## A Platform for Dutch in Human Language Technologies

#### Elisabeth D'Halleweyn

Nederlandse Taalunie Postbus 10595 2501 HN Den Haag The Netherlands Edhalleweyn@ntu.nl http://www.taalunie.org

#### **Erwin Dewallef**

Nederlandse Taalunie C/o Ministry of the Flemish Community Science and Innovation Administration Boudewijlaan 30 B-1000Brussel Belgium <u>Erwin.dewallef@wim.vlaanderen.be</u> http://www.innovatie.vlaanderen.be

### Jeannine Beeken

Nederlandse Taalunie Postbus 10595 2501 HN Den Haag The Netherlands jeannine.beeken@skynet.be

#### Abstract

As ICT increasingly forms a part of our daily life it becomes more and more important that all citizens can make use of their native languages in all communicative situations. For the development of successful applications and products for Dutch basic provisions are required. 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, knowledge infrastructure and industry) in the Netherlands and Flanders is required. The existence of the Dutch Language Union (Nederlandse Taalunie) facilitates this co-operation. The responsible ministers decided to set up a Dutch-Flemish platform for Dutch in Human Language Technologies. The purpose of the platform is the further construction of an adequate digital language infrastructure for Dutch so that the industry develops the required applications which must guarantee that the citizens in Holland and Flanders can use their own language in their communication within the information society and the Dutch language area remains a full player in a multi-lingual Europe. This paper will show some of the efforts that have been taken.

#### 1. Introduction

The far-reaching integration between information and communication technology (ICT) has led to a society in which information has become a central element. ICT is of crucial importance for future competitive capability in all branches of industry, trade and service provision. Human Language Technologies (HLT) enable the user to use natural language to communicate with the computer. As ICT increasingly takes a part in the daily life it becomes more and more important that the users can make use of their native languages.

The fast moving ICT developments affect language both as a *tool* and as a *medium* (i.e. the language in which information is provided). As such, ICT developments can affect in the long run the international *position* of the Dutch language in relation to other languages.

For the development of successful applications and products for Dutch basic provisions are required such as corpora, electronic dictionaries, terminology databases, etc. In order to allow Dutch to play a full role in the multilingual information society it is important that an effort is made to develop the basic resources that are lacking. This is an expensive undertaking which exceeds the capacity of the individuals involved. Collaboration between the various agents (policy, knowledge infrastructure and industry) in the Netherlands and Flanders is required.

This paper will show how the collaboration between the Netherlands and Flanders at a policy level has taken shape, how the existence of the Dutch Language Union (Nederlandse Taalunie) facilitated this co-operation and what concrete actions have been/will be taken.

### 2. Platform for Duch in Human Language Technologies

### 2.1 The need for Flemish-Dutch co-operation

As the geographical border between the Netherlands and the Flemish part of Belgium is not a linguistic border, the governments on both sides share a concern for the Dutch language. The observation that language policy decisions taken on one side of the national border affect citizens at the other side of the border led to the creation of the Dutch Language Union (Nederlandse Taalunie – abbreviated NTU). The NTU is an intergovernmental organisation, based on the Language Union Treatment between Belgium and the Netherlands in 1980. The NTU's mission is to deal with all issues concerning the position of the Dutch language, which includes Dutch language and literature as a scientific subject, Dutch as a cultural language, Dutch as an administrative language, Dutch as a means of communication, and, more in general, the Dutch language as a tool for social interaction. The NTU enables Flanders and The Netherlands to speak with a single voice in the international arena. The Committee of Ministers, composed of the Flemish and Dutch ministers for Education and Culture, is responsible for the policy of the NTU. When establishing its current long-range policy plan (1998 - 2002), the NTU has taken the following major developments into consideration: the increasing internationalisation, highly influencing language, education and culture; the rapid developments in the field of ICT, with a major impact on language issues.

In view of the preservation of the Dutch language's position in our rapidly evolving multilingual and international information society, the governments of the Netherlands and Flanders recognised the growing importance of Human Language Technologies (HLT) as a specific part of information technology. Keeping up with the technological developments in this field implies major investments and the commitment of those involved, notably the policy makers on a national and European level, the knowledge infrastructure and the business community. Collaboration between all these actors is of utmost importance, and given the size of the Dutch language region, this collaboration also needs to be expanded to a cross-border Flemish-Dutch level. Building on this awarness, two large HLT projects that were initiated over the last years not only have a Flemish-Dutch character but also try to combine expertise from the research community as well as of the business community.

The *Spoken Dutch Corpus Project* is a five-year project aimed at the compilation and annotation of a 10-million-word corpus of contemporary standard Dutch as spoken in the Netherlands and Flanders. The project is funded jointly by the Dutch and Flemish governments. Project activities are co-ordinated from two sites: one in Flanders and one in the Netherlands. The copyright to the Spoken Dutch Corpus is owned by the NTU who will be responsible for the exploitation of the results.

*NL-Translex* is a project aimed at the development of the language pairs Dutch - English/French and English/French - Dutch in systems for machine translation for unrestricted text. The development of these components takes place within the framework of MLIS. The project is funded jointly by the European Commission, the Dutch Language Union, the Dutch Ministry of Education, Culture and Science, the Dutch Ministry of Economic Affairs and the Flemish Institute for the Promotion of Scientific and Technological Research in Industry. The components to be developed are intended for use by the translation services of official bodies of the EU Member States and by the translation services of the European Commission.

Other papers of LREC2000 provide a more detailed description of these two projects.

# **2.2** The need for structural co-operation: the creation of a platform

In the project preparation of the Spoken Dutch Corpus as well as of NL-Translex much time was invested in the search for the appropriate responsible (funding) bodies as it was not clear who was responsible for the construction of a digital language infrastructure for Dutch. This observation was confirmed in several surveys that were conducted over the last years. The market research carried out in the Netherlands and in Flanders in the of EUROMAP framework and the research commissioned by the NTU into the position of Dutch in Language and Speech Technology (report Bouma & Schuurman, 1998) pointed out that the fragmentation of responsibilities made it difficult to conduct a coherent policy and meant that the field lacked transparency for interested parties.

In order to create more transparency and to give shape to the collaboration in the field of HLT the NTU took the initiative to install a *Flemish-Dutch platform to support the Dutch language in HLT* (hereafter referred to as the **HLT-Platform**).

Apart from the NTU, the HLT-Platform consists of the following partners. Within Flanders: the Ministry of the Flemish Community, represented by the Science and Innovation Administration (Administratie Wetenschap en Innovatie, abbreviated AWI), the Flemish Institute for the Promotion of Scientific-technological Research in Industry (Vlaams Instituut ter Bevordering van het Wetenschappelijk-technologisch Onderzoek in de Industrie, abbreviated IWT) and the Fund for Scientific Research - Flanders (Fonds voor Wetenschappelijk Onderzoek - Vlaanderen, abbreviated FWO-Vlaanderen). Within The Netherlands: the Ministry of Education, Culture and Sciences, the Ministry of Economic Affairs, the Netherlands Organisation for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, abbreviated NWO) and Senter.

The different organisations approach HLT from different angles as they each have their own aims and responsibilities: e.g. the competitive position of Dutch or Flemish companies, the competitive position of Dutch or Flemish research, the concern for the Dutch speaking user. Together they cover the broad spectrum of angles involved.

In creating the HLT-Platform it was not the intention to create a new structure but to co-ordinate existing structures. The platform is to be seen as a flexible construction within which the partners adjust the agendas in relation to HLT to each other and decide whether to place new subjects on a common agenda. The necessary collaboration with the field is designed by creating a permanent advisory body and setting up ad hoc work groups. In the first instance the platform is set up for a period of five years (1999-2004).

### 2.3 Purpose and objectives of the HLT-Platform

The purpose of the platform is the further development of an adequate digital language infrastructure for Dutch so that the industry develops the required applications which must guarantee that the citizens in the Netherlands and Flanders can use their own language in their communication within the information society and that the Dutch language area remains a full player in a multi-lingual Europe. More specifically, the HLT-Platform has the following objectives:

- To promote the position of the Dutch language in HLT developments, so that the Dutch language can become and remain a "full" language within a multilingual European information society;
- To establish the proper conditions for a successful management and maintenance of basic HLT resources developed with governmental funding;
- To promote and stimulate the collaboration between research community and business community in the field of HLT;
- To contribute to the realisation of European collaboration in HLT-relevant areas;
- To establish a network both electronically and personally that brings together demand and supply of knowledge, products and services.

### 3. Action Plan

Even if the Netherlands and Flanders co-operate in funding the development of basic language resources, the investments for the different partners involved remain substantial. Therefore the need is felt to have an insight in the kind of essential provisions that are needed for a language in general and to draw up a plan for the development of provisions that are lacking for Dutch.

Apart from investing in the development of basic language resources, attention should also be given to such matters as evaluation of project results, standardisation, maintenance, distribution etc. This means that, in order to maximise the existing and future results, the necessary preconditions must be created. Therefore the HLT-Platform defined an *Action plan for Dutch in language and speech technology* aimed at initiating a number of short-term actions to this end. The Action plan is funded jointly by the different partners of the HLT-platform.

#### **3.1** Key Action A: construction of a 'bridgebuilding' function

As was already mentioned above, the lack of collaboration between the various parties is experienced as a serious difficulty. The HLT-Platform will wherever

possible put the importance of the position of Dutch HLT on the agenda and organise working meetings, colloquia and information days. In this way the platform can be instrumental in far-reaching collaboration between knowledge infrastructure and industry in the area of HLT. The aim is to construct a network, both electronically and in the personal sphere, that effectively brings together demand and supply of knowledge, products and services.

The NTU, as a representative of the HLT-platform, is one of the focal points of HOPE (Human Language Technology Opportunity Promotion in Europe).

HOPE is a multi-country, shared-cost accompanying measure project of the IST-Programme of the European Commission, spanning two years. It aims to provide awareness, bridge-building and market-enabling services to boost opportunities for market take-up for the results of national and European HLT RTD. The key focus is on helping to accelerate the volume of HLT transfer from the research base to the market by creating communities of interest between the critical players in the development and value chain. The aims of HOPE clearly coincide with the aims of Key Action A.

One of the main activities of both programmes (i.e. of Key Action A and of HOPE) is the construction of an extensive informational website on the HLT sector in Flanders and the Netherlands - with multiple European links, a/o related to the Fifth Framework Programme of the European Commission. According to the planning the website should be opened towards the beginning of May 2000. The website will provide actual information on all relevant actors in the field of HLT (i.e. R&D people, technology and application developers, end users and policy makers) on a cross-border Flemish-Dutch as well as a European level. This includes detailed information on how the HLT sector evolves (implying a continuous follow-up of national and international HLT-related news originating from various channels) and on HLT-related events throughout Europe. For - obvious - reasons of efficiency, this website builds on the Flemish HLTrelated website that was developed during the EUROMAP II project and that will be lifted to a crossborder Flemish-Dutch as well as a European level. The website - presented in Dutch and English - will serve as the core HLT Infodesk of the NTU, complying with the objectives and tasks ensuing from both the HLT-Platform and HOPE.

Related to both fields, the NTU will establish and maintain specific contacts databases, for which purpose existing field-related contacts that are already included in the Flemish and Dutch HLT contacts databases will be utilised in a network-building way.

## **3.2 Key Action B: plan to strengthen the digital language infrastructure**

Both for the development of successful (commercial) applications and the carrying out of (strategic research) basic materials are required. The purpose of Action line B is to define what the so-called BLARK (Basis LAnguage Resources Kit) for Dutch must contain, what materials must preferentially be developed and what costs are associated with this. Priority lists will be drawn up which can serve as a policy guideline.

## **3.3 Key Action C: working out standards and evaluation criteria**

It is not sufficient to observe that a certain basic provision is present: the basic provision must also meet requirements of formal and content quality, availability (free of rights or under certain conditions), multifunctionality and re-use (not only useable for the application for which it was originally developed) and it should be documented. For the evaluation in principle one should be able to use the results of different European and American projects and organisations (EAGLES, TSNLP, ELSE, ARPA/DARPA etc.). Key Action C aims at providing a modular set of evaluation instruments which can be used to assess project applications and evaluate project results. A list will be drawn up of criteria that basic materials must satisfy. Furthermore standards should be drawn up for the validation of Dutch language materials.

Key Actions B and C will be carried out in an integrated way. A Flemish-Dutch steering committee, consisting of experts from different disciplines in HLT, will draw up lists of provisions (resources and tools) that should be developed by priority with different essential applications in mind. At the back of these Key Actions is the question: what is essentially needed to ensure that citizens with Dutch as their mother tongue, can use their own language in the different communicative situations where ICT is involved– e.g. multilingual Internet browsers, GPS systems in cars, domotica.

On the basis of a qualitative evaluation it should be evident whether the provisions that already exist are qualitatively satisfactory and are available. This gives a clearer view of what can be included in the existing material infrastructure. On the basis of the results it may be evident which materials are suitable, unsuitable (for example not multifunctional or not available) or are only suitable after adaptation as part of the infrastructure. In this way a realistic summary of the present state of affairs is created and an impression is created of the knowledge and time investment for the modular adaptation and subsequently making available of existing materials.

Comparing the actual existing basic materials with the minimal infrastructure that is required for a language should give input for the above mentioned lists of provisions that should be developed by priority. For drawing up these lists a dialogue is required between industry (what is needed?), knowledge infrastructure (what is needed? what is possible?) and policy (how to facilitate this).

## **3.4 Key Action D: management, maintenance and distribution plan**

Bodies often only finance the development of materials and no longer feel responsible once the project has been finished. However, materials that are not maintained quickly lose value. Unclear intellectual property right arrangements can create difficulties for exploitation. The purpose of Key Action D is the drawing up of a blueprint for the management, availability, and distribution of basic materials developed with government money. The management includes among other things reaching agreements with owners of the material on intellectual property rights and distribution, the possible adaptation of the material so that it can be used on different computer systems, making available documentation, the help desk function, maintaining and updating the material. Much attention will be given to legal issues. Models of different contracts and licence agreements can give guidance in different situations. In the blueprint attention must then also be given to ways in which organisations/persons who manage materials can collaborate structurally.

The blueprint may offer points of connection for various organisations confronted with the same problems. When new projects are being set up, the recommendations which consider at an early stage matters such as intellectual property, financing maintenance, distribution possibilities can be taken into account. In drawing up the blueprint there will be collaboration with ELRA.

### **4** Conclusion

In the Dutch speaking area a growing awareness of the importance of HLT as a key to participation in the information society can be observed. We feel that solid foundations have been led for a structural cross-border co-operation between de various actors involved. This cooperation combined with the creation of the right preconditions, enhances the opportunities for HLTprojects and contributes to strengthen the position of Dutch in the multilingual information society.

#### **5.** Acknowledgements

We like to thank drs. A. Dijkstra of the Netherlands Organisation for Scientific Research (NWO) for her comments on an earlier version of this paper.

#### 6. References

- Bouma, G. and I. Schuurman, 1998a. Intergovernmental Language Policy for Dutch and the Language and Speech Technology Infrastructure. In Rubio, A., N. Gallardo, R. Castro and A. Tejada, Eds., *Proceedings First International Conference on Language Resources* & *Evaluation*, Granada, Spain 28-30 May 1998. 509-513.
- Bouma, G. and I. Schuurman, 1998b. *De positie van het Nederlands in Taal- en Spraaktechnologie*. Report for the Dutch Language Union.
- Erwin Dewallef, 1998. Language Engineering in Flanders, Project EUROMAP-Flanders.
- Ministerie van economische zaken, 1997, Data- en Kennissystemen, Multimedia en interactie. In *TechnologieRadar 3 – Technologie Profielen*.

Nederlandse Taalunie (1999). Action plan for Dutch in language and speech technology.

http://www.taalunie.org/\_/werkt/technologie.html

Van Staden, M., 1997. *Charter the future of Language engineering*, Project EUROMAP-Netherlands.