The Dutch-Flemish HLT Programme STEVIN: Essential Speech and Language Technology Resources

Elisabeth D'Halleweyn, Jan Odijk*, Lisanne Teunissen, Catia Cucchiarini

Nederlandse Taalunie
Lange Voorhout 19
2514 EB Den Haag
The Netherlands
edhalleweyn@taalunie.org; lteunissen@taalunie.org; ccucchiarini@taalunie.org

*: Universiteit Utrecht
Faculteit der Letteren, UiL-OTS
Trans 10
3512 JK Utrecht
The Netherlands
jan.odijk@let.uu.nl

Abstract

In 2004 a consortium of ministries and organizations in the Netherlands and Flanders launched the comprehensive Dutch-Flemish HLT programme STEVIN (a Dutch acronym for "Essential Speech and Language Technology Resources"). To guarantee its Dutch-Flemish character, this large-scale programme is carried out under the auspices of the intergovernmental Dutch Language Union (NTU). The aim of STEVIN is to contribute to the further progress of HLT for the Dutch language, by raising awareness of HLT results, stimulating the demand of HLT products, promoting strategic research in HLT, and developing HLT resources that are essential and are known to be missing. Furthermore, a structure was set up for the management, maintenance and distribution of HLT resources. The STEVIN programme, which will run from 2004 to 2009, resulted from HLT activities in the Dutch language area, which were reported on at previous LREC conferences (2000, 2002, 2004). In this paper we will explain how different activities are combined in one comprehensive programme. We will show how cooperation can successfully be realized between different parties (language and speech technology, Flanders and the Netherlands, academia, industry and policy institutions) so as to achieve one common goal: progress in HLT.

1. Introduction

In order to stimulate the technology sector so as to increase its innovative power, and in order to improve or at least secure the position of the Dutch language in the modern information and communication society, the Dutch and Flemish governments have launched the 5-year (2004-2009) Dutch-Flemish Language and Speech Technology Programme *STEVIN* (Spraak- en Taaltechnologische Essentiele Voorzieningen In het Nederlands).[1]

The STEVIN programme for Dutch language and speech technology is a coordinated effort of the Dutch Ministry of Economic Affairs, SenterNovem (an agency of the Dutch Ministry of Economic Affairs), the Netherlands Organisation for Scientific Research (NWO), the Dutch Ministry of Education, Culture and Science, the Flemish Institute for the Promotion of Innovation by Science and Technology (IWT), the Administration for Science and Innovation (AWI) of the Ministry of the Flemish Community, and the Flemish Fund for Scientific Research (FWO). To guarantee its Dutch-Flemish character, this programme is carried out under the auspices of the Nederlandse Taalunie (Dutch Language Union, an inter-governmental organization dedicated to promoting the Dutch language).

These funding parties, together with experts in the field, form the STEVIN Board responsible for supervision of the programme and for the funding decisions. The Taalunie co-ordinates the STEVIN programme and is responsible for the financial management. The Programme Committee consists of Dutch and Flemish experts in the field of language and speech technology, both from academia and industry. The Committee watches over the content: its members defined the STEVIN multi-year programme (Odijk et al., 2004), write the calls for proposals, and advise the Board on their funding decisions. The STEVIN Programme Office is run jointly by NWO Humanities Division and SenterNovem – it is responsible for the practical programme management.

The aim of the STEVIN programme is to contribute to the further progress of HLT for the Dutch language, more specifically to realise an appropriate digital language infrastructure for the Dutch language, based on well-defined priorities; carry out strategic research in the domains of language and speech technology, in particular in areas for which there is a large demand from specific applications and technologies; create networks and core research areas; promote the embedding of research and educate new generations of experts; encourage demand and knowledge transfer.

In trying to achieve these goals, the programme partners make a concerted effort based on three pillars:

1. raising awareness of HLT results and stimulating the demand of HLT products;
2. promoting strategic research in HLT and developing HLT resources that are essential and are known to be missing;
3. organising the management, maintenance and distribution of HLT resources once they are developed.

In section 3 of this paper we will explain how these different activities are combined in one comprehensive

[1] http://taalunieversum.org/stevin

programme. But first, in section 2, we will show how the programme is related to previous actions in the field of HLT that were carried out in the Dutch language area (the Netherlands and Flanders) in the past five years. The overview of the activities which were undertaken may be useful for other countries that are now starting similar initiatives. It shows how cooperation can successfully be realized between different parties (language and speech technology, Flanders and the Netherlands, academia, industry and policy institutions) so as to achieve one common goal: progress in HLT.

2. Background

The STEVIN programme is a direct result of the Dutch-Flemish HLT Platform project, which we reported on at previous editions of LREC (Beeken, Dewallef & D'Halleweyn, 2000; Binnenpoorte et al., 2002; Cucchiarini & D'Halleweyn, 2004). The roots were laid in 1999 by the decision of the Dutch and Flemish government to work closely together in matters concerning HLT for Dutch. The Nederlandse Taalunie took the initiative to install an HLT Platform, bringing together all Flemish and Dutch government bodies involved. Within the framework of the HLT Platform project a number of action lines were carried out which culminated in concrete achievements.

- The action line *creating a 'market place'* envisaged to encourage cooperation between all parties involved (Dutch and Flemish industry, academia and policy institutions), to raise awareness and give publicity to the results of HLT research so as to stimulate market take-up. Part of the activities were carried out in the framework of the European Euromap project.
- The action line *defining the BLARK (Basic LAnguage Resources Kit) for Dutch* envisaged to identify the basic elements that are required for a language to be HLT-enabled and to determine which of these essential elements were missing for Dutch. It resulted in a priority list for language technology and one for speech technology, specifying which parts of the BLARK appeared to be missing and should be developed. A complementary study, ordered by the Dutch Ministry of Economic Affairs, looked into the economic potential of HLT and investigated which forms of support would, in combination with the BLARK proposal, be most beneficial to the HLT sector.
- The action line *developing a blueprint for the management, maintenance and distribution of language resources* envisaged to determine what is needed to facilitate the re-use of digital language resources developed with government money.

The respective Dutch and Flemish policy institutions acknowledged the recommendations that resulted from the various action lines, and budgets were assigned for a comprehensive new HLT Programme, STEVIN. In line with the structure of the HLT Platform project, the STEVIN programme is organised around three action lines:

1. raising awareness;
2. strategic research and resource development to fill the infrastructural gaps;
3. management, maintenance and distribution of HLT resources.

These three lines of activities are presented in sections 3.1, 3.2, and 3.3, respectively.

3. The Dutch-Flemish HLT Programme STEVIN

3.1. Raising awareness and stimulating demand of HLT products

The first pillar of STEVIN focuses on creating visibility of the HLT sector, promoting cooperation, information exchange, and dissemination of research results. Thanks to the creation of the HLT Platform in 1999 and to the activities related to the Dutch-Flemish Spoken Dutch Corpus project and the Euromap project, cooperation and visibility have increased considerably in the Dutch-Flemish HLT field. A cooperative framework is now available that provides a forum for discussing, exchanging and sharing experiences, best practices, information, data and tools.

In 2001, a number of Dutch HLT providers and knowledge institutes established a foundation (NOTaS) to jointly protect their interests. The organization is very committed and functions as an important contact and sounding board for the Nederlandse Taalunie and the other government bodies.

Furthermore, Dutch and Flemish official bodies have been able to formulate a common agenda and to launch new initiatives together. The former partners of the HLT Platform now meet in the board of the STEVIN programme. Within STEVIN, a substantial budget is allocated for 'accompanying measures', e.g. conferences, demonstration projects, and network subsidies. The already existing cooperation is maintained and intensified through instruments that have proven to be successful, such as the HLT website[2], newsletters, and seminars.

3.1.1. Conferences and seminars

The challenge is to narrow the gap between technology and the market, and to address the end user. A large conference – a Dutch-Flemish counterpart of LangTech (2002 in Berlin, 2003 in Paris) – *Taal in Bedrijf* (intentionally ambiguous between "Language in Business" and "Language in Action"), was organized in November 2005 to bring together HLT and related-fields companies, as well as actual and potential users of speech and language technologies. In close cooperation with NOTaS, the conference programme was designed in such a way as to attract as many different potential professional users of HLT as possible, with 24 different business cases from sectors such as media, education, health care, transportation and logistics, tourism and recreation, public administration, telecom, and finance. The exhibition gave a large number of Dutch and Flemish HLT companies an important opportunity to showcase their businesses.

3.1.2. Demonstration projects

The aim of the demonstration projects is to set an example of successful HLT applications and thus stimulate the demand of Dutch HLT applications. These projects have a short duration (maximum 15 months).

[2] http://taalunieversum.org/tst

Essential characteristics of such projects are that they make use of "proven technology", that they try to access new markets for already established products or that they port established technologies to new domains. One of the key deliverables of each project is a sound and detailed dissemination plan.

A total budget of 1 million euros has been made available for such projects with a maximum of 100.000 euros STEVIN funding per project. Three calls for projects are scheduled. The first one with a maximum budget of 200.000 euros was launched in the fall of 2005. A second call (budget: 500.000 euros) will be launched in 2006 and a third one (budget: 300.000 euros) in 2007.

Nine proposals were submitted for the first call, three were selected for funding. The projects selected are:

- SNRT (Dutch acronym for "Speech-driven License Plate Retrieval Tool") aims to demonstrate that automating vehicle license plate retrieval using speech recognition technology can lead to improved service with reduced human effort.
- C-Content aims to extend the existing search facilities on a portal for legislation with language-supporting techniques so that the user can pose simpler questions, gets better answers and useful suggestions for related topics.
- Gemeenteconnect ("City Connect") aims to demonstrate that dialogue systems developed by the participating partners on the basis of Dutch language and speech technology are indeed suited to optimize the city's information provision to its citizens.

3.1.3. Market studies

A number of market studies will be carried out in the framework of the STEVIN programme so as to investigate new sectors in which HLT could play an important role. The surveys will look into the specific needs of different sectors with respect to HLT applications.

A first study, *Human language technologies and communicative disabilities*, was carried out in 2005 (Rietveld & Stolte, 2005). It was aimed at identifying the specific HLT-based tools that restricted language users require to improve their communicative capabilities, i.e. tools that assist verbal dialogue, reading and writing, and communication with machines. The long-term goal is to try to improve the position of these specific groups of users of Dutch. The study shows a world of very diverse desires, requirements, and possibilities – which helps explain why communicative disabilities arouse so little interest in the business sector. The diversity of disorders and requirements makes it impossible to develop products that everyone can use. Similar studies will be carried out for the education sector and public administration.

Another important goal of the accompanying measures is to try to motivate students to choose HLT-oriented studies, since one of the most pressing needs within the HLT sector is the need for highly qualified personnel. To this end, a market study was carried out to establish the best methods for introducing teenagers to the fascinating world of HLT and its various applications, which they often use in their everyday life without being aware of the HLT technologies inside. High school teachers from different disciplines as well as other education experts were interviewed to find out how HLT could be integrated within the school curriculum.

3.2. Strategic research, resource development, and application development

The first open call for strategic research, resource development, and application development was launched in September 2004, with a maximum budget of 2 M€. This was a call for focused short term projects with a self-contained result, and a maximum duration of two years.

A second open call was issued in the spring of 2005 for more complex projects with a longer time frame (maximum duration four years), combined with a call for tender for projects targeting some specific priorities (maximum duration two years). The total maximum budget was 3.4 M€.

Though originally three calls were planned, the remaining budget of 3.1M€ will possibly be split over a third and a fourth call (planned for late 2006 and 2007). The specific content of these calls remains to be determined by the Programme Committee.

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.

Only senior researchers at Flemish or Dutch knowledge institutions are eligible to apply, but cooperation with industry is encouraged. Cross-border consortiums are stimulated by increasing the standard bench fee by 50%.

The research proposals are assessed and ranked independently first by an International Advisory Panel (IAP) and then by the Programme Committee. Evaluation criteria are quality, innovative features and economic aspects of the project proposal, contribution to the STEVIN Programme, proper treatment of IPR, use of standards, prevention of duplication. Based on the Programme Committee's recommendations, the STEVIN Board finally formulates a binding advice to the Nederlandse Taalunie as to which projects will be funded.

3.2.1. First call for projects

The first call for projects was an open call. In total, 19 project proposals were submitted. The following five projects were elected for funding:

- AUTONOMATA (Automata for deriving phoneme transcriptions of Dutch and Flemish names) 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 (Yang et al., 2006).
- COREA (COreference Resolution for Extracting Answers) aims to develop a robust system for the resolution of coreferential relations in text.
- D-Coi (Dutch Language Corpus Initiative) 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 (van Noord, Schuurman & Vandeghinste, 2006; Oostdijk & Boves, 2006; Reynaert, 2006; Schuurman, Vandeghinste & van den Bosch, 2006).

- IRME (Identification and Representation of Multi-word Expressions) aims to develop innovative methods and tools for the automatic identification and lexical representation of multi-word expressions. (Grégoire 2006, Villada-Moirón 2005, Villada-Moirón & Tiedemann 2006, Villada-Moirón & Grégoire 2006).
- JASMIN-CGN aims at extending the Spoken Dutch Corpus (CGN) with read speech and human-machine dialogues of children, non-natives and elderly people in the Netherlands and Flanders (Cucchiarini et al., 2006).

In section 3.2.3, we will pay attention to the relevance of these projects with respect to the goals and criteria set out in the document describing the STEVIN multi-year programme (Odijk et al., 2004): distribution over language v. speech technology, distribution over the Netherlands v. Flanders, distribution over areas such as research v. resources v. applications, degree of participation of industry, et cetera. More details about the projects themselves, their content and methodology can be found on the STEVIN website (see footnote 1) and in the dedicated papers referred to above.

3.2.2. Second call for projects

The second call (total budget maximum 4.2 million euro) was launched in the spring of 2005. This call consisted of two parts: a call for tender targeting two specific priorities and an open call.

The priorities selected for the call for tender (max. 0,8 million euro) were the development of a speech recognition toolkit for Dutch and the development of a lexical resource for the semantic processing of Dutch. Four project proposals were submitted. Two projects (maximum duration two years) were selected: SPRAAK (speech recognition toolkit) en CORNETTO (semantic lexical resource).

The open call aimed at more complex projects with a longer time frame (maximum duration four years). In a first stage, 34 pre-proposals were submitted. Of these, 17 were recommended for elaboration into a full project proposal. In total, 18 full project proposals were submitted. Six projects were selected for funding:

- DAESO (Detecting And Exploiting Semantic Overlap) aims at developing technology for detecting semantic overlap between sentences and exploiting such technology in a range of NLP applications. The project's goal is to use parallel translations of texts into the same language to semi-automatically discover alternative formulations for the same semantic content, aiming at increased coverage and accuracy in three NLP applications: question answering, information extraction and advanced summarisation. The project will result in a set of tools and a 1 million monolingual parallel/comparable corpus.
- DPC (Dutch Parallel Corpus) aims at building a high-quality multifunctional and multilingual parallel corpus. DPC promises two sentence aligned bilingual corpora (Dutch-English and Dutch-French) with a portion aligned at a sub-sentential level as well.
- LASSY (LArge Scale SYntactic annotation of written Dutch) aims at developing a large (1 million words) corpus of written Dutch texts that is syntactically annotated and manually corrected. In addition, the complete D-Coi corpus (see first call) will be syntactically annotated automatically.
- MIDAS (MIssing DAta Solutions) aims at tackling the noise robustness problem in automatic speech recognition (ASR) through missing data techniques which enable masking out 'unreliable' parts of the speech signal (due to noise etc.) during the recognition process, assuming pre-defined models to make a better fit than matching data known to be corrupted.
- NBest (Nederlandse Benchmark Evaluatie van Spraakherkennings Technologie, "Dutch Benchmark Evaluation of Speech Recognition Technology") aims at developing an evaluation benchmark for large vocabulary continuous speech recognition in Dutch as spoken in Flanders and the Netherlands in two evaluation conditions: broadcast news and conversational telephony. This is an essential and recognised step to secure the position of the Dutch language in systems with speech recognition, and to give industrial partners a better insight into the position on the state of the art of the different systems for Dutch.
- STEVINcanPRAAT aims at the development of a number of improvements and added functionality to the PRAAT program that will become additionally and freely available for speech technologists via this open source program.

More details about these projects can soon be found on the STEVIN website.

3.2.3. A balanced programme

STEVIN aims to be a balanced programme, where "balancing" can and must be measured along various dimensions.

First, since STEVIN is a programme financed both by the Netherlands (2/3) and by Flanders (1/3), there should be a balance in this respect. In fact, with the current projects, the number of participants from the Netherlands is perfectly in balance with the number of participants from Flanders (67% v. 33%). In terms of person months (58% v. 42%) and money (60% v. 40%) we are close to the target but Flanders is slightly overrepresented. This does not mean that all is well, since Flanders is in fact overrepresented in the projects aiming at research and resource creation, but currently not represented at all in demonstration projects.

Second, STEVIN aims to stimulate both language technology and speech technology. The number of projects focusing on language technology is actually equal to the number of projects focusing on speech technology (50% v. 50%), but in terms of person months and money more is being spent on language technology than on speech technology (58% v. 42%). This probably reflects the fact that there are more researchers and companies in the Netherlands and Flanders focusing on language technology than on speech technology.

Third, STEVIN focuses on funding knowledge institutions but aims to stimulate a close cooperation

between knowledge institutions and industry. Indeed, 70% of the participants in the running projects are knowledge institutes, and 30% from industry, which shows a high degree of participation of industry in the programme. In terms of person months and money, most is spent on knowledge institutions (person months: 91% v. 9%; money: 88% v. 12 %), as was intended.

Fourth, STEVIN aims at stimulating strategic research, resource creation, and application development, preferably all together in a single project. Though strategic research (23%) and resource creation (54%), and also combinations of these in a single project (23%), are well-represented in the current programme, application development is not represented at all yet. Though it is natural that application development projects will be submitted only later in the programme after a number of strategic research and resource creation projects have yielded their results, this imbalance has the special attention of the Programme Committee and the Board.

Finally, also on a more detailed level a balance should be sought for in coverage of research topics, resource types, and applications. The STEVIN Programme Committee is currently assessing the situation in this respect, so that it can formulate a specific focus for the remaining call(s), planned for late 2006 and/or 2007. These new calls, together with the calls for demonstration projects, will also create opportunities to straighten out any other imbalances that currently exist.

3.3. Maintenance, distribution, IPR

To ensure that HLT resources developed with public funding become available for interested users (academia and companies) the Nederlandse Taalunie, as the owner of a number of these resources, took the initiative to set up the HLT Agency. The aim was to combine the infrastructures required for different projects, thus reducing the costs for equipment, data, software, licences, experts, and personnel, and at the same time to ensure optimal visibility and accessibility by offering resources through a one-stop-shop supplier.

In addition, to prevent HLT resources developed with public funding from lying unused on the shelf, it is necessary to make sure that they stay usable, which may entail debugging or updating for new platforms. All these activities concerning management, maintenance and distribution are carried out by the HLT Agency, which is hosted by the Institute for Dutch Lexicology. For further information on the activities of the HLT Agency, see other contributions at this conference (Boekestein, Depoorter & van Veenendaal, 2006; Wittenburg et al., 2006; Krauwer et al., 2006; Maks & Boelhouwer, 2006).

The HLT Agency also takes care of IPR issues. For this reason, it is involved in the evaluation and negotiation procedures concerning the STEVIN projects. As a matter of fact, all resources that are developed within the framework of STEVIN will become property of the Taalunie, which will make them available through the HLT Agency. In this way it can be ensured that all developed resources will become available for the whole language community in the Netherlands and Flanders.

4. Conclusions

Through cross-border cooperation within the STEVIN Programme, the Netherlands and Flanders are making significant progress in building a full-fledged language technology infrastructure. The combination of stimulating research and development, creating a landing site for the results, and raising awareness of these results amongst the prospective users turns out to be a fruitful and effective approach. The implementation of the programme with a Board, a Programme Committee, an International Advisory Panel and a Programme Bureau may seem somewhat heavy at first sight, but in fact provides a sound structure for dividing responsibilities and making conscientious decisions. Halfway through the programme, STEVIN turns out to be well-balanced in most respects; any remaining imbalances will be straightened out in the calls yet to come. We hope that our efforts in developing a model to make a language "technology-ready" may prove useful to other countries.

5. References

Beeken, J., Dewallef, E., D'Halleweyn, E. (2000). A Platform for Dutch in Human Language Technologies, In *Proceedings LREC 2000*.

Binnenpoorte, D., Cucchiarini, C., D'Halleweyn, E., Sturm, J., Vriend, F. de (2002). Towards a roadmap for Human Language Technologies: Dutch-Flemish experience. In *Proceedings LREC 2002*.

Boekestein, M., Depoorter, G., Veenendaal, R. van (2006) Functioning of the Centre for Dutch Language and Speech Technology (TST Centre). In *Proceedings LREC 2006.*

Cucchiarini, C., D'Halleweyn, E. (2004). The new Dutch-Flemish HLT Programme: a concerted effort to stimulate the HLT sector. In *Proceedings LREC 2004*.

Cucchiarini, C., Van hamme, H., Houben-van Herwijnen, O., Smits, F. (2006). JASMIN-CGN: Extension of the Spoken Dutch Corpus with speech of elderly people, children and non-natives in the human-machine interaction modality. In *Proceedings LREC 2006*.

Grégoire, N., Elaborating the Parameterized Equivalence Class Method for Dutch. In *Proceedings LREC 2006.*

Krauwer, S., Quasthoff, U., Goddijn, S., Odijk, J., Choukri, K., Calzolari, N., Maegaard, B., Cieri, Ch., Huang, C., Tokunaga, T., Hoege, H., Heuvel, H. van den, Gibbon, D., Choi, K.S., Asmussen, J. (2006) Quality assurance and quality measurement for language and speech resources. In *Proceedings LREC 2006*.

Maks, I., Boelhouwer, B. (2006) Exploring opportunities for enrichment by linking lexical databases. In *Proceedings LREC 2006.*

Noord, G. van, Schuurman, I., Vandeghinste, V. (2006). Syntactic Annotation of Large Corpora in STEVIN. In *Proceedings LREC 2006*.

Odijk, J. et al. (2004). Vlaams-Nederlands meerjarenprogramma voor Nederlandstalige taal- en spraaktechnologie STEVIN: Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands. Nederlandse Taalunie, Den Haag.

Oostdijk, N., Boves, L. (2006). User requirements analysis for the design of a reference corpus of written Dutch. In *Proceedings LREC 2006*.

Reynaert, M. (2006). Corpus-Induced Corpus Clean-up. In *Proceedings LREC 2006.*

Rietveld, T., Stolte, I. (2005). Taal- en spraaktechnologie en communicatieve beperkingen. Nederlandse Taalunie, Den Haag.

Schuurman, I., Vandeghinste, V., Bosch, A. van den (2006). Transferring PoS-tagging and lemmatization tools from spoken to written Dutch corpus development. In *Proceedings LREC 2006.*

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).*

Villada Moirón, B., Grégoire, N. (2006). Quantifying and qualifying lexicalized and idiomatic expressions. Abstract accepted for the *Collocations and Idioms 1: The First Nordic Conference on Syntactic Freezes*.

Villada Moirón, B., Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word-alignment. Paper at the EACL 2006 *Workshop on Multiword expressions in a multilingual context*.

Yang, Q., Martens, J.P., Konings, N., Heuvel, H. van den (2006). Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme-to-phoneme (g2p) conversion of names. In *Proceedings LREC 2006.*

Wittenburg, P., Veenendaal, R. van, Johnson, H., Barwick, L. (2006). Towards a Research Infrastructure for Language Resources. In *Proceedings LREC 2006*.