



Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands
(Substantial language and speech technology resources for the Dutch language)

**A Flemish/Dutch programme to stimulate the language and speech technology (LST)
sector in the Dutch language area (Flanders and the Netherlands)**

Ladies and Gentlemen,

I have the rather difficult task of presenting in maximum 20 minutes the complex programme STEVIN - a Dutch-Flemish programme to stimulate the language and speech technology (LST) sector in the Dutch speaking area.

The STEVIN-programme is carried out under the auspices of the *Dutch Language Union (Nederlandse Taalunie)* so let me first introduce my organisation, as its existence turned out to be beneficial in shaping an ambitious cross-border programme.

1. The Dutch Language Union

As you know, the geographical border between the Netherlands and the Flemish part of Belgium is not a linguistic border. Language policy decisions taken on one side of the national border affect citizens at the other side of the border. This observation led to the creation of the Dutch Language Union. The Dutch Language Union is an intergovernmental organisation, based on the Language Union Treaty between Belgium and the Netherlands in 1980. The Dutch Language Union's mission is to deal with all issues concerning the position of the Dutch language. Last year the South American country Surinam also joined the Dutch Language Union, since Dutch is the official language in Surinam.

The Committee of Ministers, composed of the Flemish and Dutch ministers for Education and Culture, is responsible for the policy of the Dutch Language Union.

To carry out its mission, the Dutch Language Union follows a pragmatic, instrumental approach to language policy. In the new policy plan for the next 5 years, the emphasis is put on the language user. This means for instance that the Dutch Language Union considers it a vital task to enable all speakers of the Dutch language to use their own language in all situations, at home as well as at work.

In order to preserve the position of the Dutch language in the information society, it is therefore of utmost importance that the Dutch language can catch up with the further development and application of language and speech technologies. The Dutch Language Union therefore wants to ensure that the appropriate resources, knowledge, infrastructure and tools become available for the efficient and effective use of Dutch under all circumstances by all categories of users. It is estimated that 21 million people have Dutch as their native language. They should all be able to work in a

Dutch word processing environment, consult digital information services in Dutch or communicate in Dutch with their car or fridge...

2. Background

For the development of applications and products for Dutch, basic provisions are required. Therefore, in 1998, the Dutch language Union ordered a survey of the position of Dutch in LST. The survey made clear that the development of the basic material that is lacking, is an expensive undertaking which exceeds the capacity of the individuals involved. Collaboration between the various agents (policy, academia and industry) in the Netherlands and Flanders is required.

The Dutch Language Union consequently took the initiative to install a platform for Dutch in language and speech technology (abbreviated *LST-platform*), that brought together all Flemish and Dutch government bodies involved. This LST-platform provided the necessary Flemish/Dutch organisational and financial framework to carry out the “Action plan for Dutch in language and speech technology”, that was outlined following the results of the survey mentioned above.

Within this action plan, three concrete and clearly defined action lines were to be realised:

1. *Networking and creating a market place.*
2. *Defining the BLARK (Basic LAnguage Resources Kit) for Dutch”.*
3. *Developing a blueprint for the management, maintenance and distribution of language resources developed with government money.*

- 1- The first action line “*Networking and creating a market place*” envisaged to encourage co-operation between all parties involved (industry, academics and policy institutions), to raise awareness and give publicity to the results of LST-research so as to stimulate market take-up. It contributed to creating transparency in the LST-field in Flanders and the Netherlands, and managed to improve communication between interested partners. A co-operative framework is since then available (and kept on growing ever since), providing a forum for discussing, exchanging and sharing experiences, best practices, information, data and tools. For this purpose, a centralised *LST-infodesk* and website (www.taalunieversum.org/tst - information in Dutch only) is being managed and

maintained by the Dutch Language Union on a continuous basis.

- 2- The second action line focused on *“Defining the BLARK”*. A BLARK is a so-called *Basic Language Resources Kit*, the total of digital language resources that should be available for a language in order to play a part in LST. Traditionally a distinction is made between language technology resources (such as text corpora, lexica or parsers) and speech technology resources (such as a speech recogniser or a text-to-speech converter). The purpose of the action line was to identify to what extent the BLARK exists for Dutch and which elements are still missing. It resulted in a BLARK for language technology and a BLARK for speech technology. Priority then was assigned to the development of those parts of the BLARKs that are known to be crucial and appeared to be missing. Final priority lists were submitted to the various policy institutions involved.

- 3- The third action line *“Developing a blueprint for the management, maintenance and distribution of language resources developed with government money”* aimed at establishing a technological, organisational and legal framework in order to guarantee proper maintenance and distribution of LST-resources developed by means of governmental funding. Particular focal points were technological standards and legal issues such as Intellectual Property Rights (IPR).

The second and third action lines resulted in reports that were both finalised in 2002.

In addition to these reports, the Dutch Ministry of Economic Affairs issued a study particularly aiming at determining additional forms of economic support for the LST-sector. The following questions were to be answered.

- To what extent does the LST-sector contribute to achieving sustainable economic growth in Flanders and the Netherlands?
- Does this technology offer opportunities for the Flemish and Dutch economies?
- How does the LST-innovation system work?
- How can government intervene so as to improve its functioning?
- Which form of governmental intervention will be supported by both the industrial and the academic world?

The study concluded that the LST-sector does in fact have potential for the economy in the Dutch language area. The analysis of the LST-innovation system revealed that the optimal form of support for the LST-sector should envisage three types of funding flows:

- one to realise the prioritised resources and thus complete the digital language infrastructure for Dutch, as defined in the “BLARK for Dutch” survey;
- one to stimulate the demand for LST-products and applications. Since Flemish and Dutch service organisations do not appear to be eager to implement LST-products unless they are very convinced that the gains in efficiency will be at least twice the investment costs, the government should take measures to stimulate the demand of such products;
- one to stimulate academic LST-research along the lines indicated by LST-companies. Stimulating strategic research (and thus increasing expertise) will enhance the profit deriving from investments in resource development, and will lead to new ideas and insights and - possibly - to new start-ups.

In response to the outcome of both the “BLARK for Dutch” survey and the study ordered by the Dutch Ministry of Economic Affairs, the responsible Flemish and Dutch government bodies decided to combine their funding capacities for LST in one comprehensive R&D and stimulation programme, known as STEVIN.

3. STEVIN: partners and organisational structure

The STEVIN-programme was launched in 2004. The acronym stands for “Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands”, which can be described in English as “Substantial language and speech technology resources for the Dutch language”.

STEVIN is a co-ordinated effort of the following policy organisations: in Flanders:

- the Science and Innovation Administration (*administratie Wetenschap en Innovatie - AWI*) of the Ministry of Flanders,
- the Institute for the Promotion of Innovation by Science and Technology in Flanders (*Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen - IWT-Vlaanderen*) and
- the Fund for Scientific Research - Flanders (*Fonds voor Wetenschappelijk Onderzoek - Vlaanderen - FWO-Vlaanderen*);

in the Netherlands:

- the Ministry of Education, Culture and Science (*Ministerie van Onderwijs, Cultuur en Wetenschap - OCW*),
- the Ministry of Economic Affairs (*Ministerie van Economische Zaken - EZ*),
- the Dutch Organisation for Scientific Research (*Nederlandse Organisatie voor Wetenschappelijk Onderzoek - NWO*) and
- *SenterNovem*, the agency of the Ministry of Economic Affairs.

STEVIN has a five-year duration and its total budget amounts up to 11,4 million euro. The Flemish partners mentioned above provide 1/3 of this budget (or 3,8 million euro); the Dutch partners (except for SenterNovem) equally share the remaining 2/3 of the budget (or 7,6 million euro).

In order to fully preserve and guarantee its Flemish/Dutch character, STEVIN is carried out under the auspices of the Dutch Language Union. Since the Dutch Language Union is a Flemish/Dutch governmental body, a rigorous geographical distribution of financial means can be avoided, allowing for STEVIN to be managed and carried out as a full-fledged cross-border programme, thereby having regard to the 1/3 - 2/3 proportion between Flanders and the Netherlands.

The organisational structure of the STEVIN-programme further comprises of (1) the Language and Speech Technology Board (abbreviated LST-Board), (2) the STEVIN Programme Committee and (3) the STEVIN Programme Office.

- The LST-Board consists of the funding parties in Flanders and the Netherlands I mentioned before, complemented with a Flemish and Dutch senior expert (who both have an advisory function) and the chairman of the STEVIN Programme Committee (also with an advisory function). The LST-Board gives the Dutch Language Union a “binding advice” on all final funding decisions to be made, based on the input given by the STEVIN Programme Committee.
- The STEVIN Programme Committee, formally installed by the LST-Board, outlines and watches over the content of the STEVIN-programme. It elaborated the actual multi-annual programme as well as the programme’s calls for project proposals. The STEVIN Programme Committee consists of 12 LST-experts, well balanced between Flanders and the Netherlands, language technology and speech technology, and academics and industrialists.
- The STEVIN Programme Office consists of the Dutch Organisation for Scientific Research and SenterNovem. It is responsible for the practical management of the R&D-projects carried out within the

STEVIN-programme, and for the (administrative) support of the STEVIN Programme Committee and the LST-Board. It furthermore co-ordinates and organises the STEVIN-programme's "Accompanying Measures" (see section 4.1 below).

4. STEVIN: programme overview

The aim of the STEVIN-programme is to contribute to the further progress of LST for the Dutch language and in that way to stimulate the innovative power of the technology sector and to strengthen the position of the Dutch language in the modern information and communication society.

More specifically the aim of STEVIN is to:

- realise an appropriate digital language infrastructure for the Dutch language, based on the defined BLARK for Dutch mentioned earlier;
- carry out strategic research in the field of language and speech technology, particularly in areas with a high demand for concrete LST-tools and -applications;
- further stimulate the creation of LST-networks and core research areas;
- promote the embedding of LST-research, educate new generations of LST-experts, and further encourage knowledge transfer and the demand for LST.

In trying to achieve its goals, the STEVIN-partners are making a concerted effort based on three pillars:

- 1) raising awareness of LST-results;
- 2) stimulating the demand for LST-products and -applications, promoting strategic research in the field of LST, and developing LST-resources that are essential and are known to be missing (see further below);
- 3) organising the management, maintenance and distribution of LST-resources once they are developed.

I will now comment on the first two pillars. The third pillar is the responsibility of the Flemish/Dutch *Agency for Language and Speech Technology*, abbreviated *LST-Agency*. All LST-resources developed with governmental funding in Flanders and the Netherlands are to be transferred to the LST-Agency. The LST-agency provides a technological, infrastructural and legal framework for the management, maintenance and distribution of such resources. The STEVIN-programme is clearly linked with the LST-Agency since all

resources developed within STEVIN will also be transferred to the LST-Agency once they are completed. The project leader of the LST-Agency also partakes in the STEVIN Programme Committee.

4.1. Awareness

The first pillar of the STEVIN-programme focuses on co-operation, information dissemination and creating visibility. The LST-field in both Flanders and the Netherlands has seen a growing tendency to co-operation. In the nineties there still was a considerable lack of mutual visibility between Flanders and the Netherlands, and also between academia, policy and industry. A number of initiatives taken since then have led to clear changes.

- The Flemish/Dutch project *Spoken Dutch Corpus* brought together researchers from both Flanders and the Netherlands, and from both language technology and speech technology. Together they compiled a digital database for present-day Dutch as spoken by adults in Flanders and the Netherlands.
- The already mentioned LST-platform enabled the Flemish and Dutch responsible government bodies to formulate a common agenda and to commonly launch new initiatives.
- And as far as industry goes, a large part of the SME's active in the LST-field in the Netherlands set up a foundation (NOTaS) in order to jointly look after their interests. Their philosophy is that although they are competitors, they are nevertheless often each other's technology or data suppliers. Moreover, together they are stronger in a rather new, still developing and establishing market. NOTaS is currently undertaking the initiative to recruit Flemish LST-related SME's as well.

STEVIN gratefully uses all such networks and initiatives to maintain and further intensify the already existing co-operation. Instruments are the already mentioned LST-infodesk, newsletters and the organisation of seminars and conferences on a regular basis. A large part of these awareness actions is carried out by the Dutch Language Union.

A major challenge in this respect remains to further narrow the gap between technology and the market. Only if the - potential - end user can be addressed and stimulated, all prophecies of possible LST-benefits can be fulfilled, and the economic potential of LST that has been predicted can be exploited. In order to stimulate the demand for language and speech technology, and also to ensure that all players at the technological development and application layers are

represented in the STEVIN-programme, particular funding opportunity is provided within the context of the STEVIN-programme's "Accompanying Measures". These measures include the following:

- an open and continuous call for demonstration projects (with a maximum budget of 1 million euro), each envisaging the development of a specific application that will actually be deployed, possibly in a pilot/test phase. These applications should be appealing and (made) highly visible to the general public or to decision makers within companies and/or governmental bodies involved;
- the organisation on a regular basis of symposia, meetings and other publicity and/or networking events;
- the (financial) support of STEVIN-related events. Organisers of such events can apply for so-called "networking funding" within the STEVIN-programme.

4.2. Strategic research and resource development

As far as research and development goes, STEVIN is divided into three calls for project proposals. There are "open calls" for project proposals responding to required priorities and "closed calls" for tender proposals responding to a specific R&D-requirement. Each proposal can relate to basic linguistic resources (tools and data), to fundamental strategic research or to applications. Proposals can be submitted both in the area of language technology and in the area of speech technology. All projects have to contribute to an appropriate digital language infrastructure for Dutch. Exclusively senior researchers at Flemish or Dutch knowledge institutes are eligible to apply. It is particularly to the advantage of a project proposal if the expertise of Flemish and Dutch R&D-groups or companies are combined, if R&D-institutes and companies jointly make a proposal, or if the proposal relates both to language and to speech technology.

The STEVIN Programme Committee carries out the selection of research proposals. The assessment and ranking by this committee is then commented upon by an International Advisory Panel (IAP). The assessment by the STEVIN Programme Committee, as well as the comments formulated by the IAP, are the basis for the final decision by the LST-Board.

- The first call for proposals was issued on 15 September 2004. It envisaged rather small short-term projects with a self-contained

result and with a maximum duration of two years. This call resulted in a portfolio of five R&D-projects, totalising a budget of 2.052.225 euro (see annex).

- The second call for proposals was issued on 2 March 2005. It has a maximum budget of 4,2 million euro. It envisages more complex and larger consortium projects with a longer duration (maximal four years). There is an open call for project proposals and a closed call for tender proposals. All proposals submitted within both calls are currently being assessed and ranked by the STEVIN Programme Committee and the International Advisory Panel (IAP). In December 2005 the LST-Board will decide upon actual funding.
- The third call for proposals (with a maximum budget of 2,3 million euro) will be elaborated in 2007, a/o based on the results of the first two calls for proposals.

Within the STEVIN programme, there is a strong emphasis on the ability to access, use and exploit the basic resources resulting from the STEVIN R&D-projects on non-discriminative terms. The applicants have to declare themselves willing to negotiate on this matter with the LST-Agency. Conclusion of a contract on the IPR-arrangements is a necessary condition for actual funding.

5. Conclusion

I hope that my presentation has showed how cross-border co-operation can be realised in a successful way between different parties (language and speech technology, Flanders and the Netherlands, the academic world, industry and policy institutions) so as to achieve one common goal: progress in LST.

All information regarding the STEVIN-programme is available on the STEVIN-website (www.taalunieversum.org/stevin - information in Dutch only).

Annex: project portfolio resulting from the first STEVIN-call for proposals

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

Allocated budget: 322.848 euro

Project co-ordinator

Prof. dr. ir. J.-P. Martens
Universiteit Gent
ELIS Speech Lab
Sint-Pietersnieuwstraat 41
B-9000 Gent
Phone: +32 9 264 33 95
E-mail: Jean-Pierre.Martens@elis.ugent.be
URL: <http://www.elis.ugent.be/>

Project consortium

- Prof. dr. ir. J.-P. Martens (Universiteit Gent, ELIS Speech Lab)
- Dr. H. van den Heuvel (Radboud Universiteit Nijmegen, Centre for Language and Speech Technology - CLST)
- Dr. ir. G. Bloothoofd (Universiteit Utrecht, Utrecht institute of Linguistics - UiL-OTS)
- Ir. L. Peirlinckx (TeleAtlas)
- Dr. ir. J. Verhasselt (ScanSoft Belgium BVBA)

Project description

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.

Coreference Resolution for Extracting Answers (COREA)

Allocated budget: 353.875 euro

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
Phone: +31 50 363 59 37
E-mail: gosse@let.rug.nl
URL: <http://www.let.rug.nl>

Project consortium

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

Project description

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.

Dutch Language Corpus Initiative (D-coi)

Allocated budget: 566.531 euro

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
Phone: +31 24 361 57 85
E-mail: n.oostdijk@let.ru.nl
URL: www.let.ru.nl

Project consortium

- Dr. N. Oostdijk (Radboud Universiteit Nijmegen, Centre for Language and Speech Technology - CLST)
- Dr. A. van den Bosch (Universiteit Tilburg, Induction of Linguistic Knowledge - ILK)

- Prof. dr. W. Daelemans (Universiteit Antwerpen, Centrum voor Nederlandse Taal and Spraak - CNTS, en Universiteit Tilburg, Induction of Linguistic Knowledge - ILK)
- Drs. Th. van den Heuvel (Polderland Language and Speech Technology BV)
- Prof. dr. F. de Jong (Universiteit Twente, Human Media Interaction - HMI)
- Dr. P. Monachesi (Universiteit Utrecht, Utrecht institute of Linguistics - UiL-OTS)
- Dr. G. van Noord (Rijksuniversiteit Groningen, Alfa-informatica)
- Prof. dr. F. Van Eynde (Katholieke Universiteit Leuven, Centrum voor Computerlinguïstiek - CCL)

Project description

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 made available through the Flemish-Dutch HLT Agency (*TST-centrale*).

Extension of CGN with speech of children, non-natives, elderly and human-machine interaction (JASMIN-CGN)**Allocated budget: 419.471 euro****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
Phone: +31 24 361 57 85
E-mail: C.Cucchiarini@let.ru.nl
URL: www.let.ru.nl

Project consortium

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

Project description

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.

Identification and Representation of Multi-word Expressions (IRME)

Allocated budget: 389.500 euro

Project co-ordinator

Prof. Dr. J. Odijk
Universiteit Utrecht
Faculteit der Letteren
Utrecht institute of Linguistics OTS (UiL-OTS)
Trans 10
NL-3512 JK Utrecht
Phone: +31 30 253 60 76
E-mail: jan.odijk@let.uu.nl
URL: <http://www-uilots.let.uu.nl>

ScanSoft Belgium BVBA
International Headquarters
Guldensporenpark 32, Gebouw D
B-9820 Merelbeke
Phone: +32 9 239 80 00
E-mail: jan.odijk@scansoft.com
URL: <http://www.scansoft.com>

Project consortium

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

Project description

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.