Strengthening the Dutch Human Language Technology Infrastructure

Catia Cucchiarini[1,3]*, Walter Daelemans*[2] *and Helmer Strik*[3]

[1] Nederlandse Taalunie, The Hague, The Netherlands
[2] CNTS Language Technology Group, University of Antwerp, Belgium
[3] A^{2}RT, Department of Language and Speech, University of Nijmegen, The Netherlands

The ELRA Newsletter, Vol 6 (4), 3-7, 2001.

1. Introduction

The growing importance of information and communication technology (ICT) in our society has emphasized the need for Human Language Technologies (HLT), since these make it possible for people to use natural language in their communication with computers. Preferably, this language should be the user's mother tongue, since this is the only way to guarantee that all citizens can fully participate in the information society. In order to develop HLT applications that allow people to use their native language in their interactions with computers, a digital language infrastructure is required for each language. By digital language infrastructure we mean all basic software tools, language and speech data, corpora and lexicons that are necessary for conducting research and developing applications in the field of HLT. Since the costs of developing HLT resources are high, it is important that all parties involved, both in industry and academia, co-operate so as to maximise the outcome of efforts in the field of HLT. This particularly applies to languages that are commercially less interesting than English, such as Dutch.

The last few years have witnessed a growing awareness of the importance of such a digital language infrastructure, not only in the United States and in Asia, but also in Europe. This is evident from the various initiatives that have been taken at European level, such as the creation of ELRA, the organization of the LREC conferences, and the various projects funded by the European Commission, e.g. SPEECHDAT, PAROLE, SIMPLE, CLASS, EAGLES, HOPE, ISLE, to name but a few. Moreover, several projects have recently been launched by the National Authorities (Ministries or their Departments) in various European countries with the specific aim of strengthening the digital language infrastructure. Projects of this kind require that a dialogue be established between the parties involved: industry, academia and policy institutions. To establish such a dialogue is not always easy, often because the various parties have conflicting interests. Discrepancies may exist not only between industry and universities, but also between the various research groups within industry and academia. From the contacts we have had with our European colleagues, it appears that it is just these kinds of problems that have hampered the emergence and the organization of other countries' national projects aimed at providing or improving HLT resources for their respective languages

In this paper we report on one such initiative that was taken for the Dutch language by the *Dutch Language Union (Nederlandse Taalunie* – abbreviated ***NTU***): the Dutch Human Language Technologies Platform. We hope that the experiences we have had in the last two years in setting up these activities may be useful to others who are now beginning with this kind of work.

2. The Dutch Language Union (NTU) and Human Language Technologies

The plan to set up a Dutch HLT platform was launched by the NTU. This is an intergovernmental organisation established in 1980 on the basis of the Language Union Treaty between Belgium and the Netherlands, which has the mission of dealing with all issues related to strengthening the position of the Dutch language (see also www.taalunie.org). 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 of Education and Culture, is responsible for the policy of the NTU. In establishing its current long-term policy plan (1998 – 2002), the NTU has given full consideration to the rapid developments in the field of ICT that are going to have a major impact on language issues. The governments of the Netherlands and Flanders appreciate the growing importance of 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 at the national and European level, the knowledge infrastructure and the business community. Co-operation among all these actors is of utmost importance, and given the size of the Dutch language area, this co-operation needs to be expanded to a cross-border Flemish-Dutch level. Building on this awareness, 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 (see also Oostdijk, 2000). The project is funded jointly by the Dutch and Flemish governments. Project activities are coordinated 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 machine translation modules for the language pairs Dutch - English/French and English/French- Dutch (see also Cucchiarini, 2001; Goetschalckx, Cucchiarini, and Van Hoorde, 2001) . 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, the Flemish Institute for the Promotion of Scientific and Technological Research in Industry, and Systran, which is the technology provider. 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.

In the project preparation of the *Spoken Dutch Corpus* as well as of *NL-Translex* much time was spent in finding 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 few years. The market research carried out in the Netherlands and in Flanders in the framework of EUROMAP and the research commissioned by the NTU into the position of Dutch in Language and Speech Technology (report Bouma and 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 co-operation in the field of HLT, the NTU took the initiative to set up a Dutch-Flemish platform to support the Dutch language in HLT.

3. The Dutch Human Language Technologies Platform

The main purpose of the Dutch HLT Platform is to further development of an adequate digital language infrastructure for Dutch so that the applications can be developed which can guarantee that the citizens in Holland 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 strengthen the position of the Dutch language in HLT developments, so that the speakers of Dutch can fully participate in the information society;
- To establish the proper conditions for a successful management and maintenance of basic HLT resources developed through governmental funding;
- To stimulate co-operation between academia and industry in the field of HLT;
- To contribute to the realisation of European co-operation in HLT-relevant areas;
- To establish a network that brings together demand and supply of knowledge, products and services.

In addition to the NTU, the following Flemish and Dutch partners are involved in the HLT Platform:

- the Ministry of the Flemish Community,
- the Flemish Institute for the Promotion of Scientific-technological Research in Industry
- the Fund for Scientific Research – Flanders
- the Dutch Ministry of Education, Culture and Sciences,
- the Dutch Ministry of Economic Affairs,
- the Netherlands Organisation for Scientific Research (NWO)
- Senter (an agency of the Dutch Ministry of Economic Affairs)

All these organisations have their own aims and responsibilities and approach HLT accordingly. Together they provide a good coverage of the various perspectives from which HLT policy can be approached.

The rationale behind the Dutch HLT platform was not to create a new structure, but rather to co-ordinate the activities of existing structures. The platform is a flexible framework within which the various partners adjust their respective HLT agendas to each other's and decide whether to place new subjects on a common agenda. Initially, the Dutch HLT platform was set up for a period of five years (1999-2004).

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. This absolutely requires that efforts be cumulative and not duplicated, that insight be provided into the resources that are needed for a language in general and for Dutch in particular and that a plan be drawn up for the development of the resources that are totally lacking or insufficiently available for Dutch. Furthermore, attention should be paid to such matters as evaluation of resources and project results, standardisation, maintenance, distribution etc. In other words, it is necessary to create the preconditions to maximise the outcome of efforts in the field of HLT. To this end, an *Action plan for Dutch in language and speech technology* has been defined, which is funded jointly by the different partners in the HLT platform. The activities described in this action plan are organized in four action lines:

Action line A: performing a 'market place' function
The main goals of this action line are to encourage co-operation between the parties involved (industry, academia and policy institutions), to raise awareness and give publicity to the results of HLT research so as to stimulate market takeup of these results.

Action line B: strengthening the digital language infrastructure
The aims of action line B are to define what the so-called BLARK (Basic LAnguage Resources Kit) for Dutch should contain and to carry out a survey to determine what is needed to complete this BLARK and what costs are associated with the development of the material needed. These efforts should result in a priority list with cost estimates which can serve as a policy guideline.

Action line C: working out standards and evaluation criteria
This action line is aimed at drawing up a set of standards and criteria for the evaluation of the basic materials contained in the BLARK and for the assessment of project results.

Action line D: developing a management, maintenance and distribution plan
The purpose of this action line is to define a blueprint for management (including intellectual property rights), maintenance, and distribution of HLT resources.

In this paper we will focus on action lines B and C.

4. Action lines B and C: survey, evaluation and directions for future development

As explained in section 2, the purpose of action line B is to define the BLARK for Dutch and to determine what should be developed on the basis of a detailed analysis of the needs for HLT resources in the short and medium term, in comparison with the BLARK definition and the present situation.

However, it is not sufficient to acknowledge the existence of a given resource, be it a piece of language data or a tool: all HLT resources, to be really useful, have to meet requirements of formal and content quality, availability (free of rights or under certain conditions), multi-functionality and re-usability. It follows that the work to be carried out for action line B is inextricably linked to the activities in action line C. Only on the basis of a qualitative evaluation is it possible to establish whether the resources that already exist are available and qualitatively satisfactory. This gives a clearer view of what can be included in the HLT infrastructure. The results of such an analysis will reveal which materials are suitable, unsuitable (for example not multifunctional or not available) or are only suitable after adaptation. This will provide a realistic view on the present state of affairs with respect to HLT resources. For the reasons mentioned above, it was soon decided that action lines B and C would be carried out in an integrated way.

In the following sections we provide more detailed information on action lines B and C. First we describe the structure that was set up to conduct the work planned in these two action lines. We then describe the tasks of the various participants. Subsequently, we present the instruments that were developed to carry out these activities and, finally, we present the results obtained so far.

4.1. Structure

4.1.1. Steering committee

The first step in organizing the activities for action lines B and C was to set up a Flemish-Dutch steering committee. This committee is composed of experts from different disciplines in HLT and of representatives of language and research policy institutions such as NTU and NWO. The experts have been selected on the basis of their nationality and their expertise. More precisely, there are four experts from the Netherlands and four experts from Flanders. For each geographical area there are two experts on language technology and two experts on speech technology. This composition guarantees that all parties involved have a representative that will protect their interests and that will provide reliable information on the topics at issue.
The steering committee has the followings tasks:

1. to draw up a plan of the activities that should be carried out to achieve the goals of action lines B and C;
2. to develop an initial framework that will be used for surveying the current state of Dutch HLT resources;
3. to select and hire field researchers who will carry out the actual field survey (see following section);
4. to supervise the field survey of Dutch HLT resources;
5. to establish a set of standards and evaluation criteria for HLT resources;
6. to define the so-called BLARK (Basic LAnguage Resources Kit) for Dutch;
7. to draw up a list of what is needed to complete the BLARK and what costs are associated with the development of the material needed.

The framework to be used in the field survey will be presented in the following section.

4.1.2. Field researchers

Four field researchers have been appointed by the steering committee, two for language technology and two for speech technology. These researchers have the following tasks:

1. to further refine the framework that will be used for surveying the current state of Dutch HLT resources
2. to develop specific instruments for the field survey (tables and questionnaires)
3. to collect information on HLT evaluation instruments
4. to conduct the field survey
5. to write a report

4.2. Survey framework

In order to carry out a thorough survey of the current state of Dutch HLT resources adequate instruments are needed which guarantee, as much as possible, that the survey is complete, unbiased and uniform. The HLT experts in the steering committee worked out an initial framework that was further refined by the field researchers. In setting up this framework the experts have used the usual three components: 1. applications, 2. modules and 3. data.

4.2.1. Applications

In this framework, the term application refers to a class of applications rather than to a specific application or product. This is done to obtain a framework that is general enough to capture all sorts of possible applications. The applications distinguished are:

- *CALL* (Computer Assisted Language Learning)
- *Access control*

Applications in which physical characteristics such as speech signals are used for speaker verification or identification to provide access to systems, buildings etc.

- *Speech input*

Applications in which speech input is analysed and converted into text. This category also includes applications such as command and control, dictation, and automatic transcription.

- *Speech output*

Applications in which text is converted into speech, such as spoken e-mail, spoken dictionaries and aids for the blind.

- *Language and speech interfaces*

Spoken dialogue systems that constitute a natural interface to databases, expert systems, information systems and virtual reality applications in which speech interaction plays a part.

- *Document production*

All applications concerning text production, from spelling, grammar and style checking up to text generation.

- *Information access*

Applications in which text and speech analysis play a part in information localization and knowledge extraction, information retrieval, text mining, document routing, filtering and classification, question answering etc.

- *Machine translation*

Translation aids, translation memories, machine translation.

4.2.2. Modules (semi-products)

Under modules, or semi-products, we understand the basic software components of HLT applications. In general, these components do not have much commercial value as such, but they are essential in the HLT infrastructure. The list of modules identified so far is given below:

- Rule-based synthesis
- Diphone synthesis
- Unit selection
- Sentence boundary detection
- Grapheme-phoneme conversion
- Complete speech synthesis
- Complete speech recognition
- Token detection
- Lemmatizing
- Morphological analysis
- Morphological synthesis
- Part of speech tagging
- Constituent recognition
- Shallow Parsing

- Named entity recognition
- Parsers and grammars
- Prosody prediction
- Referent resolution
- Word meaning disambiguation
- Semantic analysis
- Pragmatic analysis
- Text generation
- Language-pair dependent translation modules.

4.2.3. Data

In this case the term data refers to sets of language data and descriptions in machine readable form, to be used in building, improving or evaluating natural language and speech processing systems. Examples of data are written and spoken corpora, lexical databases and terminology lists. In our scheme the following data types have been distinguished:

- *Monolingual lexicons*.

Lexicons containing orthographic, phonetic, phonological, morphological, syntactic, semantic and pragmatic knowledge about lexical entities (morphemes, word forms, collocation and special expressions).

- *Multilingual lexicons*.

Monolingual lexicons with translations of the lexical entities.

- *Thesauri*.

Lexicons with semantic and associative relations among words.

- *Annotated text corpora*.

Large (10M+) text databases with annotation tiers for orthography, phonology, morphology, syntax, semantics and pragmatics. These data are especially important for training the various modules.

- *Non-annotated text corpora*.

Large (100M+) text databases without annotation tiers, which only contain information the origin of the texts and, possibly, the typographic structure. These corpora are used for unsupervised training.

- *Speech corpora*.

Large (10M+) databases with, at least, orthographically annotated speech.

- *Multingual corpora*.

Databases that contain speech from Dutch and other languages.

- *Multimodal corpora*.

Databases that contain speech and data from other modalities

- *Multimedia speech corpora*.

Databases that contain speech from radio and TV but also information from other media (e.g. texts and figures from WWW, papers and journals)

4.2.4. Matrices

On the basis of the relationships between the three components mentioned above, applications, modules and data, three matrices were designed that address three different topics in the HLT infrastructure:

1. *Relevance of modules for applications*

This matrix shows which modules are required for the various applications.

2. *Relevance of data for modules*

This matrix shows which data are required for the various modules.

3. *Availability of data and modules*

This matrix indicates which data and modules are really available in the sense that they have an acceptable quality level.

4.3. Survey results

4.3.1. BLARK

On the basis of matrices 1 and 2 a BLARK for language technology and one for speech technology can de derived. These are shown below:

BLARK for language technology

- **Modules**
  - Robust modular text preprocessing
  - Morphological analysis and morphosyntactic disambiguation / unknown words

    - Robust syntactic analysis
    - Aspects of semantic analysis (word meaning and reference)
- **Data**
    - Monolingual lexicon
    - Annotated corpus of written Dutch
    - Benchmarks for evaluation

BLARK for speech technology

- **Modules**
    - Automatic speech recognition (module)
    - Speech synthesis system (module)
    - Tools for annotation of speech corpora
    - Confidence measures and utterance verification
    - Identification (speaker, language, dialect)
    - Evaluation of speech technology tools and applications
- **Data**
    - Monolingual speech corpora for specific applications
    - Multilingual speech corpora
    - Multimodal/medial speech corpora
    - Richly annotated speech corpora
    - Pronunciation lexicons

4.3.2. List of priorities

By analyzing the availability of modules and data, priority can be assigned to the development of those parts of the BLARK that are known to be crucial and appear to be missing. The general idea is that those components and data that are relevant for many applications and turn out to be unavailable or of low quality should be developed first. On the basis of matrix 3 such a list of priority was drawn up and was subsequently submitted to representatives from the whole HLT field. The results of this consultation are to be discussed in public during a supranational seminar to be held in The Hague on 15 November (for further information, the reader is referred to http://www.taalunieversum.org/tst/).

5. Concluding remarks

In this paper we have reported on the activities that were carried out in the two years that the Dutch HLT platform has been active. It should be noted that much effort was spent in setting up the whole platform structure, i.e. in finding the representatives of the appropriate responsible bodies and expertise centres. Owing to the fragmentation of responsibilities, it was difficult in the past to conduct a coherent HLT policy. We hope that the HLT platform will contribute to creating more transparency in this respect.

Up to now our experiences have been positive across the board. It turned out that experts from different disciplines and different countries managed to work together and could reach an agreement on a number of important matters. We can only hope that this trend will continue, since there is still much work to be done.

6. References

Bouma, G. and Schuurman, I. (1998) *De positie van het Nederlands in Taal- en Spraaktechnologie.* Report for the Dutch Language Union.

Cucchiarini, C. (2001) The Dutch Connection. A European Machine Translation project for the Dutch Language, *Language International*, Vol. 13 No. 4, 43-46.

Goetschalckx, J. Cucchiarini, C. and Van Hoorde, J. (2001) Machine Translation for Dutch: the NL-Translex Project Why Machine Translation?, Proceedings of the International Colloquium "*Trends in Special Language & Language Technology*", R. Temmerman & M. Lutjeharms (eds.), 261-280.

Oostdijk, N. (2000) The Spoken Dutch Corpus. Overview and first Evaluation, *Proceedings LREC 2000*, Athens, Greece.

7. Acknowledgements

We are indebted to the other members of the steering committee and to the field researchers for their invaluable contribution to the work presented in this paper: J. Beeken, D, Binnenpoorte, G. Bouma, E. Dewallef, E. D'Halleweyn, F. De Vriend, A. Dijkstra, A. Dirksen, D. Heylen, M. Kempen, F. de Jong, J-P. Martens, A. Nijholt, J. Sturm, D. Van Compernolle, F. Van Eynde, R. Veldhuis.