DUTCH HLT RESOURCES: FROM BLARK TO PRIORITY LISTS

H. Strik[1], W. Daelemans[2], D. Binnenpoorte[3], J. Sturm[1], F. De Vriend[1], C. Cucchiarini[1,4]

[1] Department of Language and Speech, University of Nijmegen, The Netherlands
{Strik, D.Binnenpoorte, Janienke.Sturm, F.deVriend, C.Cucchiarini}@let.kun.nl
[2] Department of CNTS Language Technology, University of Antwerp, Belgium
Walter.Daelemans@uia.ua.ac.be
[3] Speech Processing Expertise Centre (SPEX), Nijmegen, The Netherlands
[4] Nederlandse Taalunie, The Hague, The Netherlands

ABSTRACT

In this paper we report on a project about Dutch Human Language Technologies (HLT) resources. In this project we first defined a so-called BLARK (Basic LAnguage Resources Kit). Subsequently, a survey was carried out to make an inventory and evaluation of existing Dutch HLT resources. Based on the information collected in the survey, a priority list was drawn up of materials that need to be developed to complete the Dutch BLARK. Although the current project only concerns the Dutch language, the method employed and some of the results are also relevant for other languages.

1. INTRODUCTION

With information and communication technology becoming more and more important, the need for HLT also increases. HLT enable people to use natural language in their communication with computers, and for many reasons it is desirable that this natural language be the user's mother tongue. In order for people to use their native language in these applications, a set of basic provisions (such as tools, corpora, and lexicons) is required. However, 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.

For this reason, the Dutch Language Union (Nederlandse Taalunie – abbreviated NTU), which is a Dutch/Flemish intergovernmental organisation responsible for strengthening the position of the Dutch language (for further details on the NTU, see [1]), launched an initiative, the Dutch HLT Platform. This platform aims at stimulating co-operation between industry and scientific institutes and at providing an infrastructure that will make it possible to develop, maintain and distribute HLT resources for Dutch.

The work to be carried out in this project was organised along four action lines, which are described in more detail in [2]. In the present paper, action lines B and C are further outlined. Action line A is about constructing a 'broking and linking' function, and the goal of action line D is to define a blueprint for management, maintenance and distribution.

The aims of action line B are to define a set of basic HLT resources for Dutch that should be available for both academia and industry, the so-called BLARK (Basic LAnguage Resources Kit), 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 materials needed. These efforts should result in a priority list with cost estimates, which can serve as a policy guideline. Action line C 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. Obviously, the work done in action lines B and C is closely related, for determining whether materials are available cannot be done without a quality evaluation. For this reason, action lines B and C have been carried out in an integrated way.

The work in action lines B and C was carried out in three stages, which are described in more detail below:

1. defining the BLARK,
2. carrying out a field survey to make an inventory and evaluation of existing HLT resources, and
3. defining the priority list.

The project was co-ordinated by a steering committee consisting of Dutch and Flemish HLT experts.

2. DEFINING THE BLARK

The first step towards defining the BLARK was to reach consensus on the components and the instruments to be distinguished in the survey. A distinction was made between applications, modules, and data (see Table 1). 'Applications' refers to classes of applications that make use of HLT. 'Modules' are the basic software components that are essential for developing HLT applications, while 'data' refers to data sets and electronic descriptions that are used to build, improve, or evaluate modules.

In order to guarantee that the survey is complete, unbiased and uniform, a matrix was drawn up by the steering committee describing (1) which modules are required for which applications, (2) which data are required for which modules, and (3) what the relative importance is of the modules and data. This matrix (subdivided in language and speech technology) is depicted in Table 1, where "+" means important and "++" means very important.

This matrix serves as the basis for defining the BLARK. Table 1 shows for instance that monolingual lexicons and annotated corpora are required for the development of a wide range of modules; these should therefore be included in the BLARK. Furthermore, semantic analysis, syntactic analysis, and text pre-processing (for language technology) and speech

recognition, speech synthesis, and prosody prediction (for speech technology) serve a large number of applications and should therefore be part of the BLARK, as well.
Based on the data in the matrix and the additional prerequisite that the technology with which to construct the modules be available, a BLARK is proposed consisting of the following components:

For language technology:

Modules:

- Robust modular text pre-processing (tokenisation and named entity recognition)
- Morphological analysis and morpho-syntactic disambiguation
- Syntactic analysis
- Semantic analysis

Data:

- Mono-lingual lexicon
- Annotated corpus of text (a treebank with syntactic, morphological, and semantic structures)
- Benchmarks for evaluation

For speech technology:

Modules:

- Automatic speech recognition (including tools for robust speech recognition, recognition of non-natives, adaptation, and prosody recognition)
- Speech synthesis (including tools for unit selection)
- Tools for calculating confidence measures
- Tools for identification (speaker identification as well as language and dialect identification)
- Tools for (semi-) automatic annotation of speech corpora

Data:

- Speech corpora for specific applications, such as Computer Assisted Language Learning (CALL), directory assistance, etc.
- Multi-modal speech corpora
- Multi-media speech corpora
- Multi-lingual speech corpora
- Benchmarks for evaluation

3. SURVEY: INVENTORY & EVALUATION

In the second stage, a survey was carried out to establish which of the components that make up the BLARK are already available; i.e. which modules and data can be bought or are freely obtainable for example through open source. Besides being available, the components should also be (re-)usable. Note that only language specific modules and data were considered in this survey.

Obviously, components can only be considered usable if they are of sufficient quality. Therefore, a formal evaluation of the quality of all modules and data is indispensable. Evaluation of the components can be carried out on two levels: a descriptive level and a content level. Evaluation on a content level would comprise validation of data and performance validation of modules, whereas evaluation on a descriptive level would mean checking the modules and data against a list of evaluation criteria. Since there was only a limited amount of time, it was decided that only the checklist approach would be feasible. A checklist was drawn up consisting of the following items:

- Availability:
  - public domain, freeware, shareware, etc.
  - legal aspects, IPR
- Programming code:
  - language: Fortran, Pascal, C, C++, etc.
  - makefile
  - stand-alone or part of a larger module?
- Platform: Unix, Linux, Windows 95/98/NT, etc.
- Documentation
- Compatibility with standards: (S)API, SABLE
- Compatibility with standard packages: MATLAB, Praat, etc.
- Reusability / adaptability / extendibility:
  - to other tasks and applications
  - to other platforms
- Standards

As a first step in the inventory, the experts in the steering committee made an overview of the availability of components. Then the steering committee appointed four field researchers to carry out the survey. The field researchers then extended and completed this overview on the basis of information found on the internet and in the literature, and personal communication with experts.

4. PRIORITY LISTS

The survey of Dutch and Flemish HLT resources resulted in an extensive overview of the present state of HLT for the Dutch language. We then combined the BLARK with the inventory of components that were available and of sufficient quality, and drew up priority lists of the components that need to be developed to complete the BLARK. The prioritisation proposed was based on the following requirements:

- the components should be relevant (either directly or indirectly) for a large number of applications,
- the components should currently be either unavailable, inaccessible, or have insufficient quality, and
- developing the components should be feasible in the short term.

At this point, we incorporated all information gathered in a report containing the BLARK, the availability figures together with a detailed inventory of available HLT resources for Dutch, priority lists of components that need to be developed, and a number of recommendations [3]. This report was given a provisional status, as feedback on this version from a lot of actors in the field was considered desirable, since reaching consensus on the analysis and recommendations for the Dutch and Flemish HLT field is one of the main objectives.

Therefore, we consulted the whole HLT field. Using the address list compiled in Action Line A of the Platform, a first version of the priority lists, the recommendations, and a link to a pre-final version of the inventory [3] were sent to all known actors in the Dutch HLT field: a total of about 2000 researchers, commercial developers and users of commercial systems. We asked all actors to comment on the report, the priority lists, and the recommendations. Relevant comments were incorporated in the report.

Simultaneously, the same group of people was invited to a workshop that was organised to discuss the BLARK, the priority list and the recommendations. Some of the actors that had sent their comments were asked to give a presentation to make their ideas publicly known. The presentations served as an onset for a concluding discussion between the audience and a panel consisting of five experts (all members of the steering committee). A number of conclusions that could be drawn from the workshop are:

- Cooperation between universities, research institutes and companies should be stimulated.
- It should be clear for all components in the BLARK how they can be integrated with off-the-shelf software packages. Furthermore, documentation and information about performance should be readily available.
- Control and maintenance of all modules and data sets in the BLARK should be guaranteed.
- Feedback of users on the components (regarding quality and usefulness of the components) should be processed in a structured way.
- The question as to what open source / license policy should be used needs some further discussion.

On the basis of the feedback received from the Dutch HLT field, some adjustments were made to the first version of the report. The final priority lists are as follows:

For language technology:

1. Annotated corpus written Dutch: a treebank with syntactic and morphological structures
2. Syntactic analysis: robust recognition of sentence structure in texts
3. Robust text pre-processing: tokenisation and named entity recognition
4. Semantic annotations for the treebank mentioned above
5. Translation equivalents
6. Benchmarks for evaluation

For speech technology:

1. Automatic speech recognition (including modules for non-native speech recognition, robust speech recognition, adaptation, and prosody recognition)
2. Speech corpora for specific applications (e.g. directory assistance, CALL)
3. Multi-media speech corpora (speech corpora that also contain information from other media, i.e. speech together with text, html, figures, movies, etc.).
4. Tools for (semi-) automatic transcription of speech data
5. Speech synthesis (including tools for unit selection)
6. Benchmarks for evaluation

From the inventory and the reactions from the field, it can be concluded that the current HLT infrastructure is scattered, incomplete, and not sufficiently accessible. Often the available modules and applications are poorly documented. Moreover, there is a great need for objective and methodologically sound comparisons and benchmarking of the materials. The components that constitute the BLARK should be available at low cost or for free.

To overcome the problems in the development of HLT resources for Dutch the following can be recommended:

- existing parts of the BLARK should be collected, documented and maintained by some sort of HLT agency,
- the BLARK should be completed by encouraging funding bodies to finance the development of the prioritised resources,
- the BLARK should be available to academia and the HLT industry under the conditions of some sort of open source / open license development,
- benchmarks, test corpora, and a methodology for objective comparison, evaluation, and validation of parts of the BLARK should be developed.

Furthermore, it can be concluded that there is a need for well-trained HLT researchers, as this was one of the issues discussed at the workshop. Finally, enough funding should be assigned to fundamental research.

The results of the survey will be disseminated to the HLT field. The priority lists and the recommendations will be made available to funding bodies and policy institutions by the NTU. A summary of the report, containing the priority lists, the recommendations, and the BLARK will be translated into English to reach a broader public. More information can be found at [4, 5, 6].

5. ACKNOWLEDGEMENT

The following people participated in the steering committee (at various stages of the project): J. Beeken, G. Bouma, C. Cucchiarini, E. D'Halleweyn, W. Daelemans, E. Dewallef, A. Dirksen, A. Dijkstra, D. Heijlen, F. de Jong, J.P. Martens, A. Nijholt, H. Strik, L. Teunissen, D. van Compernolle, F. van Eynde, and R. Veldhuis. The four field researchers were: D. Binnenpoorte, J. Sturm, F. de Vriend, and M. Kempen. We would like to thank all of them, and all others who contributed to the work presented in this paper. Furthermore, we would like to thank an anonymous reviewer for constructive remarks on a previous version of this paper.

6. REFERENCES

[1] Beeken, J., Dewallef, E., D'Halleweyn, E. (2000), A Platform for Dutch in Human Language Technologies. Proceedings of LREC2000, Athens, Greece.

[2] Cucchiarini, C., D'Halleweyn, E. and Teunissen, L. (2002), A Human Language Technologies Platform for the Dutch language: awareness, management, maintenance and distribution. Proceedings LREC2002, Canary Islands, Spain.

[3] Daelemans, W., Strik, H. (Eds.) (2001) Het Nederlands in de taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen (versie 1), 27 sept. 2001. See http://www.taalunieversum.org/tst/actieplan/batavo-v1.pdf or http://lands.let.kun.nl/TSpublic/strik/publications/a82-batavo-v1.pdf

[4] http://www.taalunieversum.org/tst/

[5] http://www.ntu.nl/_/werkt/technologie.html

[6] http://lands.let.kun.nl/TSpublic/strik/taalunie/

Table 1. Overview of the importance of data for modules, and modules for applications.

Modules	Data									Applications							
	monoling lex	multilin lex	thesauri	anno corp	unanno corp	speech corp	multi ling corp	multi mod corp	multi media cor	CALL	access control	speech input	speech output	dialog systems	doc prod	info access	transla-tion
Language Technology																	
Grapheme-phoneme conv.	++			++						+			++	++	+	+	
Token detection	++			+	++					+		+		+	+	+	+
Sent boundary detection	+			++	++					+		++	++	+	++	++	++
Name recognition	+	+	+	++	++	++				+		++	++	+	++	++	++
Spelling correction										+							
Lemmatising	++			++	+					+		+	+	+	+	+	+
Morphological analysis	++			++	+					+		+	++	+	++	++	++
Morphological synthesis	++			++	+					+			++	+	++		++
Word sort disambig.	++			++	+					+		++	+	++	++	++	++
Parsers and grammars	++			++						+		++	++	++	++	++	++
Shallow parsing	++			++	++					+		++	++	++	++	++	++
Constituent recognition	++			++	+					+		++	++	++	++	++	++
Semantic analysis	++		++	++				++	++	+		++	++	++		++	++
Referent resolution	+		++	++	+					+		++		++	++	++	++
Word meaning disambig.	+		++	++	+					+		++	+	+	+	++	++
Pragmatic analysis	+		+	++				++	++	+		++	++	++		+	++
Text generation	++		++	++				++	++	+			++	++	++		++
Lang. dep. translation		++	++	++			++			+						++	++
Speech Technology																	
Complete speech recog.	++	+		++	+	++	+	++	++	++	++	++		++	++	++	++
Acoustic models	++	+		++	+	++	+	+	+	++	+	++		++	+	+	+
Language models	+			++	+	+	+	+	+	++	+	++		++	++	++	++
Pronunciation lexicon	++	+		+		++	+	+	+	++	+	++	+	++	+	++	++
Robust speech recognition	+			+	+	+	+	+	++	+	+	++		++	+	+	+
Non-native speech recog.	+	++		+		++	++	+	+	++	+	+		+		+	+
Speaker adaptation	+			+	+	++	+	+	++	+	+	++		+	+	++	+
Lexicon adaptation	++	+		+		++	+	+	+	++	+	++	+	++	+	++	++
Prosody recognition	+	+		++	+	++	+	+	+	++	+	++		++	++	++	++
Complete speech synth.	++	+		+		+		+		+			++	++	+	+	++
Allophone synthesis	+	+		+		+		+		+			+		+	+	+
Di-phone synthesis	++	+		+		+		+		+			++	++	+	+	+
Unit selection	++	+		+		+		+		+			++	++	+	+	+
Prosody prediction for Text-to-Speech	++	+		+		+		+	+	++			++	++		+	++
Aut. phon. transcription	++	++		+	+	++	+	+	+	++	+	+	+	+	+	+	+
Aut. phon. segmentation	++	++		+	+	++	+	+	+	++	+	+	+	+	+	+	+
Phoneme alignment	+	+		+		++	+	+	+	++	+	+		+			+
Distance calc. phonemes	+	+		+		++	+	+	+	++	+	+		+			+
Speaker identification	+			++	++	++	+	++	+	+	++	+		+		+	+
Speaker verification	+			++	++	++	+	++		+	++	+		+		+	+
Speaker tracking	+			++		++			++	+	++	+		+	+	+	+
Language identification	+	++		+	+	++	++	+	+	+	+	+		+		+	+
Dialect identification	+	++		+	+	++	++	+	+	+	+	+		+		+	+
Confidence measures	+			+	+	++	+	++	+	++	++	++		++	+	+	+
Utterance verification	+			+	+	++	+	+	+	+	+	++		++	+	+	+