Strengthening the Dutch Language and Speech Technology Infrastructure

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

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

c.cucchiarini@let.kun.nl, daelem@uia.ua.ac.be, strik@let.kun.nl

Abstract

In this paper we report on a project that was launched by the Dutch Language Union (Nederlandse Taalunie) with the aim of strengthening the position of Dutch in language and speech technology (Human Language Technologies, HLT). In particular we report on the activities aimed at surveying and evaluating HLT resources to establish priorities for future developments.

1. Introduction

The term "language resources" was originally used to refer to sets of written and spoken language data that constitute the basis for developing all sorts of language processing systems. More recently, the use of this term has been extended to include basic software tools that, together with language data, form the digital language infrastructure, necessary for conducting research and developing applications in the field of language and speech technology.

The availability of such a digital language and speech infrastructure is a pre-requisite for the participation of a language, and of the citizens speaking this language, in what has come to be known as the information society. In this society, information and communication technologies (ICT) play a vital role in guaranteeing competitiveness in all branches of industry, trade and service provision. Human language technologies (HLT) are an essential part of many ICT applications. Thanks to HLT it is possible for users to address computers in natural language. 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. Every language wishing to occupy a full position in the information society has to keep up with the developments in HLT. This requires that the digital language infrastructure needed for the development of applications be present and of satisfactory quality.

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, MATE, DISC, 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 has been taken for the Dutch language: the Dutch Human Language Technologies platform. More specifically, we will report on the activities that have been carried out within two of the action lines in the platform work plan: those aimed at surveying and evaluating language resources to establish priorities for future developments. We hope that the experiences we 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 Human Language Technologies Platform

The Dutch HLT Platform cannot actually be characterised as a national initiative, but rather as a supranational one, because it goes beyond national borders and concerns the Netherlands and the Flemish part of Belgium. The plan to set up a Dutch HLT platform was launched by the *Dutch Language Union (Nederlandse Taalunie* – abbreviated ***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 (for further details on the NTU, the reader is referred to [1]).

The main purpose of the Dutch HLT Platform is to contribute to the further development of an adequate language and speech technology infrastructure for Dutch. 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: construction of 'broking and linking' 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: plan to strengthen 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: 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.

3. 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.

3.1. Structure

3.1.1. 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. *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 first three tasks have already been carried out, while 4 is now well under way. The framework to be used in the field survey will be presented in the following section.

3.1.2. 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

and, possibly, to carry out evaluation tests.

3.2. Survey instruments

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. Up to now the HLT experts in the steering committee have worked out an initial framework that will be further refined in the coming months by the field researchers. In setting up this framework the experts have analyzed the three usual components in the HLT infrastructure for Dutch:

1. Applications
2. Modules (semi-products)
3. Data

Each of these three components will be described in detail in one the following subsections. Applications, modules and data are then combined into three different matrices, described in 3.2.4, which constitute the initial survey instruments.

By analyzing the importance of modules and data for applications, a BLARK can be proposed for Dutch HLT. Subsequently, 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.

3.2.1. 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 distinguished applications are:

- *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, pronunciation 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.

3.2.2. Modules

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. A provisional list of the modules identified so far is given below. This list will be adjusted, if necessary, by the field researchers on the basis of their findings.

- 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.

3.2.3. 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:

1. *Monolingual lexicons*.

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

2. *Multilingual lexicons*.

Monolingual lexicons with translations of the lexical entities.

3. *Thesauri*.

Lexicons with semantic and associative relations among words.

4. *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.

5. *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.

6. *Speech corpora*.

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

3.2.4. Matrices

On the basis of the relationships between the three components mentioned above, applications, modules and data, three matrices have been 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. This matrix can be properly filled in only after evaluation has been carried out.

Together, matrices 1 and 2, with the necessary adjustments, will form the BLARK, while matrix 3 will show to what extent the BLARK is now available and what should be developed to complete it.

3.3. Evaluation instruments

As explained above, it was soon decided that survey and evaluation be carried out simultaneously because the actual availability of a product is not determined merely by its existence, but depends heavily on the quality of the product itself. It follows that to complete matrix 3 a thorough analysis of all data and modules should be carried out.

However, the complexity of such an enterprise should not be underestimated. Although it is possible to refer to the work carried out by various European and American projects and organisations (EAGLES, TSNLP, ELSE, ARPA/DARPA etc.) it is clear that for many modules and applications there are no standard evaluation instruments.

In this respect a distinction can be drawn between validation, which is usually applied to language and speech data, and evaluation, which is more complex and is applied to software modules or complete applications. As explained in [2], validation can refer to a variety of actions, but, in most cases it refers to a procedure in which the quality of a particular piece of data is checked against a set of requirements. Evaluation can also be of different types. In this project we are interested in what is known as performance evaluation [3], which is aimed at measuring the performance of a particular software module or application on the basis of a criterion, a measure and a method. The major difficulty of this type of evaluation lies in finding the adequate criterion, measure and method for each component to be evaluated, see also [4].

4. 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.

5. References

[1] Beeken, J., Dewaellef, E., and D' Halleweyn E. "A Platform for Dutch in Human Language Technologies", *Proceedings LREC 2000,* Athens, Greece, 63-66, 2000.

[2] Van den Heuvel, H., Boves, L., Choukri, K., Goddijn, S.M.A. and Sanders, E. "SLR Validation: Present State of Affairs and Prospects". *Proceedings LREC 2000,* Athens, Greece, Vol. I, 435-440, 2000.

[3] Hirschman, L., and Thompson, H.S. “Overview of evaluation in speech and natural language processing”, In: R. Cole et al. (eds.) “*Survey of the State of the Art in Human Language Technology*”, Cambridge University Press, 1997.

[4] Strik, H., Cucchiarini, C. and Kessens, J. "Comparing the recognition performance of CSRs: In search of an adequate metric and statistical significance test", *Proceedings of ICSLP' 00*, Beijing, 740-744, 2000.

6. Acknowledgements

We are indebted to the other members of the steering committee for their invaluable contribution to the work presented in this paper: J. Beeken, G. Bouma, E. Dewaellef, E. D' Halleweyn,A. Dijkstra, D. Heylen, F. de Jong, P. Martens, A. Nijholt, D. Van Compernolle, F. Van Eynde, R. Veldhuis.