word expressions for use in NLP; (ii) the lack of proper methods and tools to extend the lexicon of an NLP-system for multi-word expressions given a text corpus in a maximally automated manner. Therefore, the project aims to develop innovative methods and tools for the automatic identification and lexical representation of multi-word expressions. Concomitantly, a 5.000 entry corpus-based multi-word expression lexical database for Dutch will be developed. The database will be externally validated, and its usability will be evaluated in two independent NLP-systems for Dutch. The project contributes to the development of electronic lexicons, in particular for Dutch. The MWE database to be developed fills a gap in existing lexical resources for Dutch. The project carries out strategic research into generic methods and tools for MWE identification and lexical representation, focusing on Dutch, but these tools will be largely language-independent and can also be used for other languages, new domains, and beyond this project. In this way the project contributes directly to strengthening the digital infrastructure for Dutch.

IRME website: <http://www-uilots.let.uu.nl/irme/>

Extension of CGN with speech of children, non-natives, elderly and human-machine interaction (JASMIN-CGN)

Project co-ordinator

Dr. C. Cucchiarini
Radboud Universiteit Nijmegen
Faculteit der Letteren
Centre for Language and Speech Technology (CLST)
Postbus 9103
NL-6500 HD Nijmegen
Telephone: +31 24 361 57 85
E-mail: C.Cucchiarini@let.ru.nl
URL: www.let.ru.nl

Project consortium

1. Dr. C. Cucchiarini (Radboud Universiteit Nijmegen, Centre for Language and Speech Technology - CLST)
2. Prof. dr. H. Van hamme (Katholieke Universiteit Leuven, ESAT/PSI Speech Group)
3. Dr. ir. F.M.A. Smits (TalkingHome)

STEVIN funding: € 419.471

Duration: 01/04/2005 – 30/09/2007

Project summary

Large speech corpora (LSC) constitute an indispensable resource for conducting research in speech processing and for developing real-life speech applications. In 2004 the Spoken Dutch Corpus (Corpus Gesproken Nederlands - CGN) became available, which constitutes a plausible sample of standard Dutch as spoken by adult natives in the Netherlands and Flanders. Owing to budget constraints, CGN does not include speech of children, non-natives, elderly people and recordings of speech produced in human-machine interactions.

Since such recordings would be extremely useful for conducting research and for developing HLT applications for these specific groups of speakers of Dutch, the present proposal aims at extending CGN in three dimensions. First, by collecting a corpus of contemporary Dutch as spoken by children of different age groups, non-natives with different mother tongues and elderly people in the Netherlands and Flanders (JASMIN-CGN), we aim at an extension along the age and mother tongue dimensions. In addition, we intend to collect speech material in a communication setting that was not envisaged in CGN: human-machine interaction. Therefore, in this project part of the speech material from the three speaker groups will be collected in a setting of human-machine communication. We expect that the knowledge gathered from these data can be generalized to developing appropriate systems also for other speaker groups (i.e. adult natives). One third of the data will be collected in Flanders and two thirds in the Netherlands.

JASMIN-CGN website: <http://www.esat.kuleuven.be/psi/spraak/projects/JASMIN/>

Overview proposals funded in the 2nd Call for Proposals for strategic research proposals and HLT resources (data & tools)
(max. budget 3,4 M€)

<i>acronym</i>	<i>coordinating institute and other academic partners</i>	<i>industrial partners</i>	<i>VL/NL consortium nationality</i>	<i>STEVIN priorities addressed (subject)</i>	<i>planned duration</i>	<i>funding</i>
DAESO	Tilburg University (Emiel Kraemer) Antwerpen University Universiteit van Amsterdam	Textkernel	NL/VL	Language research Language resources (Semantic / discourse annotation)	36 mnths	€ 487.000
DPC	KU Leuven (Piet Desmet) Hogeschool Gent		VL	Language resources (Multilingual corpora / translational equivalents)	34 mnths	€ 498.000
LASSY	Groningen University (Gertjan van Noord) KU Leuven		NL/VL	Language resources (Syntactic treebank)	36 mnths	€ 496.000
MIDAS	KU Leuven (Hugo Van hamme) Radboud Univ. Nijmegen	Nuance	VL/NL	Speech research (Robust ASR)	48 mnths	€ 499.000
NBest	TNO-TM (David van Leeuwen) KU Leuven, Twente University, Radboud Univ. Nijmegen, Ghent University, SPEX, TU Delft		NL/VL	Speech resources (ASR benchmarks for evaluation)	29 mnths	€ 470.000
STEVIN can PRAAT	Universiteit van Amsterdam (Paul Boersma) Leiden University SPEX	Speech-Minded	NL	Speech resources (ASR, annotation tool)	24 mnths	€ 114.000

Detecting and Exploiting Semantic Overlap (Daeso)

Project co-ordinator

dr. E. Krahmer
Tilburg University
Faculteit Communicatie en Cultuur
Taal en Informatica
Warandelaan 2
5037 AB Tilburg
Telephone: +31 13-466 25 68
E-mail: e.j.krahmer@uvt.nl
URL: <http://let.uvt.nl/research/ti>

Project consortium

1. dr. E. Krahmer (Tilburg University)
2. prof. dr. W. Daelemans (Antwerp University)
3. prof. dr. M. de Rijke (University of Amsterdam)
4. drs. J. Zavrel (Textkernel)

STEVIN funding: € 487.000

Duration: 01/10/2006 – 30/09/2009

Project summary

The well-known fact that similar information can be expressed in many different ways is one of the major challenges in building robust NLP applications. It is commonly assumed that such applications can be improved with knowledge of how natural language expressions relate to each other, for instance in terms of paraphrases (same semantic content, different wording) or entailments (one expression implied by the other). DAESO investigates the detection of semantic overlap between Dutch sentences and the exploitation of this knowledge in a range of NLP applications. For this purpose, tools will be developed for the automatic alignment and classification of semantic relations (between words, phrases and sentences) for Dutch, as well as for a Dutch text-to-text generation application which fuses related sentences into a single grammatical sentence, which may be a generalization, a specification or a reformulation of the input sentences. To facilitate development and testing of these tools, an annotated monolingual Dutch parallel/comparable corpus of 1M words will be developed, consisting of pairs of texts that express comparable information. The utility of the resources and tools will be demonstrated in the context of three applications: (1) question-answering systems (improved recall, more complete answers), (2) information extraction (improved recall), and (3) summarization (beyond extraction: sentence compression, sentence fusion, anaphora resolution).

Daeso website: <http://daeso.uvt.nl/>

Dutch Parallel Corpus (DPC)

Project co-ordinator

Prof. dr. Piet Desmet
Katholieke Universiteit Leuven Campus Kortrijk
Etienne Sabbelaan 53
B-8500 Kortrijk
Telephone: +32 (0) 56 24 61 85
E-mail: Piet.Desmet@kuleuven-kortrijk.be
URL: http://wwwling.arts.kuleuven.ac.be/franling_n/pdesmet

Project consortium

1. Prof. Dr. Piet Desmet (Katholieke Universiteit Leuven Campus Kortrijk)
2. Prof. Dr. Willy Vandeweghe (Hogeschool Gent, School of Translation Studies)
3. Dr. Hans Paulussen (Katholieke Universiteit Leuven Campus Kortrijk)
4. Dra. Lieve Macken (Hogeschool Gent, School of Translation Studies)

STEVIN funding: € 498.000

Duration: 01/05/2006 – 28/02/2009

Project summary

Aligned parallel corpora form an indispensable resource for a wide range of multilingual applications, a.o. machine translation (especially corpus-based MT such as statistical and example-based MT), computer-assisted translation tools, cross-lingual information extraction, multilingual terminology extraction, and computer-assisted language learning. Since high-quality parallel corpora with Dutch as the central language do not exist or are not accessible for the research community due to copyright restrictions, the compilation of aligned parallel corpora is one of the priorities of the STEVIN program. In this project, we want to construct a 10-million-word, high-quality, sentence-aligned parallel corpus for the language pairs Dutch-English and Dutch-French.

As the corpus will be bidirectional (Dutch as source and target language), the corpus can also be used as a comparable corpus (to compare texts originally written in Dutch with translated Dutch texts). A part of the corpus will be trilingual and will contain Dutch texts translated into both English and French. The corpus will be enriched with linguistic annotations. To guarantee the quality of the corpus and its multifunctional availability for the wide research community, each step in compiling, structuring and annotating the corpus will be validated by a user group of specialists in linguistics and language technology. Dutch being the pivotal language, we will collaborate closely with the researchers of the D-COI project, who are compiling a 50-million-word pilot corpus of contemporary written Dutch. In order to make the corpus accessible for the whole research community, we intend to obtain copyright clearance for all samples included in the corpus.

DPC website: <http://www.kuleuven-kortrijk.be/DPC>

Large Scale Syntactic Annotation of written Dutch (Lassy)

Project co-ordinator

dr. G.J.M. van Noord
Rijksuniversiteit Groningen
Faculteit der Letteren - Alfa-informatica
Oude Kijk in 't Jatstraat 26
Postbus 716
9700 AS Groningen
Telephone: +31-50-3637811
E-mail: vannoord@let.rug.nl
URL: <http://www.rug.nl/let/onderzoek/onderzoekinstututen/clcg/onderzoek/computing>

Project consortium

1. Dr. G.J.M. van Noord (Alfa-informatica Groningen)
2. Drs. I. Schuurman (CCL Leuven)
3. Prof. dr. F. van Eynde (CCL Leuven)
4. Dr. G. Bouma (Alfa-informatica Groningen)

STEVIN funding: € 496.000

Duration: 01/11/2006 – 31/10/2009

Project summary

A large corpus of written Dutch texts (1,000,000 words) is syntactically annotated (manually corrected), based on D-COI. In addition, the full D-COI corpus is syntactically annotated automatically. The project aims to extend the available syntactically annotated corpora for Dutch both in size as well as with respect to the various text genres and topical domains. In addition, various browse and search tools for syntactically annotated corpora will be further developed and made available. Their potential for applications in corpus linguistics and information extraction will be illustrated and evaluated.

Lassy website: <http://www.let.rug.nl/~vannoord/Lassy/>

Missing Data Solutions (Midas)

Project co-ordinator

Prof. dr. ir. H. Van hamme
Katholieke Universiteit Leuven
ESAT - PSI
Kasteelpark Arenberg 10
3001 Heverlee
Telephone: + 32 16 32 18 42
E-mail: hugo.vanhamme@esat.kuleuven.be
URL: <http://www.esat.kuleuven.be/psi/spraak/>

Project consortium

1. Prof. dr. ir. H. Van hamme (Katholieke Universiteit Leuven)
2. Dr. ir. B. Cranen (Radboud Universiteit Nijmegen)
3. Dr. J. De Veth (Radboud Universiteit Nijmegen)
4. Ir. B. D'hoore (Nuance Communications International)

STEVIN funding: € 499.000

Duration: 01/10/2006 – 30/09/2010

Project summary

Robustness to noise in automatic speech recognition is essential for the development of successful applications. Noise reduction techniques have been applied with some success in the past, but there remains a large performance gap between the best ASR implementations and human recognition, especially when the noise is non-stationary. This project tackles the noise robustness problem in ASR through missing data techniques (MDT) by addressing important open R&D issues for accuracy improvement and computational efficiency. Detectors of missing data will make minimal assumptions on the noise, while incorporating more knowledge about speech. The acoustic model in the recognizer's back-end will be refined and its evaluation will be made faster through algorithmic research. The developed algorithms will be integrated in the result of the STEVIN "call for tender - speech recognizer" (referred to as CFT-system) and made available through its distribution channels. This project contains language-independent research as well as work that is specific for Dutch, which both are of interest to the STEVIN program. It addresses three STEVIN priorities: 1) robustness of speech recognition, 2) tools and data for the development of robust speech recognition, and 3) confidence measures. How to account best for realistic environmental noise is largely language independent. However, the search for representations of speech that lead to better missing data implementations requires building new acoustic models that are language specific. In this project we will base our research on a "real-life" test suite that contains test material from the Dutch SpeechDat Car and Speecon databases.

Midas website: <http://www.esat.kuleuven.be/psi/spraak/projects/index.php?proj=MIDAS>



Northern and Southern Dutch Benchmark Evaluation of Speech recognition Technology (NBest)

Project co-ordinator

Ir. D.A. van Leeuwen
Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek
Technische Menskunde - Cognitieve Psychologie
Postbus 23
Kampweg 5
3769 ZG De Soesterberg
Telephone: +31 346 356 235
E-mail: david.vanleeuwen@tno.nl
URL: <http://www.tno.nl>

Project consortium

1. Dr. D. A. van Leeuwen (TNO Coordination)
2. Dr. H. van den Heuvel (SPEX Database recording)
3. Prof. L. Boves (CLST, RU Nijmegen)
4. Dr. R. J. F. Ordeman (HMI, Twente University)
5. Prof. dr. P. Wambacq (ESAT, Leuven University)
6. Prof. dr. J.-P. Martens (ELIS, Gent University)
7. Dr. L. J. M. Rothkrantz (EWI Delft University)

STEVIN funding: € 470.000

Duration: 01/05/2006 – 30/09/2008

Project summary

Over the years, standardised benchmark evaluation tests have proved indispensable for the development of several techniques in speech technology. In N-Best we will organise and execute an evaluation of large vocabulary speech recognition systems trained for Dutch (both Northern and Southern Dutch) in two evaluation conditions (Broadcast News and Conversational Telephony Speech). The goals of the project are the definition of a proper evaluation setup and a corresponding set of benchmark results. The evaluation framework can serve both as a basis for future evaluations, which can probe the progress in large vocabulary speech recognition for Dutch, and as an aid for the development of new speech recognition technologies for the Dutch language. Participants will use a common speech database, the Corpus Gesproken Nederlands (CGN), for acoustic training of their systems, as well as other common resources for language modeling and pronunciation modeling. They will co-operate through exchange of intermediate experiences, results and models of sub-technologies. The evaluation will be open to researchers outside the project, who will benefit from the common training and evaluation resources and the development experiences of the project partners. Intermediate and final exchange of experimental results and findings will be consolidated in workshops. The evaluation will be based on new speech material that will be collected and annotated for the purpose of this evaluation. All evaluation resources, materials and results will be made available via the TST-centrale.

NBest website: <http://speech.tm.tno.nl/n-best/>

STEVIN can PRAAT

Project co-ordinator

Prof. dr. P.P.G. Boersma
Universiteit van Amsterdam
Faculteit der Geesteswetenschappen - Fonetiek
Herengracht 338
1016 CG Amsterdam
Telephone: +31 20 525 2183
E-mail: paul.boersma@uva.nl
URL: <http://www.fon.hum.uva.nl/praat>

Project consortium

1. Prof. dr. P. Boersma (ACLC, University of Amsterdam)
2. Prof. dr. F. Hilgers (ACLC, University of Amsterdam / Nederlands Kanker Instituut - Anthonie van Leeuwenhoekziekenhuis)
3. Prof. dr. V. van Heuven (University of Leiden)
4. Dr. H. van den Heuvel (SPEX: Speech Processing EXpertise centre)
5. Dr. D.J.M. Weenink (ACLC, University of Amsterdam / SpeechMinded)

STEVIN funding: € 114.000

Duration: 01/01/2008 – 30/09/2008

Project summary

Appropriate tools are indispensable for the scientist to perform his/her work. This holds true for speech science as well. The PRAAT program¹ is an extensive application for language, music and speech research that is used by approximately 10,000 scientists and students around the globe. Some characteristics that explain its success right from the beginning, are the wide range of features, the user-friendliness and the scriptability, i.e. the possibility to create one's own processing for a series of inputs. The other aspect that adds to the enthusiastic and widespread use is the careful support available. This encompasses user help on diverse levels online, quick response to any questions by email, immediate handling of incidents and solving of problems, and last but not least, an infrastructure for user groups. The knowledge that the PRAAT program entails, in this means passed on to many colleagues and students. Also, users have a way to relate to one another and share their insights with regard to the possibilities the PRAAT program offers. The software is freely available for all current computer platforms like Linux, Windows and Macintosh. The manuals, FAQ and help menu are included in the package; the user group is available on the internet. Despite the multitude of features already present in the application, some important functionality is still missing. We propose to develop a number of improvements and added functionality that will then additionally and freely become available for speech scientists via the PRAAT program. This project matches the STEVIN objectives since it delivers important tools to all speech scientists who need state of the art technology to tackle the newest ideas and the largest datasets.

STEVINcanPRAAT website: <http://www.fon.hum.uva.nl/praat/>

Overview proposals funded in Call for proposals for application-oriented research
(max. budget 2,3 M€)

<i>acronym</i>	<i>coordinating institute and other academic partners</i>	<i>industrial partners</i>	<i>VL/NL consortium nationality</i>	<i>STEVIN priorities addressed (subject)</i>	<i>planned duration</i>	<i>funding</i>
AUTONOMATA TOO	Radboud University Nijmegen – CLST (Henk van den Heuvel) Ghent University Utrecht University	TeleAtlas Nuance	NL/VL	Speech Research Speech Application (ASR)	24 mnths	€ 416.750
DAISY	KU Leuven (Sien Moens) Groningen University	Q-go R&D	VL/NL	Language Research Language Application (Summarization)	36 mnths	€ 457.300
DISCO	Radboud University Nijmegen – CLST (Helmer Strik) Antwerpen University Radboud University Nijmegen – UTC	Polderland Language & Speech Technology	NL/VL	Speech research Speech Application (Computer assisted language learning)	36 mnths	€ 495.419
DuOMan	Universiteit van Amsterdam (Maarten de Rijke) Groningen University Hogeschool Gent	TrendLight GridLine	NL/VL	Language Research Language Application (Opinion and sentiment mining)	36 mnths	€ 440.447
PaCo-MT	KU Leuven – CCL (Frank Van Eynde) Groningen University	OneLiner Language & eBusiness Solutions BVBA	VL/NL	Language Research Language Application (Machine translation)	36 mnths	€ 494.575

Autonomata TOO

Project co-ordinator

Dr. H. van den Heuvel
Radboud Universiteit Nijmegen
Faculteit der Letteren
Taalwetenschap
Postbus 9103
6500 HD Nijmegen
Telephone: +31-24-3611686
E-mail: heuvel@let.ru.nl
URL: www.let.ru.nl

Project consortium

1. Dr H. van den Heuvel (CLST, Radboud University Nijmegen)
2. Prof. Dr J-P. Martens (ELIS, Ghent University)
3. Dr Ir G. Bloothoof (Utrecht institute of Linguistics (UiL-OTS), Utrecht University)
4. Ir L. Peirlinckx (TeleAtlas, Ghent)
5. Ir B. D'hoore (Nuance Communications International, Merelbeke)

STEVIN funding: € 416.750

Duration: 01/02/2008 – 31/01/2010

Project summary

The aim of this application-oriented research project is to build a demonstrator version of a Dutch/Flemish Points of Interest (POI) information providing business service, and to investigate new pronunciation modeling technologies that can help to bring the spoken name recognition component of such a service to the required level of accuracy. The demonstrator service (running on a PC) will contain a simple user interface and a restricted but realistic database of POI information. It will give a flavour of what the envisaged service can offer to the user, and it will also be used as a vehicle for testing the benefits of the newly developed speech technology in a realistic setting, involving tests with end users at strategic moments during the project.

AUTONOMATA TOO website: <http://lands.let.ru.nl/projects/AutonomataToo/>

Dutch IAngeage Investigation of Summarization technologY (DAISY)

Project co-ordinator

Prof. dr. M.F. Moens
Katholieke Universiteit Leuven
Departement Computerwetenschappen
Celestijnenlaan 200 A
B-3001 Heverlee
Telephone:
E-mail: sien@cs.kuleuven.be
URL: <http://www.cs.kuleuven.be/~sien/>

Project consortium

1. Prof. dr. M.-F. Moens (Department of Computer Science, K.U.Leuven)
2. Dr. G.J.M. van Noord (CLCG/Computational Linguistics, RuG University of Groningen)
3. Dr. Leonoor van der Beek (Q-go Research & Development)

STEVIN funding: € 457.300

Duration: 01/05/2008 – 30/04/2011

Project summary

Summarization of text is often a necessity when searching and selecting information from document repositories. However, current summarization technology is for a large part restricted to the extraction of sentences. Summarization technology for Dutch is very scarce. The aim of DAISY is to develop and evaluate essential technology for automatic summarization of Dutch informative texts. Innovative algorithms for topic salience detection, topic discrimination, rhetorical classification of content, sentence compression and text generation will be implemented. In addition, a demonstrator will be developed in collaboration with the company Q-Go.

The summarization demonstrator will be tested and evaluated in multiple ways in the QA environment of Q-go on documents in the financial and social security domains. Firstly, the system output will be compared against hand-made abstracts of the documents. Secondly, the effect of adding system-generated headline abstracts on retrieval will be measured. Finally, if suitable training and testing material can be obtained, tests will be done with automated email answering, where the summary of the email is used as input for the Q-go QA system.

DAISY website: <http://www.cs.kuleuven.be/~liir/projects.php?project=172>

Development and Integration of Speech technology into COurseware for language learning (DISCO)

Project co-ordinator

Dr. W.A.J. Strik
Radboud Universiteit Nijmegen
Faculteit der Letteren
Taalwetenschap
Postbus 9103
6500 HD Nijmegen
Telephone: +31-24-361 61 04
E-mail: strik@let.ru.nl
URL: www.let.ru.nl

Project consortium

1. Dr. H. Strik (Centre for Language and Speech Technology, Radboud University Nijmegen)
2. Prof. Dr. J. Colpaert (Linguapolis, Universiteit Antwerpen)
3. Drs. J. Bakx (Universitair Taal- en Communicatiecentrum Nijmegen)
4. Dr. I. de Mönnink (Polderland Language & Speech Technology)

STEVIN funding: € 495.419

Duration: 01/02/2008 – 31/01/2011

Project summary

Language learners are known to fare best in one-on-one interactive learning situations in which they receive optimal corrective feedback. However, providing this type of tutoring by trained language instructors is time-consuming and costly, and therefore not feasible for the majority of language learners. This particularly applies to oral proficiency, where corrective feedback has to be provided immediately after the utterance has been spoken, thus making it even more difficult to provide sufficient practice in the classroom. The recent appearance of Computer Assisted Language Learning (CALL) systems that make use of Automatic Speech Recognition (ASR) and other advanced automatic techniques offers new perspectives for training oral proficiency in a second language (L2).

The present project aims to develop and test a prototype of an ASR-based CALL application for training oral proficiency for Dutch as a second language (DL2). The application optimizes learning through interaction in realistic communication situations and provides intelligent feedback on various aspects of DL2 speaking, viz. pronunciation, morphology and syntax. The communicative settings employed in Nieuwe Buren (New Neighbours, a method for DL2 training developed by Malmberg publishers) will constitute the starting point for the application.

Disco website: <http://lands.let.ru.nl/~strik/research/DISCO/index.html>

Dutch Online Media Analysis (DuOMAn)

Project co-ordinator

Prof. dr. M. de Rijke
Universiteit van Amsterdam
Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Instituut voor Informatica, Information & Language Processing Systems
Telephone: +31-20-5255358
E-mail: mdr@science.uva.nl
URL: <http://staff.science.uva.nl/~mdr/>

Project consortium

1. Prof. dr. M. de Rijke (University of Amsterdam (UvA))
2. R. Franz (TrendLight)
3. T. Spaan (GridLine)
4. Dr. G. van Noord (Rijksuniversiteit Groningen (RuG))
5. Dr. V. Hoste (Dept. Vertaalkunde, Hogeschool Gent (HoGent))

STEVIN funding: € 440.447

Duration: 01/04/2008 – 31/03/2011

Project summary

When marketing campaigns or policies on sensitive or broad-ranging issues need to be defined or revised, access to the opinion of the target group is vital. An explosion in online content---both edited and user-generated---has vastly increased the range of opinions potentially available to media analysts and the general public alike, but efficient and effective access methods are needed to unlock this potential. The DuOMAn project will carry out an ambitious research agenda that will result in the development of a set of Dutch language resources and tools for identifying and aggregating sentiments in online data sources.

DuOMAn aims to transform the volumes of online information that threaten to leave media analysts information-bound into aggregates of attitudes organized by topic by employing classification, information extraction, and cross-document linking. DuOMAn will provide media analysts and members of the general public with focused access to opinionated information on people, products and topics through an online demonstrator for the general public and through integration of the tools and resources it develops into the workflow of professional media analysts. Key research contributions include sentiment-oriented lexical resources and advancement in the areas of automated sentiment analysis, parsing, and entity detection and coreference resolution. Applied research on robustness and adaptability receives central emphasis.

DuOMAN website: <http://staff.science.uva.nl/~mdr/Research/Projects/index.html>

Parse and Corpus based Machine Translation (PaCo-MT)

Project co-ordinator

Prof. dr. F. van Eynde
Katholieke Universiteit Leuven
Faculteit Letteren
Centrum voor Computerlinguïstiek
Maria Theresiastraat 21, postbus
B-3000 Leuven
Telephone: +32-16-325084
E-mail: frank.vaneynde@ccl.kuleuven.be
URL: www.ccl.kuleuven.be

Project consortium

1. Prof. Dr. F. Van Eynde (Centre for Computational Linguistics (CCL), K.U.Leuven)
2. Dr. J. Tiedemann (Alfa-informatica, Rijksuniversiteit Groningen (RUG))
3. Drs. K. Desmet (OneLiner Language & eBusiness Solutions BVBA)

STEVIN funding: € 494.575

Duration: 01/02/2008 – 31/01/2010

Project summary

In this project, we aim at building a hybrid machine translation system combining the positive features of corpus based and rule based systems. The primary goal is to develop an open-domain MT system for Dutch-English and Dutch-French (in both directions) integrating proper linguistic analysis and syntactic transfer into a data-driven approach. Compared to other data-driven approaches, we emphasise the improvement of translation quality and the adaptability of the system to the users requirements. This will result in a flexible MT system that is accepted by professional translators. Adaptability to users needs will be supported by a post editing interface, making the system very flexible and able to improve gradually. This novel feature increases the acceptability of the system by professional users. An evaluation of the system by human judgement and automated scores like BLEU/NIST and edit distance will be made, as well as a user test in which the translation speed will be tested.

PaCo-MT website: <http://www.ccl.kuleuven.be/~frank/Projects.html>

Overview proposals funded in three Calls for tender for specific HLT resources
(max budget 1,6 M€)

<i>acronym</i>	<i>coordinating institute and other academic partners</i>	<i>industrial partners</i>	<i>VL/NL consortium nationality</i>	<i>STEVIN priorities addressed (subject)</i>	<i>planned duration</i>	<i>funding</i>
SPRAAK	KU Leuven (Patrick Wambacq) Radboud Universiteit Nijmegen – CLST TNO Human Factors Universiteit Twente - HMI		VL/NL	Speech resources (ASR)	26 mnths	€ 400.000
CORNETTO	Free University Amsterdam (Piek Vossen) Universiteit van Amsterdam KU Leuven	Irion Technologies bv		Language resources (Semantic lexicon)	24 mnths	€ 400.000
SoNaR	Radboud Universiteit Nijmegen – CLST (Nelleke Oostdijk) Antwerpen University Hogeschool Gent Leuven University Instituut voor Nederlandse Lexicografie Groningen University Tilburg University Twente University Utrecht University Universiteit van Amsterdam SPEX, Nijmegen	Polderland Logica-CMG Dutchear Nuance IRION Van Dale Lexicografie Dutch HLT Agency	NL/VL	Language resources speech resources (Annotated written Dutch corpus)	36 mnths	€ 936.000

Speech Processing, Recognition & Automatic Annotation Kit (SpraaK)

Project co-ordinator

Prof. dr. Wambacq
Katholieke Universiteit Leuven
ESAT - PSI
Kasteelpark Arenberg 10
3001 Heverlee
Telephone: + 32 16 32 10 57
E-mail: patrick.wambacq@esat.kuleuven.be
URL: <http://www.esat.kuleuven.be/psi/spraak>

Project consortium

1. Prof. P. Wambacq (Katholieke Universiteit Leuven - ESAT/PSI)
2. Prof. L.W.J. Boves (Radboud Universiteit Nijmegen - Language and Speech RU)
3. Dr. Ir. D.A. van Leeuwen (TNO Human Factors (Soesterberg) TNO)
4. Dr. R. Ordeman (Universiteit Twente - Human Media Interaction UT)

STEVIN funding: € 400.000

Duration: 01/02/2006 – 31/05/2008

Project summary

The availability of a speech recognition system for Dutch is mentioned as one of the essential requirements for the language and speech technology (LST) community. Indeed, researchers now are faced with the problem that no good speech recognition tool is available for their purposes or existing tools lack functionality or flexibility. This project has two primary goals that will be accomplished within a single software framework. The first goal is to develop a highly modular toolkit for research into speech recognition algorithms. It allows researchers to focus on one particular aspect of speech recognition technology without needing to worry about the details of the other components. The second goal is to provide a state-of-the-art recogniser for Dutch with a simple interface, so that it can be used by non-specialists with a minimum of programming requirements. Next to speech recognition, the resulting software will enable applications in related fields as well. Examples are linguistic and phonetic research where the software can be used to segment large speech databases or to provide high quality automatic transcriptions. We choose the existing ESAT recogniser, augmented with knowledge and code from the other partners in this project, as a starting point. This code base will be transformed to meet the specified requirements. The transformation is accomplished by improving the software interfaces to make the software package more user friendly and adapted for usage in a large user community, and by providing adequate user and developer documentation written in English, so as to make it easily accessible to the international LST community as well. Next to providing a reference speech recognition platform for the Dutch speaking community, this project also encompasses knowledge transfer between the different partners, hence strengthening the ties between the Netherlands and Flanders, and between research institutions and application developers.

SPRAAK website: <http://www.esat.kuleuven.be/psi/spraak/projects/index.php?proj=SPRAAK>

Combinatorial and Relational Network as Toolkit for Dutch Language Technology (Cornetto)

Project co-ordinator

Prof. dr. P. Vossen
Vrije Universiteit Amsterdam
Onderzoeksgroep Lexicologie/Terminologie
Faculteit der Letteren
De Boelelaan 1105, Kamer 11A-24
Telephone: +31 20-5986466,
E-mail: p.vossen@let.vu.nl
URL: <http://www.let.vu.nl/organisatie/medewerkers.htm>

Project consortium

1. Prof. Dr. W. Martin (Vrije Universiteit Amsterdam)
2. Prof. Dr. M. de Rijke (Universiteit van Amsterdam)
3. Prof. Dr. M.-F. Moens (Katholieke Universiteit Leuven)
4. Prof. Dr. P. Vossen (Irion Technologies BV)

STEVIN funding: € 400.000

Duration: 01/04/2006 – 31/03/2008

Project summary

Cornetto will build a lexical semantic database for Dutch, covering 40K entries, including the most generic and central part of the language and a specialized database for the legal and finance domain. The database will go beyond the structure and content of Wordnet and FrameNet. It will contain both vertical and horizontal semantic relations and combinatorial lexical constraints such as multiword expressions, idioms and collocations on the one hand, and lexical functions and frames on the other. The concepts will be aligned with the English Wordnet so that ontologies and domain labels can be imported. The semantic layer will be validated with a formal ontology, to make it usable in Semantic Web environments. In addition, Cornetto will develop a toolkit for the acquisition of new concepts and relations and the tuning and extraction of a domain specific sub-lexicon from a compiled corpus. A sub-lexicon will be extracted for the legal and finance domain. The lexical database will be evaluated by integration in IR and QA applications and the sub-lexicon will be evaluated by a user-group of language technology companies.

Cornetto website: <http://www.let.vu.nl/onderzoek/projectsites/cornetto/>

Stevin Nederlandstalig Referentiecorpus (SoNaR)

Project co-ordinator

Dr. N. Oostdijk
Radboud Universiteit Nijmegen
Faculteit der Letteren
Centre for Language and Speech Technology (CLST)
Erasmusplein 1
Postbus 9103
NL-6500 HD Nijmegen
Telephone: +31 24 361 27 65
E-mail: n.oostdijk@let.ru.nl
URL: www.let.ru.nl

Project consortium

1. Dr. N. Oostdijk (CLST, Radboud University Nijmegen)
2. Dr. V. Hoste (Dept. Vertaalkunde, Hogeschool Gent (HoGent))
3. Prof. dr. F. de Jong (Human Media Interaction (HMI), Twente University)
4. Dr. M. Reynaert (Induction of Linguistic Knowledge (ILK), Tilburg University)
5. Dr. H. van den Heuvel (CLST, Radboud University Nijmegen)
6. Dr. P. Monachesi (UIL-OTS, Utrecht University)
7. Dr. I. Schuurman, (CCL, Leuven University)

Members advisory Board

Beeken, INL; van den Bosch, Tilburg; Daelemans, Antwerpen; Moens & Van Eynde, Leuven; van Noord, Groningen; Vandeweghe, Gent.

Members User Group

Bouma, Groningen; Boves, Nijmegen; Geeraerts, Leuven; van den Heuvel, Polderland; Iskra, Logica; Jongebloed, Duthear; Odijk, Nuance; de Rijke, Amsterdam; van Veenendaal, HLT Agency; Vossen, Irion; Zuidema, Van Dale

STEVIN funding: € 936.000

Duration: 01/01/2008 – 01/12/2011

Project summary

The project aims at the construction of a 500-million-word reference corpus of contemporary written Dutch for use in different types of linguistic (incl. lexicographic) and HLT research and the development of applications. The project will build on the results obtained in the D-COI and COREA projects which were awarded funding in the first call of proposals within the STEVIN programme. In the light of the budgetary constraints of the present call and the work conducted within other STEVIN projects (especially the LASSY project), the present project will focus on the compilation of the corpus, while the entire corpus will be (automatically) POS tagged and lemmatized by means of the D-COI tagger/lemmatizer. In addition, for a one-million-word subset of the corpus different types of semantic annotation will be provided, viz. named entity labelling, annotation of co-reference relations, semantic role labelling and annotation of spatial and temporal relations. The corpus will be made available through the Dutch HLT Agency (TST-Centrale)

SoNaR website: <http://lands.let.ru.nl/projects/SoNaR/>

**Overview proposals funded in three Calls for demonstration projects
(max. budget 1,0 M€)**

<i>acronym</i>	<i>coordinating SME and other partners</i>	<i>VL/NL consortium nationality</i>	<i>STEVIN priorities addressed</i>	<i>planned duration</i>	<i>funding</i>
Rechtsorde	C-CONTENT b.v. Polderland Language & Speech Technology b.v.	NL	Language technology demonstrator system	15 mnths	€ 90.000
Gemeente-Connect!	IRION Technologies b.v. Duthear b.v. Gemeente Gilze en Rijen	NL	Language and speech technology demonstrator system	10 mnths	€ 60.000
De Kentekenlijn: Spraak-gestuurde Nummerbord Retrieval Tool	Politie Utrecht Duthear b.v.	NL	Speech technology demonstrator system	10 mnths	€ 50.315
Audiokrant	Sensotec NV De Braillekrant vzw KU Leuven-SCD	VL	Speech technology demonstrator system	15 mnths	€ 92.400
PRIMUS: Spelling- en grammatica-controle voor dyslectische gebruikers	Polderland Language & Speech Technology b.v. Technologie & Integratie bvba Die-'s-lekti-kus vzw	NL/VL	Language technology demonstrator system	12 mnths	€ 96.730
Rechtspraak-herkenning	Telecats BV Carp Technologies BV	NL	Language and speech technology demonstrator system	16 mnths	€ 97.000
Klinkende Taal	Gridline BV Utrecht University KU Leuven STIL, Tilburg Provincie Brabant Gemeente Den Haag	NL	Language technology demonstrator system	15 mnths	€ 92.200
SpelSpiek	INL ELITECH Polderland Language & Speech Technology b.v. Van Dale Lexicografie	VL/NL	Language technology demonstrator system	12 mnths	€ 72.387
Web Assess	Telecats BV VO Consulting	NL	Speech technology demonstrator system	5 mnths	€ 45.000

<i>acronym</i>	<i>coordinating SME and other partners</i>	<i>VL/NL consortium nationality</i>	<i>STEVIN priorities addressed</i>	<i>planned duration</i>	<i>Funding</i>
Alfabetisering Anderstaligen Plan (AAP)	Polderland Language & Speech Technology b.v. BEMO-materiaalontwikkeling Uitgeverij Boom Radboud University Nijmegen	NL	Language technology demonstrator system	15 mnths	€ 56.000
Your News	Irion Technologie BV Carp Technologies MD Info	NL	Language technology demonstrator system	6 mnths	€ 19.800
Hulp bij Auditieve Training na Cochleaire Implantatie (HATCI)	Advanced Bionics NV ONICI KU Leuven	VL	Speech technology demonstrator system	8 mnths	€ 48.482
Nederlands-talige Ondertiteling (Neon)	Telecats bv, Vlaamse Radio en Televisie, Ned. Publi. Omroep, K.U.Leuven - ESAT/PSI, Universiteit Gent - ELIS, Universiteit Antwerpen - CNTS	NL/VL	Language and speech technology demonstrator system	12 mnths	€ 85.730
Sprekende zelfcorrigerende woordvoorspeller voor dyslectische gebruikers (WooDy)	Sensotec NV Lexima bv	VL	Language technology demonstrator system	15 mnths	€ 90.000

Rechtsorde

STEVIN funding: € 90.000

Consortium partners

1. C-CONTENT b.v., contact person: Marcel Mooren, m.mooren@c-content.nl
2. Polderland Language & Speech Technology b.v., contact person: Wilko Apperloo

Duration: 01/01/2006 – 01/04/2007 (15 months)

Project summary (in Dutch)

De Nederlandse overheid is er de laatste jaren meer en meer toe overgegaan om elektronische informatie op het gebied van wet- en regelgeving (W&R) publiek toegankelijk te maken. Helaas wordt deze informatie verspreid over vele (niet gestandaardiseerde) websites van de overheid gepubliceerd. Dit maakt het haast onmogelijk voor een professionele gebruiker om de gezochte informatie snel boven water te krijgen. Er is daarom grote behoefte aan één centrale ingang waar alle openbare W&R informatie volledig en snel doorzocht kan worden. C-CONTENT is begin 2005 in dit "gat" gestapt en heeft een systeem "Rechtsorde.nl" gebouwd dat dagelijks, (geautomatiseerd) alle wet- en regelgeving informatie vergaart van verschillende vrij toegankelijke overheidssites en deze informatie vervolgens middels één portaal, www.rechtsorde.nl, doorzoekbaar maakt. Rechtsorde.nl is gericht op de professionele eindgebruiker en bevat o.a. wetten, jurisprudentie, CAO's, ministeriele regelingen, officiële publicaties, verordeningen van lokale overheden etc. In dit demonstratieproject zal de zoekfunctionaliteit van Rechtsorde.nl uitgebreid worden met tal van taalondersteunende gereedschappen van Polderland. Het doel is dat de gezochte documenten gebruiksvriendelijker en efficiënter gevonden kunnen worden en dat de gebruiker middels suggesties meer geholpen wordt bij het vinden van de juiste documenten.

Project summary (in English)

Twenty years of experience in the field of language technology is embedded in RECHTSORDE.NL, the Dutch information portal for legal professionals. C-CONTENT is one of the longest standing suppliers of electronic publishing and information retrieval solutions and the initiator of Rechtsorde b.v. The demonstrator system is a portal, accessible via the internet, that provides user-friendly access to information about laws and local and government regulations as available in official legal publications.

Rechtsorde website and demonstration (in Dutch): <http://www.rechtsorde.nl/>

GemeenteConnect!

STEVIN funding: € 60.000

Consortium partners

1. Irion Technologies BV, contact person: Joop van Gent: gent@irion.nl
2. Dutchear BV, contact person: Victor Huisman
3. Gemeente Gilze en Rijen, contact person: Frank Meulendijks

Duration: 01/01/2006 – 01/11/2006 (10 months)

Project summary (in Dutch)

Gemeentes in Nederland werken aan een overbrugging van de kloof tussen overheid en burger. Zij kampen echter alle met een groot probleem: de hoeveelheid vragen die telefonisch of in direct baliecontact op ze afkomen is dermate groot dat de vraag vaak de capaciteit overstijgt. Het project GemeenteConnect! wordt opgezet om aan te tonen dat een slimme combinatie van spraak- en taaltechnologie dit probleem voor een fors deel kan oplossen: de meest voorkomende telefonische burgervragen aan gemeentes moeten ermee kunnen worden afgehandeld. Irion en Dutchear, beide spin-offs van TNO gevestigd in Delft, hebben een systeem ontwikkeld, waarmee via de telefoon interactief en op natuurlijke wijze informatie kan worden opgevraagd uit grote databases, zonder dat de gebruikers steeds met menu-toetsen worden geconfronteerd. De voordelen van het systeem voor een gemeente zijn onder andere:

- (a) geen wachttijden voor de burger;
- (b) geen menu-toetsen;
- (c) het systeem heeft verstand van alle onderwerpen, dus er hoeft niet te worden doorverbonden;
- (d) het systeem is zelflerend, op basis van gesprekken met burgers, en kan dus steeds beter antwoord geven;
- (e) het systeem kan omgaan met emoties;
- (f) het systeem kan ook als digitaal loket op de website worden geplaatst, waardoor een "chat"-functie ontstaat.

Een belangrijk onderdeel van het project betreft PR-werkzaamheden om deze specifieke en succesvolle combinatie van taal- en spraaktechnologie voor gemeentes landelijke bekendheid te geven in zowel Nederland als Vlaanderen.

Project summary (in English)

GemeenteConnect is a phone dialogue system that allows for free speech input, initiated by the caller; it uses a combination of proven state-of-the-art speech recognition, classification and computational linguistics based dialogue management. Users can freely provide information in their own wording, and in an interaction (dialogue) with the user the system combines all pieces of information given by the user to a unique query which leads to a unique answer.

Gemeenteconnect website and demonstration (in Dutch): <http://www.gemeenteconnect.nl/>



De Kentekenlijn: Spraakgestuurde Nummerbord Retrieval Tool

STEVIN funding: € 50.315

Consortium partners

1. Politie Utrecht, contactpersonen: Janneke Huijssoon, René Anker
2. Dutchear BV, contact person: Els Nachtegaal, E.Nachtegal@dutchear.nl

Duration: 15/03/2006 – 30/09/2006 (10 months)

Project summary (in Dutch)

Dutchear ontwerpt in samenwerking met de Politie Utrecht de Nummerbord Retrieval Tool. De Nummerbord Retrieval Tool zorgt ervoor dat agenten van Politie Utrecht altijd op een snelle, gemakkelijke en veilige manier voertuiginformatie kunnen krijgen. Momenteel belt een agent met zijn GSM naar de meldkamer of naar de infodesk, wanneer hij een kentekenplaat wil natrekken. De snelheid waarmee hij geholpen wordt is geheel afhankelijk van de beschikbaarheid van medewerkers op de meldkamer of bij de infodesk. De lijnen zijn echter regelmatig bezet waardoor de wachttijd voor de agent oploopt. De huidige situatie is daarom onwenselijk. Hoe sneller een agent over de relevante informatie beschikt, hoe veiliger de situatie voor hem en de maatschappij is. In de tijd dat de agent moet wachten op de informatie blijft mogelijk de onverzekerde auto doorrijden, of laat de agent een bestuurder van een gestolen auto weggrijpen. Agenten kunnen lopend, op de mountainbike, in de auto en op de motor bellen met de Nummerbord Retrieval Tool (NRT). De agent spreekt het kenteken in en krijgt informatie (naam eigenaar, APK, verzekering, gestolen) over het betreffende voertuig teruggekoppeld via een Text-To-Speech engine (spreekende computer). Naast de terugkoppeling van de informatie door de telefoon ontvangt de agent bovendien een SMS met de aan hem voorgelezen informatie.

Project summary (in English)

Dutchear and the Utrecht police have jointly developed a demonstrator that automating vehicle license plate retrieval using proven state-of-the-art speech recognition technology can lead to improved service with reduced human effort.

For privacy reasons no live demonstration is available, a movie is which the system is re-enacted cab be seen on YouTube: <http://nl.youtube.com/watch?v=1Q54vvkeKGY>

Audiokrانت

STEVIN funding: € 92.400

Consortium partners

1. Sensotec NV, contact person: Frank Allereersch, frank.allemeersch@sensotec.be
2. De Braillekrانت vzw, contact person: Katty Kloeck
3. Katholieke Universiteit Leuven-SCD, contact person: Jan Engelen

Duration: 15/02/2007 – 14/05/2008 (15 months)

Project summary (in Dutch)

Voor personen met een leeshandicap is de toegankelijkheid tot krانتeninformatie allesbehalve evident. Er bestaan dan ook al sinds enkele jaren speciale voorzieningen om deze toegankelijkheid te bewerkstelligen. In Vlaanderen zijn dat de initiatieven Braillekrانت en DiGiKrانت (gecoördineerd door De Braillekrانت vzw), waarbij respectievelijk een extractie van de krانت in Braille en een volledige krانت in digitale vorm wordt aangeboden. Voor het lezen van de krانت in digitale vorm dient men te beschikken over een pc uitgerust met vergrotingssoftware, synthetische spraakoutput en/of een braille leeslijn. De beperking tot lezers met kennis van braille of die kunnen beschikken over pc met extra uitrusting gecombineerd met een voldoende basiskennis in pc gebruik heeft als gevolg dat het gedeelte van de doelgroep dat de krانت kan lezen toch nog vrij beperkt blijft. Anderzijds is er sinds 2004 voor wat betreft de gesproken boeken voor personen met een leeshandicap zowel in Vlaanderen als in Nederland de overstap gemaakt van verspreiding op cassette naar verspreiding op data-CD. Voor de verstrekking op CD maakt men gebruik van de internationale DAISY standaard, waarmee zowel audio als tekst op eenzelfde drager kan geplaatst worden. Voor het beluisteren van de Daisy cd's bestaan er specifieke voorleesapparaten en ongeveer iedere regelmatige gebruiker van gesproken boeken in Vlaanderen en Nederland beschikt ondertussen over zo'n (draagbaar) voorleesapparaat. Het gaat hierbij om een paar tienduizend dergelijke apparaten. Binnen het AudioKrانت project zullen we dagelijks een versie van de krانت produceren die conform is met de Daisy standaard en kan voorgelezen worden met die voorleesapparaten. Vanwege het tijdskritische karakter van de productie van een krانت, is het uitgesloten dat we, zoals voor de productie van gesproken boeken, gaan gebruik maken van voorlezers. Naar onze overtuiging kan de aanwending van spraaktechnologie (synthetische spraak) en hoogtechnologische taaltechnologie (voor de optimalisatie ervan) hier echter de oplossing brengen.

Project summary (in English)

In this project the daily Belgian newspapers "Het Nieuwsblad" and "De Standaard" will be made available in Daisy format. This format contains both the written information and a spoken version. The demonstrator system uses proven speech technology (synthetic speech) and language technology to automatically produce a spoken version of the daily newspaper.

Audiokrانت website and demonstration (in Dutch): http://www.braillekrانت.be/nieuws_detail.php?nr=13

PRIMUS: Spelling- en grammaticacontrole voor dyslectische gebruikers

STEVIN funding: € 96.730

Consortium partners

1. Polderland Language & Speech Technology bv, contact person: Inge de Mönnink, inge@polderland.nl
2. Technologie & Integratie b.v.b.a., contact person: Jo Cremelie
3. Die-'s-lekti-kus vzw, contact person: Dirk Callebaut

Duration: 15/02/2007 – 14/02/2008 (12 months)

Project summary (in Dutch)

Het resultaat van dit project is een spellingcontrole en een grammaticacontrole aangepast voor dyslectische gebruikers. De standaard spelling- en grammaticacontrole in Microsoft® Office worden in dit project zodanig aangepast dat ze beter aansluiten bij de typische fouten die dyslectische gebruikers maken (bijvoorbeeld 'eemoscho nele' i.p.v. 'emotionele' en 'brugste' i.p.v. 'beruchtste'). Bovendien wordt aan de spelling- en grammaticacontrole de mogelijkheid toegevoegd om suggesties voorgelezen te krijgen door een spraaksyntheseprogramma. Omdat dyslectische gebruikers behalve spellingproblemen ook leesproblemen hebben ondersteunt de combinatie van een aangepaste spellingcontrole en een spraaksyntheseprogramma de dyslectische gebruiker maximaal in hun schrijfproces. Als laatste wordt ook de interface van de spellingcontrole aangepast op dyslectische gebruikers. Het project richt zich op dyslectische kinderen. Hierdoor kan het product al in de onderwijssituatie optimaal worden ingezet en zal het aantal kinderen dat door hun taalbeperking in het onderwijs buiten de boot valt verder beperkt kunnen worden. Aangezien de spelling- en grammaticacontrole ingebed zitten in Office en dus onder andere te gebruiken zijn in Word en Outlook is het eindresultaat ook zeer nuttig te gebruiken door volwassen dyslectische en door anderen meet een taalbeperking zoals niet-moedertaalsprekers van het Nederlands, slechtzienden en kinderen met leerproblemen.

Project summary (in English)

The result of this demonstrator project is a specialised version of Microsoft's spelling and grammar checking software for Dutch adapted for dyslectic users. The system has knowledge about errors made by this type of users and also allows suggestions for corrections to be read aloud.

Spelling- en grammaticacontrole voor dyslectische gebruiker website: www.polderland.nl

Rechtspraakherkenning

STEVIN funding: € 97.000

Consortium partners

1. Telecats BV, contact person: W. Luimes, w.luimes@telecats.nl
2. Carp Technologies BV, contact person: D. Lie

Duration: 15/02/2007 – 14/06/2008 (16 months)

Project summary (in Dutch)

Rechtbanken in Nederland zien zich in toenemende mate verplicht om de geluidsopnamen in de rechtszaal volledig uit te schrijven. Met behulp van bestaande taal- en spraaktechnologie is het mogelijk hulpmiddelen te ontwikkelen die de tijd die gemoeid is met het uitschrijven van gesproken geluidsopnamen, aanzienlijk kan verkorten. Bovendien kan vervolgens op relatief eenvoudige wijze de eenmaal uitgeschreven tekst doorzoekbaar worden gemaakt zodat gevonden passages dmv een muisklik ook beluisterbaar worden. Bijkomend voordeel van het inzetten van deze technologie is dat daarmee een goede basis wordt gelegd voor additionele toepassingen en innovaties, zoals bijvoorbeeld het (semi-) automatisch samenvatten van conversaties. Centraal in dit voorstel is dat technologie hier moet worden ingezet als hulpmiddel en niet als substitutie. Dat houdt in dat het werk nog steeds door (dezelfde) mensen wordt gedaan, maar dat door het inzetten van hulpmiddelen de benodigde tijd en dus werkdruk sterk verlaagd wordt.

Project summary (in English)

Courts in the Netherlands are increasingly required to transcribe the recordings made in the courtroom. This project aims to develop tools that can significantly cut down the amount of time needed to do so by using existing proven language and speech technology. These transcriptions will not be perfect but will reduce human effort in producing the full transcriptions. Additionally, the project will make the recordings searchable.

Rechtspraakherkenning website (in Dutch) <http://www.telecats.nl/klanten/Rechtbank/>

For privacy reasons no live demonstration will be made available, a movie in which the system is re-enacted is available on youtube: <http://www.youtube.com/watch?v=Ti9pMVhEsAo>

Klinkende Taal

STEVIN funding: € 92.200

Consortium partners

1. GridLine BV, contact person: Oele Koornwinder, oele@gridline.nl
2. Faculteit der Letteren van de Universiteit Utrecht - UiL OTS, contact person: H. Pander Maat
3. Faculteit Letteren van de Katholieke Universiteit Leuven - Centrum voor Computerlinguïstiek, contact person: Frank Van Eynde
4. Stichting Toepassing Inductieve Leertechnieken, contact person: Antal Van den Bosch
5. Provincie Brabant, contact person: H. Maaskant
6. Gemeente Den Haag - Dienst Voorlichting en Ext. Betrekkingen, contact person: H. De Kievith

Duration: 15/02/2007 – 14/05/2008 (15 months)

Project summary (in Dutch)

Van de Nederlandse overheid wordt in toenemende mate verwacht dat zij klare taal spreekt. Overheidsinstellingen produceren veel publieksgerichte teksten, in brochures en brieven en op websites. De leesbaarheid van de publieksgerichte communicatie kan worden verbeterd door de teksten van ambtelijk jargon te ontdoen. Het demonstratieproject speelt in op deze opgave door een dynamische jargon-bewaker op de markt te brengen. Het betreft een op maat aangeboden toepassing, die overheidsinstellingen in staat stelt hun teksten begrijpelijker te maken, namelijk door de opsporing en vervanging van termen die de doelgroep als jargon zal ervaren. Deze dynamische Jargonbewaker onderscheidt zich van bestaande woordkeuzetools doordat hij zich automatisch aanpast aan het kennisdomein van de organisatie en de doelgroep, alsmede aan de veranderingen die hierin optreden. De tool wordt aangeboden in een laagdrempelige vorm die aansluit op de bestaande werkwijze van de gebruiker. Het project richt zich speciaal op jargon-bewaking in publieksteksten van de lagere overheid, te weten provincies en gemeenten. Om deze lagere overheden te overtuigen van het nut van de applicatie zal een Jargonbewaker-op-maat worden gebouwd voor twee proefgebruikers, te weten de provincie Brabant en de gemeente Den Haag. De effectiviteit van deze demonstrators wordt aangetoond door middel van een leesexperiment met proefpersonen. Het project voorziet tot slot in een grootscheeps marketing-offensief, waarbij overheidsinstellingen en communicatie-adviesbureaus via presentaties en workshops kennis zullen maken met de doeltreffendheid van automatische jargon-opsporing.

Project summary (in English)

A dynamic jargon detection system is developed which automatically adapts to an organisation or target group. The demonstrator will be tested on public texts produced by local government organisations.

Klinkende Taal website and demonstration (in Dutch): <http://www.klinkendetaal.nl/>

SpelSpiek

STEVIN funding: € 72.387

Consortium partners

1. Instituut voor Nederlandse Lexicologie, dependance Vlaanderen, contact person: Katrien Van pellicom, pellicom@inl.nl
2. Elitech, contact person: J. Brouwers
3. Polderland Language & Speech Technology bv, contact person: Inge de Mönnink
4. Van Dale Lexicografie bv, contact person: Johan Zuidema

Duration: 15/02/2007 – 14/02/2008 (12 months)

Project summary (in Dutch)

Op 1 augustus 2006 is de nieuwe spelling ingegaan. De spellingregels en meest recente bijstellingen aan die regels zijn lang niet bij iedereen bekend. Vooral jongeren zijn vaak niet op de hoogte van de spellingregels, maar ook de professionele taalgebruiker heeft wel eens zijn twijfels over de manier waarop je een bepaald woord moet schrijven. Er bestaan al verschillende kanalen via welke je de spelling van woorden kunt opzoeken, of de officiële regels van de spelling van de Nederlandse taal kunt bestuderen. De Taalunie heeft een website waar je de woorden uit de Woordenlijst van de Nederlandse Taal kunt opzoeken, en waar je de regels kunt lezen. Het Groene Boekje bestaat bovendien zowel in boekvorm als op cd-rom, en er is bovendien een elektronische versie van het Groene Boekje gratis online beschikbaar. Dynamische communicatiemiddelen als MSN en sms zijn erg populair, vooral onder jongeren. Het hierboven beschreven project maakt het mogelijk om deze communicatiemiddelen te gebruiken als spellinghulp, door het inzetten van een chatbot. Dat is een robot waarmee je via MSN kunt chatten. In dit geval is het een spellingchatbot: je kunt er bijvoorbeeld aan vragen: "Hoe spel je bjoetiekees?" De chatbot geeft dan direct het juiste antwoord terug. Op die manier heb je een snelle feedback over de juiste spelling van een woord. Zowel achter de computer als onderweg, want dezelfde service stellen we ook via sms beschikbaar. Daarnaast is de service ook gewoon via de webbrowser te bereiken. Drie moderne, populaire communicatiemiddelen dus. Bovendien wordt de bot door de tijd heen slimmer: woorden die de bot niet kent (of foutieve spellingen daarvan), worden bekeken door een spellingdeskundige, waarna die informatie wordt toegevoegd aan de bot. Op die manier wordt hij dus steeds beter in het corrigeren van woorden.

Project summary (in English)

Many youngsters and preadolescents have troubles with spelling, but also professional language users regularly have questions concerning the spelling of certain words and could benefit from an easily accessible tool. Several different ways to check the spelling of words already exist, but SpelSpiek adds a dynamic feature: you can use SpelSpiek through modern communication interfaces like web browsers, MSN or SMS. Moreover, it is possible to ask spelling questions using natural language. SpelSpiek uses lexical data provided by INL and Van Dale. The SpelSpiek software integrates spelling correction software by Polderland and chatbot software by Elitech

SpelSpiek website and demonstration (in Dutch): <http://www.spelspiek.nl/>

Web Assess

STEVIN funding: € 45.000

Consortium partners

1. Telecats BV, contact person: W. Luimes, w.luimes@telecats.nl
2. VO Consulting, contact person: Geert van Ouwkerk

Duration: 15/02/2007 – 14/07/2007 (5 months)

Project summary (in Dutch)

Bedrijven besteden erg veel tijd en geld aan het selecteren van geschikte kandidaten voor het werken in call centers omdat slechts 10% van degenen die zich aanmelden daadwerkelijk geschikt blijkt te zijn. Een goede automatische voorselectie geeft bedrijven de mogelijkheid om meer tijd en aandacht te besteden aan de geschiktheid van de geselecteerde kandidaten. Om dit te kunnen doen wordt een applicatie gemaakt die geheel automatisch een (min-of-meer voorgebakken) conversatie met de kandidaten aangaat.

Spraakherkenning wordt gebruikt om te meten of bepaalde essentiële woorden wel of niet gezegd zijn. De dialoog verloopt op basis van de gegeven antwoorden omdat een vraag nogmaals (op een andere wijze) wordt gesteld wanneer één of meerdere sleutelwoorden ontbreken. De kandidaten die door het systeem gebeld worden, moeten eerst een reeds bestaande web-applicatie met goed gevolg doorlopen hebben. Deze web-applicatie die een gedegen uitleg geeft over het werken in het call center, is er op gericht de kandidaten te testen op hun kennis van de verschillende telefoniesystemen die ze gaan gebruiken. Als de kandidaten de web-applicatie met goed gevolg doorlopen hebben, kunnen ze het telefoonnummer invullen waarop ze bereikbaar zijn. De hier voorgestelde applicatie gaat ze dan op dat nummer bellen en begint dan de dialoog. Op deze gecombineerde manier (web en telefonie) kunnen veel kandidaten snel en tegen geringe kosten beoordeeld worden op hun mogelijke geschiktheid om als call center medewerker aan de slag te gaan. De applicatie is dus bedoeld voor de voorselectie om het kaf van het koren te scheiden. De eigenlijke selectie gebeurt daarna op de "ouderwetse" manier.

Project summary (in English)

Companies spend much time and money on selecting suitable candidates for working in call centres, given that only 10% of the applicants prove to be suitable for the job. This demonstrator project aims to develop a tool that allows for the automatic preselection of candidates that uses both proven speech recognition and a web application.

Web Assess website and demonstration (in Dutch): <http://www.webassess.nl/>

Alfabetisering Anderstaligen Plan (AAP)

STEVIN funding: € 56.000

Consortium partners

1. Polderland Language & Speech technology bv, contact person: Peter Beinema, peter@polderland.nl
2. BEMO-materiaalontwikkeling, contact person: Ad Bakker
3. Uitgeverij Boom, contact person: Geert van der Meulen
4. Radboud Universiteit Nijmegen, contact person: I. Van de Craats

Duration: 01/04/2008 – 30/06/2009 (15 months)

Project summary (in Dutch)

Dit project implementeert een demonstrator die bestaande spraaktechnologie toepast in het kader van alfabetisering. Hierbij is onmiddellijke feedback essentieel. De methode AAP (alfabetisering anderstaligen plan) wordt hiervoor gevolgd. De technologie zal kunnen geïntegreerd worden in toepassingen van derden.

Project summary (in English)

A demonstrator system is implemented that uses proven speech technology to produce feedback to second language learners. The system will be integrated and tested in an existing language learning application.

Your News (voorheen Easy Info)

STEVIN funding: € 19.800

Consortium partners

1. Irion Technologies bv, contact person: Joop van Gent, gent@irion.nl
2. Carp Technologies, contact person: Danny Lie
3. MD Info contact person: Bert Ponsen

Duration: 15/02/2008 – 14/08/2008 (6 months)

Project summary (in Dutch)

In het kader van de nieuwsvoorziening is er een tendens naar dienstverlening zoals “news brokers” of knipseldiensten. Klanten van deze dienstverlening kunnen een profiel opgeven in de vorm van trefwoorden. Dat profiel wordt dan gebruikt om een selectie te maken uit de actuele nieuwsberichten. Het aanmaken van profielen op basis van trefwoorden vereist veel handwerk en de “matching” blijft laag. Automatische methoden daarentegen falen dikwijls omdat er gebruik gemaakt wordt van eenvoudige zoektechnologie of statistische methodes. Dit project zal een betere “matching” verwezenlijken. Als demo koppelt men een classificatiesysteem en een samenvattingsgenerator aan het standaardplatform van een aanbieder van gepersonaliseerde informatie. Met behulp van een testgroep worden er evaluaties uitgevoerd om de kwaliteit van het systeem te testen.

Project summary (in English)

To get information “news brokers” or “news clipping services” are regularly made use of. Customers of these services have to define a profile by adding keywords. The profile is used to determine a selection of news items. The project aims at using proven language technology to help create better profiles and reduce effort in producing them. The system will be evaluated on a real user group.

Hulp bij Auditieve Training na Cochleaire Implantatie (HATCI)

STEVIN funding: € 48.482

Consortium partners

1. Advanced Bionics NV, contact person: Filiep Vanpoucke, filiepv@abionics.fr
2. ONICI contact person: Leo De Raeve
3. K.U.Leuven - ESAT/PSI, contact person: Hugo Van hamme

Duration: 01/04/2008 – 30/11/2008 (8 months)

Project summary (in Dutch)

Tijdens dit project wordt een applicatie gebouwd die m.b.v. een automatische spraakbeoordeling een therapeut ondersteunt bij het toepassen van de "speech tracking" als hoortherapie en -evaluatie bij revalidatie na cochleaire implantatie. Na cochleaire implementatie dient de patiënt te leren spreken en horen met zijn nieuwe implantaat. De doelgroep zijn vooral patiënten die reeds tot een goede articulatie komen, maar voor wie de hoornauwkeurigheid, het taalgevoel en de grammaticaverwerving verder gestimuleerd moeten worden. De demonstrator zal vooraf opgenomen teksten aan de patiënt aanbieden en hij/zij moet de tekst herhalen. De correctheid van deze herhaling wordt beoordeeld d.m.v. automatische spraakherkenning revalidatiestap.

Project summary (in English)

This project aims at building an automatic speech assessment system that can support a speech therapist in helping a patient with a cochlear implant to learn to speak.

Nederlandstalige Ondertiteling (NEON)

STEVIN funding: € 85.730

Consortium partners

1. Telecats bv, contact person: Michel Boedeltje, m.boedeltje@telecats.nl
2. Vlaamse Radio en Televisie, contact person: Bernard Dewulf
3. Nederlandse Publieke Omroep, contact person: Jurgen Lentz
4. K.U.Leuven - ESAT/PSI, contact person: Patrick Wambacq
5. Universiteit Gent - ELIS, contact person: Jean-Pierre Martens
6. Universiteit Antwerpen - CNTS, contact person: Walter Daelemans

Duration: 01/04/2008 – 31/03/2009 (12 months)

Project summary (in Dutch)

In dit project zal een geavanceerde en minder arbeidsintensieve spraakherkenningstoepassing geïmplementeerd worden voor ondertiteling van televisieprogramma's, met name gerealiseerd door het gecondenseerd aligneren van bestaande teksten of scripts met gesproken audio. Dit zal leiden tot een (semi-)automatische ondertiteling in het Nederlands. Dit gebeurt m.b.v. een spraakherkenningssysteem, waardoor automatisch rechtstreekse transcriptie van de audiostream (het resultaat van de spraakherkenning) altijd in de achtergrond aanwezig is om op terug te vallen.

Project summary (in English)

This project will implement a less labour intensive application of speech recognition for television subtitling, in particular by using condensed alignment of existing texts or scripts with the speech audio. This should lead to (semi-)automatic subtitling in Dutch. The use of speech recognition provides an automatic direct transcription of the speech in the background, to replace the text or script as a fall back.



Spreekende zelfcorrigerende woordvoorspeller voor dyslectische gebruikers (WooDy)

STEVIN funding: € 90.000

Consortium partners

1. Sensotec NV, contact person: Frank Allemeersch, frank.allemeersch@sensotec.be
2. Lexima bv, contact person: Ria Janssen

Duration: 15/02/2008 – 14/05/2009 (15 months)

Project summary (in Dutch)

Dit project bouwt een spreekende zelfcorrigerende woordvoorspeller voor dyslectische gebruikers d.m.v. van een combinatie van zelfcorrectie en woordvoorspelling. De kern bestaat uit de ontwikkeling van een basisset van woordenlijsten waaruit voorspelling wordt afgeleid, en van algoritmes ter bepaling van welke woorden aangereikt zullen worden rekening houdend met persoon-specifieke beperkingen. Dit alles wordt geïmplementeerd en gedemonstreerd met een prototype spreekende woordvoorspeller. Doelgroepen zijn individuele gebruikers met lees- en taalbeperkingen, en omkaderende dienstverlening.

Project summary (in English)

A self-correcting speaking word prediction system is built for dyslectic users by utilising proven language and speech technology.

**Overview proposals funded in educational/masterclass projects
(max. budget 192 k€)**

<i>acronym</i>	<i>coordinating institute and other partners</i>	<i>VL/NL consortium nationality</i>	<i>STEVIN priorities addressed (subject)</i>	<i>planned duration</i>	<i>funding</i>
Vooronderzoek TST in het voortgezet onderwijs**	Stichting Studio Taalwetenschap	NL	raising awareness for HLT	9 months	€ 15.000
<i>educational</i>					
TST op Kennislink	Utrecht University Kennislink	NL	raising awareness for HLT	12 months	€ 27.500
DiaDemo	KU Leuven – ESAT-PSI Technopolis	VL	raising awareness for HLT	8 months	€ 32.113
TST op Kennislink2	Utrecht University Kennislink	NL	raising awareness for HLT	12 months	€ 25.500
<i>masterclass</i>					
ICT & Dyslexie	Dedicon Expertise Centrum Nederland	NL	raising awareness for HLT		€ 27.500
TST voor Nederlandstalige overheidsdiensten	Gridline Telecats BV	NL	raising awareness for HLT		€ 19.000

** project funded in pre-call activity.

TST-pagina's voor Kennislink

STEVIN funding: € 27.500 (educa project)

Consortium partners

1. Landelijke Onderzoekschool Taalkunde (LOT), contact person: Mw. drs. M.M. Jansen, redacteur taalwetenschappen, Mathilde.Jansen@kennislink.nl
2. Kennislink (Stichting Nationaal Centrum voor Wetenschap en Technologie): contact person R. Smallenburg, manager & projectleider, ronald@ijburglaan.nl

Duration: 01/03/2008 – 28/02/2009 (12 months)

Project summary (in Dutch)

Het project 'Taal- en Spraaktechnologie op Kennislink' bestaat uit twee componenten:

1. Het populariseren van beschikbaar materiaal uit het vakgebied. Uit eerdere gesprekken tussen A. van Hessen (Notas) en Kennislinkredacteur M. Jansen is gebleken dat er veel materiaal voorhanden is binnen de Taal- en Spraaktechnologie dat geschikt is om toegankelijk gemaakt te worden voor een breed publiek. Als voorbeeld kunnen de artikelen uit het tijdschrift over Toegepaste Taal- en Spraaktechnologie, Dixit, genoemd worden. Het onderzoek binnen de TST leent zich erg goed voor popularisering, omdat veel van de onderwerpen een groot publiek aanspreken. De redacteur TST zal zich daarom voornamelijk bezig houden met het beschikbaar maken binnen Kennislink van reeds voorhanden materiaal. Daarnaast zal een netwerk van correspondenten binnen de Taal- en Spraaktechnologie worden opgezet. Van de correspondenten wordt gevraagd artikelen aan te dragen over het eigen onderzoek, dat door de redacteur TST zal worden geredigeerd. Naast het schrijven en redigeren van artikelen, houdt de redacteur zich bezig met het samenstellen van themagestuurde dossiers. De dossiers vormen een aparte categorie op Kennislink, en vormen een introductie op een bepaald thema. De dossiers worden vooral gebruikt door scholieren voor thema- en profielwerkstukken.

2. Een verkennend onderzoek naar de aansluiting van TST op een brede doelgroep. De doelgroep van Kennislink bestaat uit een gevarieerd publiek, van scholieren tot beleidsmakers. Om de aansluiting van TST op deze markt te verkennen, schrijft de redacteur TST-nieuwsberichten over recente ontwikkelingen binnen het vakgebied. Binnen het jaar waarin de redacteur TST is aangesteld, zullen regelmatig evaluatiemomenten worden ingebouwd, waarin bepaald moet worden of de artikelen aanslaan bij de verschillende doelgroepen van Kennislink. Hieruit zal worden afgeleid op welke wijze Taal- en Spraaktechnologie het beste een blijvend onderdeel kan gaan uitmaken van de vakpagina Taalwetenschappen (mogelijk in combinatie met de vakpagina Techniek).

Project summary (in English)

Kennislink is a popular website used by students and teachers to find information about recent scientific developments. Information about the state-of-the-art in human language technology will be added.

Kennislink website: <http://www.kennislink.nl/web/show>

DiaDemo: Dialectenherkenner en -demonstrator

STEVIN funding: € 32.113 (educa project)

Consortium partners

1. K.U. Leuven – ESAT-PSI, contact person: prof. dr. D. Van Compernelle, Dirk.VanCompernelle@esat.kuleuven.be
2. Technopolis.

Duration: 01/10/2008 – 31/05/2009 (8 months)

Project summary (in Dutch)

Het DIADEMO-project bouwt een demonstrator die gesproken dialecten herkent. Deze demonstrator zal worden opgesteld in Technopolis (Mechelen). Technopolis, het Vlaams doe-centrum voor wetenschap en technologie, krijgt jaarlijks ca. 280.000 bezoekers over de vloer, schoolgroepen zowel als families. Op deze manier wil DIADEMO de resultaten uit het spraakonderzoek op een speelse wijze toegankelijk maken voor een breed publiek in Vlaanderen. Meer infoamantie is te vinden in de presentatie te vinden op: http://taalunieversum.org/taal/technologie/stevin/documenten/diademmo_04092009.pdf .

Project summary (in English)

DiaDemo is a demonstrator system that recognises spoken Flemish dialects. The demonstrator is available at Technopolis. Technopolis is a Flemish interactive science center in Mechelen, which annually has about 280.000 visitors, school children and families. Via DiaDemo results from speech technology will be made accessible to a wide audience in Flanders.

DiaDemo information on de Technopolis website:

<http://www.technopolis.be/nl/?n=1&e=21&s=168&exhibit=341&&thema=4>

Project: TST op Kennislink 2

STEVIN funding: € 25.500 (educa project)

Consortium partners

1. Landelijk Onderzoekschool Taalwetenschap (LOT),
contactpersonen: Mathilde Jansen (aanvrager, Mathilde.Jansen@kennislink.nl) Erica Renckens
(uitvoerder, Erica.Renckens@kennislink.nl)
2. Kennislink, contactpersoon Carl Koppenschaar (hoofdredacteur Kennisling, Carlkop@xs4all.nl)

Duration: 01/12/2009 – 30/11/2010 (12 months)

Project summary (in Dutch)

Dit project is een voortzetting van het in 2008 gehonoreerde STEVIN-project 'Taal- en spraaktechnologie op de populair-wetenschappelijke website Kennislink'. Uit de evaluatie van dit project is gebleken dat de toevoeging van TST-artikelen aan de vakpagina's Taal & Spraak en Techniek succesvol is. Een continuering van de aanstelling van de redacteur TST is wenselijk om ruime aandacht voor taal- en spraaktechnologie op Kennislink te kunnen garanderen. Kennislink wordt in opdracht van het Ministerie van Onderwijs, Cultuur en Wetenschap uitgevoerd door Stichting Nationaal Centrum voor Wetenschap en Technologie. Kennislink maakt wetenschappelijke informatie toegankelijk voor een breed publiek. Vooral middelbare scholieren en studenten behoren tot de doelgroep. Sinds haar onlinegang op 15 april 2002 is Kennislink met inmiddels gemiddeld 12.000 unieke bezoekers per dag uitgegroeid tot één van de meest bezochte populairwetenschappelijke websites in het Nederlandse taalgebied. In april 2009 is de website en het CMS van Kennislink geheel vernieuwd. Sindsdien is de site interactiever, kan multimedia makkelijker geplaatst worden en kunnen bezoekerscijfers nauwkeuriger worden bijgehouden.

De redacteur TST zal gedurende dit project ook worden ingezet als eindredacteur van het STEVINproject 'TST op Wikipedia', en zal in die functie een belangrijke bijdrage leveren aan het Wikipediaproject. Met de aanwezigheid van een eindredacteur in dit project kunnen de artikelen over taal- en spraaktechnologie op Wikipedia een eenduidige structuur krijgen en is een hoge kwaliteit gegarandeerd.

Contactpersoon "TST op Wikipedia": Arjan van Hessen, Universiteit Twente, A.J.vanHessen@utwente.nl.

Project summary (in English)

Kennislink is a popular website used by students and teachers to find information about recent scientific developments. Information about the state-of-the-art in human language technology will be added. Information about HLT will also be transferred to Wikipedia.

Kennislink website: <http://www.kennislink.nl/web/show>

ICT & Dyslexie

STEVIN funding: € 17.500 (Masterclass)

Consortium partners

1. Dedicon, contact person: Mw. drs. I. de Mönnink: ldmonnik@dedicon.nl
2. Expertisecentrum Nederlands, contact person: Evelien Krikhaar

Duration: 01/08/2009 – 01/10/2010 (14 months)

Project summary (in Dutch)

Voor kinderen met dyslexie is lezen, en daarmee leren, een probleem. Er zijn vele ICT-hulpmiddelen beschikbaar om deze kinderen te ondersteunen in het onderwijs. Veel van deze producten bevatten taal- en/of spraaktechnologie. De beschikbare hulpmiddelen worden tot nu toe slechts beperkt ingezet in het onderwijs. Docenten zijn onvoldoende geïnformeerd over het bestaan van de producten en hebben behoefte aan voorbeelden van goed gebruik. De masterclass ICT en Dyslexie geeft een overzicht van beschikbare hulpmiddelen, stelt leerkrachten uit primair en voortgezet onderwijs in de gelegenheid zelf met de hulpmiddelen aan de slag te gaan en stimuleert leerkrachten door succesverhalen van collega's vanuit de praktijk.

Project summary (in English)

The Masterclass ICT & Dyslexie will increase awareness of the available language and speech technology tools that can support the education of children with reading disabilities.

TST voor Nederlandstalige overheidsdiensten

STEVIN funding: € 19.000 (Masterclass)

Consortium partners

1. Gridline, contact person: Dr. O. Koornwinder: oele@gridline.nl
2. Telecats BV, contact person, DR. A.J. van Hessen, arjan@telecats.nl.

Duration: 01/01/2010 – 31/08/2010 (8 months)

Project summary (in Dutch)

GridLine en Telecats organiseren in samenwerking met Business Universiteit Nyenrode een Master Class over Taal- en Spraaktechnologie (TST) voor Nederlandstalige Overheidsdiensten. De Master Class richt zich op bestuurders en beleidsmanagers in het Openbaar bestuur en de hieraan verbonden publieksdiensten (van de Belastingdienst en het UWV tot Politie en Justitie). Na afloop zullen de deelnemers een goed beeld hebben van de wijze waarop Taal- en Spraaktechnologie hun bestaande dienstverlening kan ondersteunen, verbeteren en uitbreiden. Het doel is dus markteducatie.

In de Master Class zullen de deelnemers in één dag kennismaken met de stand van zaken in de Taal- en Spraaktechnologie (TST) en haar toepassingsmogelijkheden. De Master Class bestaat uit op de doelgroep toegesneden expertcolleges over belangrijke basismethodes en hun toepassingsmogelijkheden, praktijkpresentaties (met aandacht voor producten, Business Cases, do's en don'ts, implementatietrajecten en praktijkstories) en hands on practica. De besproken toepassingen zullen live of door middel van filmpjes worden getoond en betreffen drie hoofdthema's: Tekstanalyse, Spraakanalyse en Zoeken. Hierbij zullen niet alleen TST-methodes aan de orde komen (van parseren en lexicale gegevensextractie tot signaalconversie en spraaksynthese), maar ook methodes uit verwante disciplines (Machine Learning, Classificatie, Information Retrieval en Multimedia-analyse).

Centrale vragen:

- Wat is Taal- en Spraaktechnologie en waarvoor kun je het gebruiken?
- Wordt het al gebruikt en zo ja wat is de ervaring hiermee?
- Hoe moet een organisatie het invoeren?
- Wat zijn voorwaarden voor een geslaagde business case?

De Master Class vindt plaats op Business Universiteit Nyenrode, krijgt een marktconforme inschrijfprijs van €450 en wordt afgesloten met een diner. De Class biedt plaats aan maximaal 20 deelnemers, waarvan 5 plaatsen worden gereserveerd voor deelnemers op uitnodiging (bijv. voor vertegenwoordigers van stichtingen met een beperkt budget). Na afloop zullen de cursusmaterialen online beschikbaar worden gesteld aan de deelnemers en andere belangstellenden.

Het idee is om de Master Class meerdere keren te herhalen en daarbij steeds op een andere sector mikken (o.a. de Juridische Sector, de Financiële Sector en de Zorgsector). Door beslissers inzicht te geven in de concrete mogelijkheden en kansen die taal- en spraaktechnologie biedt, denken wij het speelveld voor ons vak aanzienlijk te verruimen en aldus steeds meer aandacht te genereren.

onderzoek & ontwikkeling

STATUS per 31 december 2009

 = formeel afgerond

 = formele afronding gestart

 = voortgangsverslag moet aangevuld

 voortgang goedgekeurd

STE	titel	PL	PF houder	totaal budget	voorschotten	saldo	agreement	voortgang	start	formeel einde	oplevering	werkelijk einde
04014	AUTONOMATA	JP.Martens	Bruwaene, Hessen	322.848	322.848	0	OK	mei-07	1-6-2005	31-5-2007	OK	5-31-2007
04005	COREA	G.Bouma	Sas, Eynde	353.875	353.875	0	OK	okt-07	1-5-2005	30-4-2007	OK	10-31-2007
04008	D-Coi	N.Oostdijk	Odijk, Veenendaal	566.531	566.531	0	OK	dec-06	1-6-2005	31-12-2006	OK	12-31-2006
04019	IRME	J.Odijk	Eynde, Kenyon-Jac	389.500	389.500	0	OK	aug-07	1-6-2005	31-8-2007	OK	8-31-2007
04017	JASMIN-CGN	C.Cucchiarini	Martens, Kenyon-J	419.471	419.471	0	OK	dec-07	1-4-2005	30-9-2007	OK	12-31-2007
	TOTAAL			2.052.225	2.052.225	0						

2de oproep

tender

STE		PL	PF houder	totaal budget	voorschotten	saldo	agreement	voortgang	start	formeel einde	oplevering	werkelijk einde
05038	SPRAAK	P.Wambacq	Martens, Sas	400.000	400.000	0	OK	jun-08	1-2-2006	31-5-2008	OK	
05039	CORNETTO	P. Vossen	Eynde, Veenendaal	400.000	300.000	100.000	OK	jul-08	1-4-2006	31-3-2008	OK	
	TOTAAL			800.000	700.000	100.000						

2de oproep

open oproep

STE		PL	PF houder	totaal budget	voorschotten	saldo	agreement	voortgang	start	formeel einde	oplevering	werkelijk einde
05024	DAESO	E. Krahrmer	Eynde, Boves	487.000	121.750	365.250	OK	nov-09	1-10-2006	31-1-2010		
05026	DPC	P. Desmet	Daelemans, Boves	498.000	373.500	124.500	OK	nov-09	1-5-2006	30-9-2009		
05020	LASSY	GJ van Noord	Daelemans, Odijk	496.000	248.000	248.000	OK	nov-09	1-11-2006	1-9-2010		
05030	MIDAS	H. Van hamme	Hessen, Martens	499.000	249.300	249.700	OK	nov-09	1-10-2006	31-12-2010		
05012	NBest	D. van Leeuwen	Smeulders, Bruwaene	470.000	352.500	117.500	OK	mei-09	1-5-2006	1-6-2008	OK	
05035	STEVINcanPRAAT	P.Boersma	Bruwaene, Boves	114.000	28.500	85.500	nvt	jun-08	1-5-2006	31-10-2008	OK	
	TOTAAL			2.564.000	1.373.550	1.190.450						

3de oproep

tender

STE		PL	PF houder	totaal budget	voorschotten	saldo	agreement	voortgang	start	formeel einde	oplevering	werkelijk einde
07014	SONAR (fase 1)	N. Oostdijk	Odijk, Daelemans	100.000	85.456	14.544	OK	okt-08	1-1-2008	1-1-2009	OK	
07014	SONAR (fase 2)	N. Oostdijk	Odijk, Daelemans	836.000	218.704	617.296	OK	nov-09	1-1-2009	1-12-2011		
	TOTAAL			100.000	304.160	631.840						


3de oproep

open oproep

STE		PL	PF houder	totaal budget	voorschotten	saldo	agreement	voortgang	start	formeel einde	oplevering	werkelijk einde
07007	PACO-MT	F. van Eynde	Sas, Odijk	494.575	123.644	370.931	OK	nov-09	1-2-2008	31-1-2011		
07010	DISCO	H. Strik	Martens, Beek	495.419	247.709	247.710	OK	nov-09	1-2-2008	31-1-2011		
07008	AUTONOMATA TOO	H. v/d Heuvel	Hessen, Bruwaene	416.750	104.188	312.563	OK	nov-09	1-2-2008	1-10-2010		
07015	DAISY	S. Moens	Odijk, Smeulders	457.300	114.325	342.975	OK	nov-09	1-5-2008	30-11-2011		
07012	DUOMAN	M. de Rijke	Beek, Boves	440.447	0	440.447	OK	nov-09	1-4-2008	31-3-2011		
	TOTAAL			2.304.491	589.865	1.714.626						

Flankerend beleid

STATUS per 31/12 2009

 = formeel afgerond
 = formele afronding gestart

1e oproep demonstratieprojecten

STE		PL	totaal bud	betaald	saldo	start	einde	demo	mnd
05103	SNRT	Nachtegaal	50.315	50.315	0	15-3-2006	30-9-2006	OK	0
05101	C-Content	Mooren	90.000	90.000	0	1-1-2006	1-4-2007	OK	0
05107	Gemeentecconnect	van Gent	60.000	60.000	0	1-1-2006	1-11-2006	OK	0
	TOTAAL		200.315	200.315	0				

2e oproep demonstratieprojecten

STE		PL	totaal bud	betaald	saldo	start	einde	demo	mnd
06102	Spellingchatbot	Van Pellicom	72.387	65.086	7.301	15-2-2007	14-2-2008	OK	0
06103	PRIMUS	De Mönnink	96.730	72.548	24.183	15-2-2007	14-2-2008		0
06107	Rechtspraakherken	Luimes	97.000	72.750	0	15-2-2007	14-6-2008		0
06112	Klare Taal	Koornwinder	92.200	92.200	0	15-2-2007	14-5-2008	OK	0
06116	Audiokrant	Allemeersch	92.400	92.400	0	15-2-2007	14-5-2008	OK	0
06119	VoiceAssess	Luimes	45.000	45.000	0	15-2-2007	14-7-2007	OK	0
	TOTAAL		495.717	439.984	31.484				

3e oproep demonstratieprojecten

STE		PL	totaal bud	betaald	saldo	start	einde	demo	mnd
07111	NEON	Boedeltje	85.730	64.298	21.433	1-4-2008	1-12-2009		14
07108	Easyinfo	van Gent	19.800	14.850	4.950	15-2-2008	14-8-2008		6
07109	HATCI	Vanpoucke	48.482	35.861	0	1-4-2008	28-2-2009		8
07104	WooDy	Allemeersch	90.000	67.500	22.500	15-2-2008	14-5-2009		15
07105	AAP	Beinema	56.000	42.000	14.000	1-4-2008	31-12-2009		15
	TOTAAL		300.012	224.509	62.883				

oproep 2007 educatieve projecten

STE		PL	totaal bud	betaald	saldo	start	einde	oplevering	mnd
07201	TST op Kennislink.r	Jansen	27.500	0	27.500	1-3-2008	7-1-2009		12
	TOTAAL		27.500	0	27.500				

oproep 2008 educatieve projecten

STE		PL	totaal bud	betaald	saldo	start	einde	oplevering	mnd
08201	Diademo	van Compernelle	32.113	0	32.113	1-10-2008	31-5-2009		12
	TOTAAL		32.113	0	32.113				

oproep 2009 educatieve projecten

STE		PL	totaal bud	betaald	saldo	start	einde	oplevering	mnd
9012	TST op Kennislink2	M. Jansen	25.500	0	25.500	1-10-2008	31-5-2009		12
	TOTAAL		25.500	0	25.500				

oproep 2008 masterclass projecten

STE		PL	totaal bud	betaald	saldo	start	einde	oplevering	mnd
9501	ICT & Dyslexie	de Mönnink	20.000	0	20.000				
	TOTAAL		20.000	0	20.000				

oproep 2009 masterclass projecten

STE		PL	totaal bud	betaald	saldo	start	einde	oplevering	mnd
9014	TST in NL Overheids diensten	Koornwinder	19.000	0	19.000	1-12-2009	1-7-2010		
	TOTAAL		19.000	0	19.000				

STEVIN IPR and Standards policy

IPR policy is an integral part of the STEVIN programme. One of the major aims of this programme is to make the basic digital language infrastructure for Dutch – and above all the results of this programme – available in a non-discriminatory way to all stakeholders. It must be considered a major challenge to formulate and implement an IPR-policy for all new language resources created with public funds acceptable to all parties involved. In the STEVIN programme the situation is more complicated as not only newly created language resources are involved but also those that have been implemented in the past either with or without national or European funding for which IPR has not been satisfactorily settled.

The basic principle of STEVIN IPR policy is that all basic resources – both new and existing ones – should be actively maintained by the *TST-Centre* (Dutch Language and Speech Technology Agency) of the Dutch Language Union. This involves both making available the language resources and protecting their IPR. The *TST-Centre* has started to define the rules and regulations to be followed by the STEVIN programme. These rules and regulations will be based on experiences gained within the context of the Dutch-Flemish Corpus of Spoken Dutch (CGN) that was recently finished and for which close cooperation was established with ELRA and LDC. Both ELRA and LDC have developed IPR-standards which allow the development of resources on the basis of existing resources and are widely accepted by all parties involved, i.e. government, research institutes and industry.

To keep IPR within the STEVIN programme as simple and transparent as possible, before the project actually starts STEVIN project partners must contractually lay down in which way project results will be made available for all stakeholders. These contracts will be based on rules and regulations that are currently being developed and formulated. If a project builds on existing resources for which industrial IPR has been established, it must be contractually stated that the existing resources will be made available against reasonable conditions comparable to the way this has been arranged for pre-existing knowledge in IPR contracts for 6th Framework projects (cf. best practice guide (www.cordis.lu/fp6/find-doc.htm#ipr)).

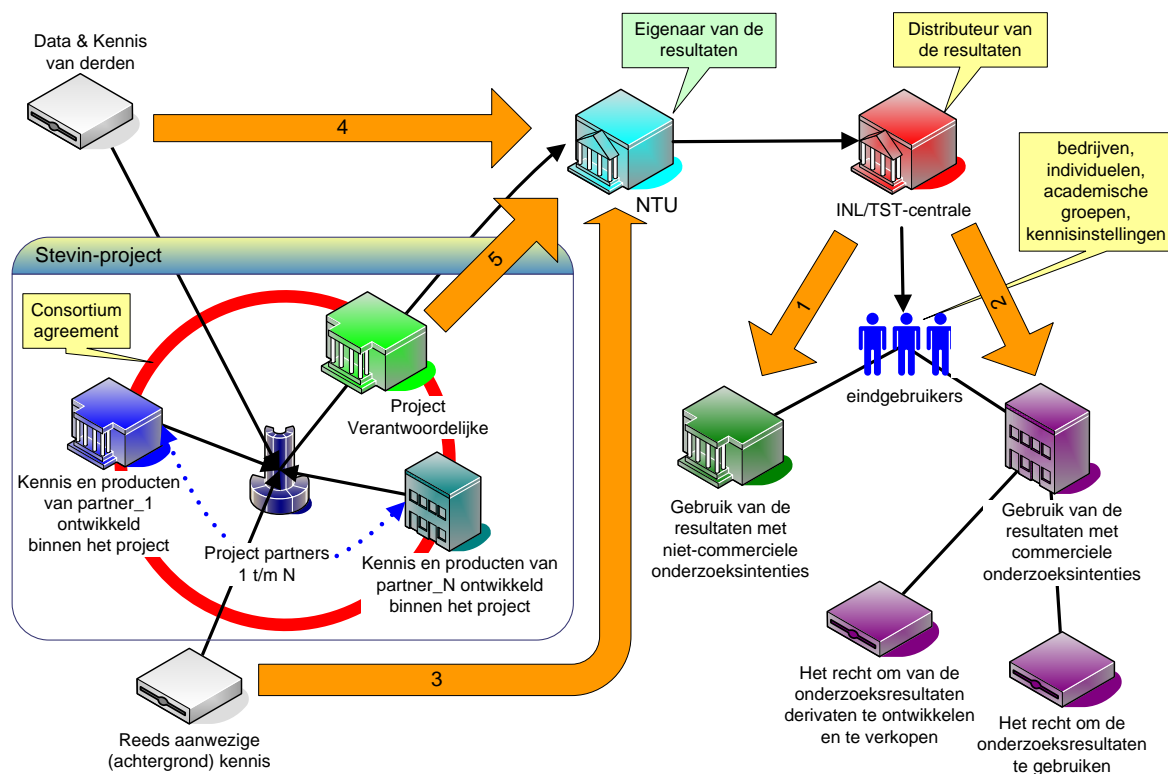
The reusability of some of the language resources developed in the past has been hindered by the use of idiosyncratic formats and data structures. Fortunately, the HLT community has been very active in developing and promoting standards. These were partly developed within European collaborative programmes such as EAGLES and ISLE. Other important institutes concerned with normalisation are ISO/T37, W3C and LISA. For the Dutch HLT-industry, with its relatively small market, it is especially important that international standards are realised and supported by the industry. The Programme Committee demands that projects apply existing standards and cooperate in developing new standards.

IPR and use and re-use of STEVIN results

To enable the use and re-use of STEVIN results, a particular IPR-arrangement has been set up. The materials (software, data etc.) must be handed over to the Dutch Language Union and will be made available to third parties through the Dutch HLT Agency ('TST Centrale' www.tst.inl.nl). The Dutch HLT Agency is responsible for IPR issues, and for the management, maintenance and distribution of HLT-related materials. In addition, the Dutch HLT Agency provides HLT-related information, advice and training to third parties.

Schematic overview HLT actors (in Dutch)

The scheme below had been produced by the STEVIN IPR working group. This committee is led by the Dutch Language Union (NTU) and consists of academic and industrial HLT experts on IPR, legal experts and representatives from the Dutch HLT Agency (TST Centrale). They advise the STEVIN PC and HLT Board in order to co-ordinate and optimize STEVIN IPR practices.



In deze figuur is getracht informatiestromen (zwarte lijnen) en de overeenkomsten die moeten worden afgesloten (oranje pijlen) naast elkaar te leggen. Binnen een Stevin-project (gesymboliseerd door het zuiltje) werken academische (en eventueel niet-academische) partners samen. Er kan door alle partijen achtergrondkennis ingebracht worden die soms ook aan de NTU gelicentieerd moet worden in het geval de achtergrondkennis onderdeel uitmaakt van de projectresultaten. Daarnaast kan een Stevin-project data en kennis van "derden", niet betrokken partijen, gebruiken (bv krantenarchieven of archieven met audio-visueel materiaal), die eveneens aan de NTU gelicentieerd moet worden in het geval deze data en kennis onderdeel uitmaken van de projectresultaten. De NTU sluit daartoe licentieovereenkomsten af met de projectpartners (pijl 3) en met de derden (pijl 4). De rechten op de binnen het Stevin-project verworven kennis en data moeten aan de NTU overgedragen worden: dat gebeurt in pijl 5.

De TST-centrale geeft namens de NTU de resultaten van de Stevin-projecten in licentie aan eindgebruikers. De eindgebruikers kunnen de resultaten (kennis, data of derivaten) voor niet-commercieel onderzoek gebruiken (pijl 1) of er commerciële bedoelingen mee hebben (pijl 2). In dat laatste geval kunnen ze de resultaten van de projecten gebruiken voor eigen onderzoek en/of om er zelf derivaten mee te maken en/of om de al gemaakte derivaten van de projecten commercieel uit te nutten.

nú | STEVIN

IPR flyer prepared by IPR working group to help convince data provider to make their data available for HLT R&D (in Dutch)

Nederlandse Taalunie Verzoek om data

U wilt efficiënt in uw archieven kunnen zoeken...
Naar alle waarschijnlijkheid beschikt u, als content provider, over grote archieven met digitale tekst, beeld of geluid. U hebt zo uw eigen methoden om efficiënt te zoeken in die data. Grote kans dat u daarbij gebruik maakt van TST: Taal- en Spraaktechnologie. Zoekt u met behulp van trefwoorden? Gebruikt u tools om lange documenten samen te vatten? Of gebruikt u geavanceerde zoektechnologie om (gesproken) documenten terug te vinden? In al deze gevallen komt de onderliggende techniek voort uit de TST.

... en TST helpt u daarbij.
Zoals dat bij iedere technologische ontwikkeling gaat, werkt ook in TST het mechanisme van voortschrijdend inzicht. Er wordt doorlopend onderzoek gedaan naar nieuwe en nóg betere toepassingen. Ideën zijn er genoeg en aan kennis ontbreekt het niet. Nederland en Vlaanderen lopen wat TST betreft voorop in Europa. Maar van één ding is er nooit genoeg: grote hoeveelheden spraak en tekst.

TST heeft veel data nodig...
Voor het ontwikkelen van betere classificatietechnieken voor gesproken en geschreven teksten, betere spraakherkenners, betere automatische samenwatters en betere zoekstrategieën is heel veel taalmateriaal nodig. De nieuwste generatie taal- en spraaksoftware leert van realistische voorbeelden: hoe schrijft een journalist voor de krant? Hoe spreekt een sportverslaggever bij Studio Sport of bij Sporza? Hoe meer voorbeeldmateriaal, hoe beter de resultaten. TST-resultaten drijven op statistische analyse en de grondstof daarvoor is veel, meer, en nog meer taal.
Teksten en spraakopnamen zijn doorgaans goed beschermd door auteursrechten. TST-onderzoekers maken alleen gebruik van rechienvrij taalmateriaal, en soms van collecties die de rechthebbenden voor onderzoeksdoelinden beschikbaar hebben gesteld.

... helpt u TST daarbij?
U kunt TST een impuls geven door taallocollecties te doneren. De Nederlandse Taalunie heeft een bronnen centrum, de TST-Centrale, ingesteld, die deze collecties zal beheren. U krijgt de garantie dat het materiaal alléén gebruikt zal worden voor het (verder) ontwikkelen en testen van TST. Teksten en geluidsopnamen mogen na gebruik voor statistische analyses niet verder verspreid worden door TST-onderzoekers. En ... doordat u de mogelijkheid krijgt de ontwikkelingen vanaf het eerste moment te volgen kunt u straks nóg sneller de weg vinden in uw digitale archieven.

De Nederlandse Taalunie is een internationale verdragsorganisatie waarin Nederland, België en Suriname samenwerken op het gebied van de Nederlandse taal en letteren en het onderwijs in en van het Nederlands.

STEVIN is een meerjarig onderzoeks- en stimuleringsprogramma met als doel het stimuleren van de taal- en spraaktechnologische sector in Vlaanderen en Nederland zodat de positie van het Nederlands in de informatie- en communicatiewereld wordt versterkt. STEVIN wordt gefinancierd en begeleid door de Nederlandse en Vlaamse overheden en uitgevoerd onder auspiciën van de Nederlandse Taalunie. Nederlandse partners zijn de Ministeries van Economische Zaken en Onderwijs, Cultuur en Wetenschappen, de Nederlandse Organisatie voor Wetenschappelijk Onderzoek en SenterNovem; Vlaamse partners zijn het Departement Economie, Wetenschap en Innovatie, het Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen en het Fonds Wetenschappelijk Onderzoek Vlaanderen.

De **TST-centrale** (Centrale voor Taal en Spraaktechnologie) is verantwoordelijk voor het verwerven, beheren, onderhouden en verspreiden van diverse spraak- en taalmaterialen die voor het Nederlandstalig taalgebied met overheidsmiddelen zijn ontwikkeld.

Voor meer informatie over STEVIN:
I <http://www.taalunieversum.org/stevin>
E stevin@taalunie.org

Voor meer informatie over de TST-centrale:
I <http://www.tst.nl>
E st@tst.nl

nú | STEVIN

Scientific outputs of STEVIN programme in international literature

Journals

1. [DISCO] Cucchiarini, C., A. Neri & H. Strik (2009), Oral Proficiency training in Dutch L2: the Contribution of ASR-based corrective feedback, *Speech Communication* 51 (10), October 2009, pp.853-863.
2. [DPC] Paulussen, Hans (2007). "Acta academica: DPC, een nieuw vertaalcorpus". *Romaneske* 2007: 1, 19-22
3. [DUOMAN] He J., Weerkamp W.W., Larson M., de Rijke M., An Effective Coherence Measure to Determine Topical Consistency in User Generated Content, *International Journal on Document Analysis and Recognition*, 2010
4. [DUOMAN] Hofmann K., Balog K., Bogers T., de Rijke M., Contextual Factors for Finding Similar Experts, *Journal of the American Society for Information Science and Technology*, 2010
5. [DUOMAN] Tsagkias E., Larson M., de Rijke M., Predicting Podcast Preference: An Analysis Framework and its Application, *Journal of the American Society for Information Science and Technology*, 2010
6. [IRME] Grégoire, N. (accepted), 'DuELME: A Dutch Electronic Lexicon of Multiword Expressions', *Journal of Language Resources and Evaluation*, special issue on Multiword Expressions.
7. [MIDAS] Gemmeke, J., H. Van hamme, B. Cranen, L. Boves (submitted), Compressive Sensing for Missing Data Imputation in Noise Robust Speech Recognition. Submitted to *IEEE-Journal of selected topics in Signal Processing*.
8. [PACOMT] Van den Bogaert, J. (2009). The emergence of hybrid machine translation systems and their integration into business processes. *Berkeley Globalization Conference. Journal of Internationalisation and Localisation*.
9. [PACOMT] Vandeghinste, V. (2009). Scaling up a Hybrid MT System: From Low to Full Resources. In *Linguistica Antverpiensia* 8/2009.

Conference proceedings

1. [AUTONOMATA TOO] Heuvel, H. van den, Reveil, B., Martens, J-P., D'hoore, B. (2009): "Pronunciation-based ASR for names", in *Proceedings Interspeech2009*, Brighton, UK
2. [AUTONOMATA TOO] Reveil, B., Martens, J-P., D'hoore, B. (2009): "How speaker tongue and name source language affect the automatic recognition of spoken names", in *Proceedings Interspeech2009*, Brighton, UK
3. [AUTONOMATA] Van den Heuvel, H., J.P. Martens, and N. Konings (2008). 'Fast and easy development of pronunciation lexicons for names', *Proceedings LangTech(Rome)*, 117-120.
4. [AUTONOMATA] Van den Heuvel, H., J.P. Martens, B. D'hoore, K. D'hanens, and N. Konings (2008). 'The Autonomata Spoken Name Corpus. Design, recording, transcription and distribution of the corpus', *Proceedings LREC (Marrakech)*.
5. [AUTONOMATA] Van den Heuvel, H., J.P. Martens, N. Konings (2007). 'G2P conversion of names. What can we do (better)?', *Proceedings Interspeech (Antwerp)*, 1773-1776
6. [AUTONOMATA] Yang, Q., J.P. Martens, N. Konings, H. van den Heuvel (2006), 'Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme-to-phoneme (g2p) conversion of names', *Proceedings LREC (Genoa)*, 287-292.
7. [COREA] Hendrickx I., Hoste V. and Daelemans W., (2007), Evaluating hybrid versus data-driven coreference resolution. In *Anaphora: Analysis, Algorithms and Application. Lecture Notes in Artificial Intelligence* 4410, pp. 137-150, Springer Verlag
8. [COREA] Hendrickx, I., G. Bouma, F. Coppens, W. Daelemans, V. Hoste, G. Kloosterman, A. Mineur, J. Van Der Vloet, J. Verschelde (to appear) 'Coreference Resolution for Extracting Answers for Dutch'. *Proceedings of LREC (Marrakech, 2008)*.

9. [COREA] Hendrickx, I., V. Hoste and W. Daelemans (2008). 'Semantic and Syntactic features for Anaphora Resolution for Dutch'. In: Springer Lecture Notes in Computer Science. Proceedings of the CILing-2008 conference, Volume 4919, pp.351-361, Haifa, Israel, 2008.
10. [COREA] Hoste, V., I. Hendrickx, L. Macken (2007). 'The Referential versus Non-referential Use of the Neuter Pronoun in Dutch and English', In: Proceedings of Corpus Linguistics 2007, Birmingham, England, 2007
11. [COREA] Hoste, V., I. Hendrickx, W. Daelemans (2007). 'Disambiguation of the neuter pronoun and its effect on pronominal coreference resolution', Proceedings TSD (Plzen), 48-55.
12. [CORNETTO] Horák, A., I. Maks, A. Rambousek, R. Segers, H. van der Vliet, P. Vossen (2008), Cornetto Tools and Methodology for Interlinking Lexical Units, Synsets and Ontology, in: Proceedings of the 18th International Congress of Linguists (CIL18), Seoul, Republic of Korea, July 21-26, 2008.
13. [CORNETTO] Horák, A., P. Vossen, A. Rambousek "A Distributed Database System for Developing Ontological and Lexical Resources in Harmony", in: Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics: CILing 2008, February 17-23, 2008, Haifa, Israel. Also to be published in the Lecture Notes on Computational Linguistics and Intelligent Text Processing in Lectures Notes in Computer Science, Volume 4919/2008, ISBN 978-3-540-78134-9, 1-15, Springer-Verlag, Berlin, 2008.
14. [CORNETTO] Horák, A., Vossen P., Rambousek A. (2008) The Development of a Complex-Structured Lexicon based on WordNet, in: Proceedings of the Fourth International GlobalWordNet Conference - GWC 2008, Szeged, Hungary, January 22-25, 2008
15. [CORNETTO] Jijkoun, V. and K. Hofmann "Generating a Non-English Subjectivity Lexicon: Relations That Matter". Submitted to EACL 2009
16. [CORNETTO] Maks I., P. Vossen, Segers R., VanderVliet H., van Zutphen H. (2008) "Encoding adjectives in the Dutch semantic lexical database Cornetto", in: Proceedings of LREC 2008, Marrakech, Morocco, May 28-30 May 2008.
17. [CORNETTO] Tjong Kim Sang, E. and K. Hofmann (2007), Automatic Extraction of Dutch, Hypernym-Hyponym Pairs. In Proceedings of CLIN-2006, Leuven, Belgium, 2007.
18. [CORNETTO] Tjong Kim Sang, E. and K. Hofmann: "Lexical Patterns or Dependency Patterns: Which Is Better for Hypernym Extraction?". Submitted to EACL 2009
19. [CORNETTO] Vossen P., Maks I., Segers R., VanderVliet H. (2008) "Integrating lexical units, synsets and ontology in the Cornetto Database", in: Proceedings of LREC 2008, Marrakech, Morocco, May 28-30 May 2008.
20. [CORNETTO] Vossen P., Maks I., Segers R., VanderVliet H., van Zutphen H. (2008) "The Cornetto Database: the architecture and alignment issues", in: Proceedings of the Fourth International GlobalWordNet Conference - GWC 2008, Szeged, Hungary, January 22-25, 2008
21. [CORNETTO] Vossen, P., Hofmann, K. de Rijke, M. Tjong Kim Sang, E. and Deschacht, K. (2007), The Cornetto Database: Architecture and User-scenarios. In Proceedings of DIR 2007, pp. 89-96.
22. [DAESO] Hendrickx, I. & W. Bosma (2008), 'Using coreference links and sentence compression in graph-based summarization'. In: Proceedings of the Text Analysis Conference 2008, Gaithersburg, USA.
23. [DAESO] Hendrickx, I., W. Daelemans, K. Luyckx, R. Morante and V. Van Asch (2008), 'CNTS: Memory-Based Learning of Generating Repeated References. In Proceedings of the 5th International Natural Language Generation Conference (INLG 2008), Salt Fork, Ohio, USA, June 12-14, 2008, s.l., Association for Computational Linguistics, 2008, p. 194-195
24. [DAESO] Kraemer, E., E. Marsi and P. van Pelt (2008), 'Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion'. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Columbus, Ohio, USA, June 15-20, 2008, pp. 193-196.
25. [DAESO] Kraemer, E., M. Theune, J. Viethen and I. Hendrickx, 'The Costs of Redundancy in Referring Expressions (GRAPH)'. Accepted for the Referring Expression Generation Challenge 2008, held in conjunction with the 5th International Natural Language Generation Conference (INLG 2008), Salt Fork, Ohio, USA, June 12-14, 2008.

26. [DAESO] Marsi, E. and E. Krahmer (to appear), 'Detecting semantic overlap: A parallel monolingual treebank for Dutch'. In: Proceedings of Computational Linguistics in the Netherlands (CLIN) 2007
27. [DAESO] Theune, M., J. Viethen, I. Hendrickx and E. Krahmer, 'GRAPH: Realizing the Costs'. Accepted for the Referring Expression Generation Challenge 2008, held in conjunction with the 5th International Natural Language Generation Conference (INLG 2008), Salt Fork, Ohio, USA, June 12-14, 2008.
28. [DCOI] Oostdijk, N., L. Boves (2006). 'User requirement analysis for the design of a reference corpus of written Dutch', Proceedings LREC (Genoa), 1206-1211.
29. [DCOI] Reynaert, M. (2006). 'Corpus-induced corpus clean-up', Proceedings LREC (Genoa), 87-92.
30. [DCOI] Schuurman, I. and P. Monachesi (2006). The contours of a semantic annotation scheme for Dutch, In Proceedings of the 16th Meeting of Computational Linguistics in the Netherlands 2005.
31. [DCOI] Van den Bosch, A., I. Schuurman, V. Vandeghinste (2006). 'Transferring POS tagging and lemmatization tools from spoken to written Dutch corpus development', Proceedings LREC (Genoa), 1807-1810.
32. [DCOI] Van Noord, G., I. Schuurman, V. Vandeghinste (2006). 'Syntactic annotation of large corpora in STEVIN', Proceedings LREC (Genoa), 1811-1814.
33. [DISCO] Cucchiari, C., J. van Doremalen, & H. Strik (2008) DISCO: Development and Integration of Speech technology into Courseware for language learning. Proceedings of Interspeech-2008, Brisbane, Australia, Sept. 26-29, 2008, pp. 2791-2794
34. [DISCO] Strik, h., A. Neri and C. Cucchiari (2008), Speech technology for language tutoring. Proceedings of LangTech-2008, Rome, February 28-29, 2008
35. [DISCO] Strik, H., J. van Doremalen & C. Cucchiari (2008) A CALL system for practicing speaking proficiency: pronunciation, morphology and syntax, CALL 2008, Proceedings of the XIIIth International CALL Conference, Antwerp.
36. [DPC] Macken, L., J. Trushkina, & L. Rura (2007). 'Dutch Parallel Corpus: MT Corpus and translator's aid'. In: Proceedings of Machine Translation Summit XI, 10-14 september 2007, Copenhagen, Denmark, 313-320.
37. [DPC] Macken, L., J. Trushkina, H. Paulussen, L. Rura, P. Desmet, & W. Vandeweghe (2007). 'Dutch Parallel Corpus. A multilingual annotated corpus'. In: On-line Proceedings of Corpus Linguistics 2007, 27-30 juli 2007, Birmingham, United Kingdom.
38. [DPC] Paulussen, H., L. Macken, J. Trushkina, P. Desmet & W. Vandeweghe (2006). 'Dutch Parallel Corpus: a multifunctional and multilingual corpus'. Cahiers de l'Institut de Linguistique de Louvain, CILL, Louvain-La-Neuve, 32.1-4 (2006), 269-285
39. [DPC] Rura, L., W. Vandeweghe & Maribel Montero Perez (2008). 'Designing a parallel corpus as a multifunctional translator's aid'. In: Proceedings of XVIII FIT World Congress, 4-7 August 2008, Shanghai, China
40. [DPC] Trushkina, J., L. Macken & H. Paulussen (2008). 'Sentence Alignment in DPC: Maximizing Precision, Minimizing Human Effort'. In: Proceedings of LREC: 6th Language Resources and Evaluation Conference, 28-30 May 2008, Marrakesh, Morocco.
41. [DUOMAN] Balog K., de Rijke M., Franz R., Peetz H., Brinkman B., Johgi I., Hirschel M., SaHaRa: Discovering Entity-Topic Associations in Online News, 8th International Semantic Web Conference (ISWC 2009): Springer, October, 2009
42. [DUOMAN] Hofmann K., Tsagkias E., Meij E J., de Rijke M., The Impact of Document Structure on Keyphrase Extraction, ACM 18th Conference on Information and Knowledge Management (CIKM 2009), Hong Kong, ACM, November, 2009
43. [DUOMAN] Jijkoun V., Hofmann K. Generating a Non-English Subjectivity Lexicon: Relations That Matter. In Proceedings of 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009), 2009
44. [DUOMAN] Khalid M. A., Jijkoun V., de Rijke M. The Impact of Named Entity Normalization on Information Retrieval for Question Answering. Proceedings of the 30th European Conference on Information Retrieval (ECIR 2008): Springer, pp. 705-710, April, 2008
45. [DUOMAN] Tsagkias E., de Rijke M., Weerkamp W.W., Predicting the Volume of Comments on Online News Stories, ACM 18th Conference on Information and Knowledge Management (CIKM 2009), Hong Kong, ACM, November, 2009.

46. [DUOMAN] Tsagkias E., Larson M., de Rijke M. Exploiting Surface Features for the Prediction of Podcast Preference. 31st European Conference on Information Retrieval Conference (ECIR 2009), April, 2009
47. [IRME] Grégoire N., (2006), Elaborating the parameterized Equivalence Class Method for Dutch. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 1894-1899
48. [IRME] Grégoire N., (2007), Design and Implementation of a Lexicon of Dutch Multiword Expressions. In Proceedings of the ACL07 Workshop on A Broader Perspective on Multiword Expressions, pp. 17-24
49. [IRME] Van de Cruys, T. and B. Villada Moirón (2007), 'Lexico-Semantic Multiword Expression Extraction'. In P. Dirix et al. (eds.), Computational Linguistics in the Netherlands 2006, pp. 175-190
50. [JASMIN] Cucchiarini, C., J. Driesen, H. Van hamme, and E. Sanders (2008) Recording Speech of Children, Non-Natives and Elderly People for HLT Applications: the JASMIN-CGN Corpus, Proceedings LREC2008, Marrakesh, Morocco.
51. [JASMIN] Cucchiarini, C., H. Van hamme, O. van Herwijnen, F. Smits (2006). 'JASMIN-CGN: Extension of the Spoken Dutch Corpus with Speech of Elderly People, Children and Non-natives in the Human-Machine Interaction Modality', Proceedings LREC (Genoa), 135-138.
52. [LASSY] Bouma G. and G. Kloosterman. Mining Syntactically Annotated Corpora with XQuery. In: LAW 2007, Prague
53. [LASSY] Van Noord, G. I. Schuurman, V. Vandeghinste. Syntactic Annotation of Large Corpora in STEVIN. In: LREC 2006
54. [LASSY] Van Noord, G. Learning Efficient Parsing. In: EACL 2009. The 12th Conference of the European Chapter of the Association for Computational Linguistics. 30 March - 3 April 2009, Athens, Greece. pp 817-825.
55. [LASSY] Van Noord, G. Using Self-Trained Bilexical Preferences to Improve Disambiguation Accuracy. In: IWPT2007, Prague.
56. [MIDAS] Gemmeke J. and B. Cranen, (2008), Noise robust digit recognition using sparse representations, In Proceedings of the International Speech Communication Association (ISCA 2008) ISCA Tutorial and Research Workshop (ITRW) "Speech Analysis and Processing for knowledge discovery",
57. [MIDAS] Gemmeke J. and Cranen B., (2008), Noise reduction through Compressed Sensing, In Proceedings of InterSpeech 08, pp. 1785-1788
58. [MIDAS] Gemmeke J. and Cranen B., (2009), Missing Data Imputation using Compressive Sensing Techniques for Connected Digit Recognition, In Proceedings of the International Conference on Digital Signal Processing (DSP 2009)
59. [MIDAS] Gemmeke J. and Cranen B., (2009), Sparse imputation for noise robust speech recognition using soft masks, in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009), pp. 4645-4648
60. [MIDAS] Gemmeke J., Cranen B. and ten Bosch L., (2008), On the relation between statistical properties of spectrographic masks and recognition accuracy, In IASTED Signal Processing, Pattern Recognition and Applications (SPPRA 2008), pp. 200-207
61. [MIDAS] Gemmeke, J. (2008), Classification on incomplete data using sparse representations: imputation is optional. In Wehenkel, L., Geurts, P. and Marée, R. (eds.), Proceedings of the 17th annual Belgian-Dutch Conference on Machine Learning (BeNeLearn 2008), pp. 71-72
62. [MIDAS] Gemmeke, J. and B. Cranen (EUSIPCO 2008), Using sparse representations for missing data imputation in noise robust speech recognition
63. [MIDAS] Gemmeke, J., L. ten Bosch, L. Boves, and B. Cranen (submitted to EUSIPCO 2009), Using sparse representations for exemplar based continuous digit recognition
64. [MIDAS] Gemmeke, J., Y. Wang, M. Van Segbroeck, B. Cranen, H. Van hamme (submitted to Interspeech 2010), Application of noise robust MDT speech recognition on the SPEECON and SpeechDat-Car databases
65. [MIDAS] Wang Y., and H. Van hamme (NAG/DAGA 2009), Speed improvements in a Missing Data-based speech recogniser by Gaussian selection. Paper No. 356

66. [MIDAS] Wang, Y., R. Vuerinckx, J. Gemmeke, B. Cranen, H. Van hamme (NAG/DAGA 2009), Evaluation of missing data techniques for in-car automatic speech recognition. Paper No. 373
67. [N-BEST] Despres, J., P. Fousek, J.-L. Gauvain, S. Gay, Y. Josse, L. Lamel, A. Messaoudi, "Modeling Northern and Southern Varieties of Dutch for STT", Proceedings ISCA Interspeech, Brighton, September 2009, pp 96-99.
68. [N-BEST] Huijbregts, M., R. Ordelman, L. van der Werff and F. de Jong, "SHoUT, the University of Twente N-Best Submission", Proceedings ISCA Interspeech, Brighton, September 2009, pp 2575--2578
69. [N-BEST] Kessens, J., D. van Leeuwen (2007), 'N-Best: The Northern and Southern Dutch Evaluation of Speech Recognition Technology', Procs Interspeech, 1354-1357.
70. [N-BEST] Van Leeuwen, D.A., J. Kessens, E. Sanders and H. van den Heuvel, "Results of the N-Best 2008 Dutch Speech Recognition Evaluation", Proceedings ISCA Interspeech, Brighton, September 2009, pp 2571-2574
71. [PACOMT] Tiedemann, J., & Kotz , G. (2009). A Discriminative Approach to Tree Alignment. Proceedings of RANLP
72. [PACOMT] Vandeghinste, V. (2007). Removing the distinction between a translation memory, a bilingual dictionary and a parallel corpus. In Proceedings of Translation and the Computer, 29. London
73. [PACOMT] Vandeghinste, V., (2009), Tree-based Target Language Modeling. In M rquez L. and Somers H. (eds.), Proceedings of the 13th Annual conference of the European Association for Machine Translation (EAMT 2009). European Association for Machine Translation, pp.152-159.
74. [SONAR/LASSY/D-COI] Oostdijk, N., M. Reynaert, P. Monachesi., G. van Noord, R. Ordelman, I. Schuurman, V. Vandeghinste (2008). From D-Coi to SoNaR: A reference corpus of Dutch. In *Proceedings LREC 2008*.
75. [SPRAAK/N-BEST] Demuyne, K., A. Puurula, D. Van Compernelle, P. Wambacq: The ESAT 2008 System for N-Best Dutch Speech Recognition Benchmark, in Proceedings IEEE ASRU 2009, Merano, Italy, 13-17 December 2009.
76. [SPRAAK] Demuyne, K., J. Roelens, D. Van Compernelle, P. Wambacq (2008), SPRAAK: an open source "Speech Recognition and Automatic Annotation Kit", In Proc. Interspeech 2008, page 495, Brisbane, Australia, September 2008
77. [STEVIN] D'Halleweyn, E., J. Odijk, L. Teunissen and C. Cucchiarini (2006), The Dutch-Flemish HLT Programme STEVIN: Essential Speech and Language Technology Resources. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 761-766.
78. [STEVIN] Spyns, P., Cucchiarini, C. and D'Halleweyn, E. (2008), The Dutch-Flemish comprehensive approach to HLT stimulation and innovation: STEVIN, HLT Agency and beyond. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Workshop proceedings

1. [?] Plank, B. and G. van Noord (2008), Exploring an Auxiliary Distribution based approach to Domain Apaptation of a Syntactic Disambiguation Model, In Proceedings of the Coling Workshop on Cross-Framework and Cross-Domain Parser Evaluation (PE), pp. 9-16.
2. [COREA] Bouma, G. & G. Kloosterman (2007). 'Mining Syntactically Annotated Corpora using Xquery', Proceedings of Linguistic Annotation Workshop, ACL 2007 (Prague).
3. [COREA] Hendrickx I. and Daelemans W., (2007), Adding Semantic Information: unsupervised Clusters for Co-reference Resolution. In Workshop on Machine Learning for Natural Language Processing.
4. [COREA] Hoste, V. & A. van den Bosch (2007). 'A Modular Approach to Learning Dutch Co-reference Resolution',_Proceedings of first WAR Colloquium. Cambridge Scholars Press (to appear)
5. [COREA] Hoste, V. & W. Daelemans (2005). 'Comparing Learning Approaches to Coreference Resolution. There is More to it Than Bias', Proceedings of Workshop on Meta-Learning (Bonn), 20-27
6. [CORNETTO] Boiy, E., K. Deschacht & M.-F. Moens (2008). Learning Visual Entities and their Visual Attributes from Text Corpora In Proceedings of the 5th International Workshop on Text-based Information Retrieval. IEEE Press.

7. [CORNETTO] Fellbaum, C. & P. Vossen (2007). 'Connecting the Universal to the Specific: Towards the Global Grid', Proceedings of First Int. workshop on Intercultural Collaboration (Kyoto) (published on the web)
8. [CORNETTO] Vossen, P., I. Maks, R. Segers & H. van der Vliet (2008). 'Cornetto: lexical units, synsets and ontological types combined', Workshop on Linguistic Studies of Ontology: From Lexical Semantics to Formal Ontologies and Back, (Seoul) (to appear)
9. [DAESO] Hendrickx, I., W. Daelemans, E. Marsi and E. Krahmer (to appear) 'Reducing Redundancy in Multi-document Summarization Using Lexical Semantic Similarity'. Proceedings of the 2009 Workshop on Language Generation and Summarisation (ULG+Sum 2009), Association for Computational Linguistics, Singapore, pp. 63-66.
10. [DAESO] Marsi E. and E. Krahmer (2007), 'Annotating a parallel monolingual treebank with semantic similarity relations'. In: The Sixth International Workshop on Treebanks and Linguistic Theories (TLT'07), Bergen, Norway, December 7-8, 2007.
11. [DAESO] Marsi, E., E. Krahmer and W. Bosma (2007), 'Dependency-based paraphrasing for recognizing textual entailment'. In: *Proceedings of ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, June 2007.
12. [DAESO] Marsi, E., E. Krahmer, I. Hendrickx, and W. Daelemans (to appear), 'Is sentence compression an NLG task?'. In: Proceedings of 12th European Workshop on Natural Language Generation (ENLG 2009), Athens, Greece, pp. 25-32
13. [DAESO] Wubben, S., A. van den Bosch, E. Krahmer, and E. Marsi (to appear), 'Clustering and Matching Headlines for Automatic Paraphrase Acquisition'. In: Proceedings of ENLG 2009, Athens, Greece, pp. 122-125.
14. [DCOI] Monachesi, P., & J. Trapman (2006). 'Merging FrameNet and PropBank in a corpus of written Dutch'. Proceedings of workshop on Merging and layering linguistic information. (Genoa), 32-39.
15. [DUOMAN] Balog K., He J., Hofmann K., Jijkoun V B., Monz C., Tsagkias E., Weerkamp W.W., de Rijke M. The University of Amsterdam at WePS2. In: Second Web People Search Evaluation Workshop (WEPS 2009), April, 2009
16. [DUOMAN] Hofmann K., de Rijke M., Huurnink B., Meij E J. A Semantic Perspective on Query Log Analysis. Working Notes for the CLEF 2009 Workshop, September, 2009
17. [DUOMAN] Jijkoun V., Khalid M. A., Marx M., de Rijke M. Named Entity Normalization in User Generated Content. Proceedings of the SIGIR 2008 Workshop on Analytics for Noisy Unstructured Text Data (AND 2008), Singapore, July, 2008
18. [DUOMAN] Jijkoun V., de Rijke M. Overview of WebCLEF 2008 (draft). Working Notes for the CLEF 2008 Workshop, Aarhus, September, 2008
19. [DUOMAN] Jijkoun V., de Rijke M. Overview of WebCLEF 2008. In Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, to appear
20. [IRME] Grégoire, N. (2007). 'Design and Implementation of a Lexicon of Dutch Multiword Expressions', Proceedings of Workshop on A Broader Perspective on Multiword Expressions (Prague), 17-24
21. [IRME] Van de Cruys, T. & B. Villada Moirón (2007). 'Semantics-based Multiword Expression Extraction', Proceedings of Workshop on A Broader Perspective on Multiword Expressions (Prague), 25-32.
22. [IRME] Villada Moirón, B. & J. Tiedemann (2006). 'Identifying idiomatic expressions using automatic word-alignment'. Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context, p.33-40. Trento, Italy.
23. [IRME] Villada Moirón, B. (2005), 'Linguistically enriched corpora for establishing variation in support verb constructions'. In Proceedings of the 6th International Workshop on Linguistically Interpreted Corpora (Linc'05) held at The 2nd International Joint Conference on Natural Language Processing (IJCNLP-05). R. of Korea
24. [LASSY] Bouma, G. and J. Spenader. The Distribution of Weak and Strong Object Reflexives in Dutch. In: Frank van Eynde, Anette Frank, Koenraad de Smedt, Gertjan van Noord (editors), Proceedings of

- the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7). January 23-24, 2009, Groningen, The Netherlands. LOT Occasional Series
25. [LASSY] Schuurman, I., V. Hoste and P. Monachesi. Cultivating Trees: Adding Several Semantic Layers to the Lassy Treebank in SoNaR. In: Frank van Eynde, Anette Frank, Koenraad de Smedt, Gertjan van Noord (editors), *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*. January 23-24, 2009, Groningen, The Netherlands. LOT Occasional Series.
 26. [LASSY] Tjong Kim Sang, E.F. To Use a Treebank or Not - Which Is Better for Hypernym Extraction? In: Frank van Eynde, Anette Frank, Koenraad de Smedt, Gertjan van Noord (editors), *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*. January 23-24, 2009, Groningen, The Netherlands. LOT Occasional Series.
 27. [LASSY] Van Noord, G. and G. Bouma. Parsed Corpora for Linguistics. In: *Proceedings of EACL Workshop The Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* Athens, 2009. pp 33-39.
 28. [LASSY] Van Noord, G. Huge Parsed Corpora in LASSY. In: Frank van Eynde, Anette Frank, Koenraad de Smedt, Gertjan van Noord (editors), *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*. January 23-24, 2009, Groningen, The Netherlands. LOT Occasional Series
 29. [LASSY] Van Noord, G. Self-trained Bilexical Preferences to Improve Disambiguation Accuracy. To appear in a book on parsing technology, based on selected papers from the IWPT 2007, CONNL 2007, and IWPT 2005 workshops, edited by Harry Bunt, Paola Merlo and Jakim Nivre, published by Springer
 30. [SONAR] Schuurman, I., V. Hoste and P. Monachesi (2009). Cultivating Trees: Adding Several Semantic Layers to the Lassy Treebank in SoNaR. In *Proceedings of the International Workshop on Treebanks and Linguistic Theories (TLT 7)*.

Book editing

1. [IRME] Grégoire N., Evert S. and KIM S.N. (eds.), (2007), *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*.
2. [IRME] Grégoire, N., S. Evert & B. Krenn (eds.) (2008), 'Proceedings of the Workshop Towards a Shared Task for Multiword Expressions', LREC 2008, Marrakech, Morocco. June 1, 2008.
3. [IRME] Villada Moirón, B., A. Villavicencio, D. McCarthy, S. Evert, & S. Stevenson (eds.) (2006). 'Proceedings of COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties' (Sydney).
4. [LASSY] Van Eynde, F., A. Frank, K. de Smedt, G. van Noord (eds), *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*. January 23-24, 2009, Groningen, The Netherlands. LOT Occasional Series

Book contributions

1. [CORNETTO] Horak A., P. Vossen, A. Rambousek (2008), "A Distributed Database System for Developing Ontological and Lexical Resources in Harmony", in *the Lecture Notes on Computational Linguistics and Intelligent Text Processing in Lectures Notes in Computer Science*, Volume 4919/2008, ISBN 978-3-540-78134-9, 1-15, Springer-Verlag, Berlin, 2008.
2. [CORNETTO] Vossen P. (fc) "WordNet: principles, developments and applications", in: *Dictionaries. An International Encyclopedia of Lexicography. Volume: Recent developments with special focus on computational lexicography*, Walter/Mouton de Gruyter, Handbooks of Linguistics and Communication Science (HSK), Berlin, 2008
3. [CORNETTO] Vossen P., Fellbaum C. (2009) "Universals and Idiosyncracies in Multilingual WordNets", in: *Handbook Multilingual Lexicography*, Oxford University Press, 2009
4. [DUOMAN] Balog K., Azzopardi L A., de Rijke M. Resolving Person Names in Web People Search. *Weaving Services, Locations, and People on the WWW*: Springer, July, 2009

5. [DUOMAN] Fissaha Adafre S., de Rijke M., Tjong Kim Sang E F. Completing Lists of Entities. In: *Recent Advances in Natural Language Processing V*: John Benjamins Publishing Company, 2009
6. [DUOMAN] Hendrickx I., Hoste V. Coreference Resolution on Blogs and Commented News. In: S. Lalitha Devi, A. Branco, and R. Mitkov (Eds.): *DAARC 2009, Lecture Notes in Artificial Intelligence* 5847, pp. 43–53, Springer-Verlag Berlin Heidelberg.
7. [PACOMT] Tiedemann, J. (2008). Prospects and Trends in Data-Driven Machine Translation. In Nivre, Joakim; Dahllöf, Mats ; Megyesi, Beáta (eds). *Resourceful Language Technology: Festschrift in Honor of Anna Sägvall Hein, 2008-06-10*, Uppsala Sweden
8. [PACOMT] Tiedemann, J. (to appear) "News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds) *Recent Advances in Natural Language Processing, Volume V*, John Benjamins, Amsterdam/Philadelphia
9. [PACOMT] Vandeghinste, V. (2008). *A Hybrid Modular Machine Translation System*. PhD. Leuven

Books

1. [IRME] Nicole Grégoire (2009 to appear), *Untangling Multiword Expressions*, PhD Thesis, Utrecht 10 november 2009
2. [PACOMT] Vandeghinste, V. (2008). *A Hybrid Modular Machine Translation System*. PhD thesis, Leuven

List of HLT activities organised or financially supported by the STEVIN programme

<i>Date</i>	<i>Location</i>	<i>HLT activity (* = organised by STEVIN)</i>
2004		
September 15	Tilburg	STEVIN Brokerage and kick off STEVIN programme (160 participants)*
2005		
March 2	Antwerpen	STEVIN Brokerage - HLT and the ICT market (163 participants)*
November 22	Eindhoven	Taal in Bedrijf (290 participants)*
December 16	Amsterdam	CLIN 16th meeting of computational linguists in the Netherlands
December	Leuven	Symp. on Speech Technology for Clinical and Educational Applications
2006		
March 13,14	Delft	DIR 2006, 6th Dutch-Belgian Information Retrieval Workshop (TNO)
June 20	Utrecht	NoTaS speed dating session
September 10	Antwerpen	STEVIN programme meeting*
December 15	Nijmegen	HLT in the care sector in St. Maartens kliniek
2007		
January 12	Leuven	CLIN 17th meeting of computational linguists in the Netherlands
May 16	Amsterdam	Machine Learning for NLP 2007
June 11-22	Leuven	2007 LOT Summerschool
August 27-31	Antwerpen	Interspeech 2007
September 21	Hoeven	STEVIN programme meeting*
November 23	Antwerpen	Dutch HLT Agency meeting: de gebruiker central
December 7	Nijmegen	CLIN 18th meeting of computational linguists in the Netherlands
2008		
January 7-18	Tilburg	2008 LOT Winterschool
May 8	Utrecht	ICT Delta
May 14	Utrecht	Symposium Begrijpelijke Taalgebruik
May 22	Soesterberg	Resonansgroep NOTaS
June 26-27	Rotterdam	STEVIN midterm event*
September 11	Hoeven	STEVIN programme meeting*
November 19	Brussels	Taal in Bedrijf (158 participants)*
2009		
January 22	Groningen	CLIN 19th meeting of computational linguists in the Netherlands
January 23-24	Groningen	TLT workshop
April 24	Nijmegen	OSTT-symposium over taaltechnologie in de zorg
June 11-12	Groningen	TABU-dag 2009
May 26	Den Haag	Zorglandschap van Morgen (Flevum NV)
September 4	Tilburg	STEVIN programme meeting*
2010		
February 5	Utrecht	CLIN 10th meeting of computational linguists in the Netherlands
April 27-29	Utrecht	Vakbeurs Overheid & ICT 2010

List of publications about the STEVIN programme

- English brochure about the Dutch Language Union and HLT for Dutch
- a bilingual Dutch-English brochure about the STEVIN-programme

Publications in Dutch

- "Computer plaatst vragen in de juiste context", SenterNovem Innovatiekrant, 26 april 2006, p.8
- "Mens en machine dicht bij elkaar", SenterNovem Monitor 2006 (4): 7-9
- "De computer begint steeds meer mee te praten" door dr. Peter-Arno Coppen in Taalschrift 20/10/06
- Het Dixiteindejaarsnummer 2006 bevat een uitgebreide thematische STEVIN-sectie. DIXIT wordt uitgegeven door en is te verkrijgen via de Stichting NoTaS
- "Innovatief spraakherkenningssysteem als back-up voor justitie", SenterNovem Innovatiekrant 24 april 2007, p.12
- "Klinkende Taal voor Ambtenaren", SenterNovem Innovatiekrant 4 december 2007, p.17
- "Digi-revolutie in de rechtbank" in De Twentsche Courant Tubantia 11 mei 2007 (pdf-bestand)
- Bea Ross, "De computer luistert beleefd en geeft netjes antwoord", NWO Hypothese 2007, p. 18-20
- Niels Dekker, "Wanbetalers sneller gepakt" in Autowereld 03/07/2008
- Het Dixiteindejaarsnummer van 2008 bevat een uitgebreide thematische STEVIN-sectie. DIXIT wordt uitgegeven door en is te verkrijgen via de Stichting NoTaS
- Dixit speciale editie "STEVIN en onderwijs" 2009
- "Experiment NL: Wetenschap in Nederland", deel 2, 2009. Uitgave van NWO in samenwerking met Quest. Hierin worden 4 STEVIN demonstratieprojecten beschreven: AAP, Spelspiek, Web Assess, Primus