A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch

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

[1]Speech Processing Expertise Centre (SPEX), Nijmegen, the Netherlands
[2]Department of Language and Speech, University of Nijmegen
Erasmusplein 1, Nijmegen, The Netherlands
{D.Binnenpoorte, F.deVriend, Janienke.Sturm, H.Strik, C.Cucchiarini}@let.kun.nl
[3]Department of CNTS Language Technology, University of Antwerp, Belgium
Walter.Daelemans@uia.ua.ac.be
[4]Nederlandse Taalunie, Postbus 10595, 2501 HN, The Hague, The Netherlands

Abstract

In this paper we describe a survey of Dutch language resources that has been carried out within the framework of a project launched by the Dutch Language Union (Nederlandse Taalunie) with the aim of strengthening the position of Dutch in Human Language Technologies (HLT). In this paper we present a so-called BLARK (Basic LAnguage Resources Kit). Based on the information collected in the survey, a priority list has been drawn up for materials that need to be developed to complete the BLARK specific for Dutch. The method employed and reported in this paper is not specific for Dutch and can be adopted for other languages.

1. Introduction

With information and communication technology (ICT) becoming more and more important, the need for language and speech technology, often referred to as Human Language Technologies (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, the reader is referred to Beeken et al (2000)), 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 done for the platform is divided into four action lines, which are described in more detail in Cucchiarini & D' Halleweyn (2002). In the present paper, action lines B and C are further outlined. 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 that reason, action lines B and C have been carried out in an integrated way.

The project was co-ordinated by a steering committee consisting of ten people that have expertise in different aspects of the HLT field. The steering committee appointed four field researchers to carry out the survey.

The present paper describes the methods and tools used for conducting the survey. A detailed description is given of the three stages in which the survey was carried out. The components that constitute the BLARK are presented together with the priority list and a number of recommendations that resulted from this survey.

2. Survey

The field survey can be best described according to the three stages that were passed through. In the first stage the BLARK for Dutch was defined. Then, in the second stage, an inventory was made of HLT resources that are already available. Finally, in the third stage the priority list was drawn up on the basis of the BLARK and the inventory. In the next sections, the three stages will be described in more detail.

2.1. 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:

Applications: refers to classes of applications that make use of HLT. The following classes were defined: CALL (Computer Assisted Language Learning), access control, speech input, speech output, dialogue systems, document production, information access, and multilingual applications or translation modules.

Modules: refers to the basic software components that are essential for developing HLT applications (e.g. grapheme-phoneme conversion, part of speech tagging, automatic speech recognition, speaker verification, text-to-speech, etc.).

Data: refers to data sets and electronic descriptions that are used to build, improve, or evaluate modules. The following data sets are important for HLT: monolingual lexicons, multi-lingual lexicons, thesauri, corpora enriched with several annotations, corpora without annotations, speech corpora with at least an orthographic transcription, multi-lingual corpora, multi-modal corpora, multi-media corpora, and test suites.

In order to guarantee that the survey is complete, unbiased and uniform, a matrix was drawn up by the

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	+			+	+	++	+	+	+	+	+	++		++	+	+	+

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

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. The 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:

- Monolingual 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 CALL, directory assistance, etc.
- Multi-modal speech corpora
- Multi-media speech corpora
- Multi-lingual speech corpora
- Benchmarks for evaluation

2.2. Inventory and evaluation

In the second stage, an inventory was made to establish which of the components - modules and data - that make up the BLARK are already available; i.e. which modules and data can be bought or are freely obtainable for example by 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 but that the BLARK is also relevant for other languages than Dutch.

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: Waves,
- MATLAB, Praat, GIPOS, etc.
- Reusability / adaptability / extendibility:
  - to other tasks and applications
  - to other platforms
  - of modules
  - part of larger module?
- Documentation
- Standards

As a first step in the inventory, the experts in the steering committee made an overview of the availability of components. The field researchers then extended and completed this overview on the basis of information found on the internet and in the literature and by personal communication with actors in the field. Subsequently, the information on availability and the matrix in Table 1 together with a preliminary version of the inventory were submitted to a group of HLT experts from both industry and academia, ensuring that a balanced picture could be obtained.

Based on the reactions of the experts and the earlier collected information a second matrix was filled in which describes the availability of the components in the BLARK (cf. Table 2). Availability in this matrix is expressed in numbers from 1 ('module or data set is unavailable') to 10 ('module or data set is easily obtainable').

At the end of the second stage, all information gathered was incorporated in a report containing the BLARK, the availability figures together with a detailed inventory of available HLT resources for Dutch, a priority list of components that need to be developed, and a number of recommendations. This report was given a provisional status as, feedback on this version from a lot of actors in the field was considered desirable.

Modules	**Availability**
Grapheme-phoneme conversion	8
Token detection	9
Sentence boundary detection	3
Name recognition	4
Spelling correction	3
Lemmatising	9
Morphological analysis	7
Morphological synthesis	9
Word sort disambiguation	7
Parsers and grammars	3
Shallow parsing	2
Constituent recognition	5
Semantic analysis	3
Referent resolution	2
Word meaning disambiguation	2
Pragmatic analysis	1
Text generation	3
Language dependent translation	3
Complete speech recognition	4
Acoustic models	8
Language models	3
Pronunciation lexicon	5
Robust speech recognition	2
Non-native speech recognition	2
Speaker adaptation	2
Lexicon adaptation	2
Prosody recognition	2
Complete speech synthesis	6
Allophone synthesis	7
Di-phone synthesis	6
Unit selection	1
Prosody prediction for Text-to-Speech	3
Autom. phonetic transcription	3
Autom. phonetic segmentation	5
Phoneme alignment	8
Distance calculation of phonemes	8
Speaker identification	2
Speaker verification	2
Speaker tracking	2
Language identification	2
Dialect identification	2
Confidence measures	2
Utterance verification	2
Data	
Unannotated corpora	9
Annotated corpora	5
Speech corpora	4
Multi lingual corpora	3
Multi modal corpora	1
Multi media corpora	1
Test corpora	1
Monolingual lexicons	8
Multilingual lexicons	6
Thesaurus	4

Table 2 Availability of modules and data

2.3. Feedback

Reaching consensus on the analysis and recommendations for the Dutch and Flemish HLT field is one of the main objectives of the survey. Therefore, in the third stage, the whole HLT field was consulted. Using the address list that has been compiled in Action Line A of the Platform, we sent the priority list, the recommendations, and a link to a pre-final version of the inventory to all known actors in the HLT field: a total of about 2000 researchers, commercial developers and users of commercial systems. All actors were asked to comment on the report, the priority list, and the recommendations by email to one of the field researchers. Relevant comments were incorporated in the report.

Simultaneously the same group of people were 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.

From the workshop we got useful advice and many additions to the matrices; these were incorporated in the final version of the report. A number of conclusions that could be drawn from the workshop:

- 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 is the effect of the open source policy on companies and their contribution to the BLARK needs some further discussion.

3. Results: inventory, priority list, and recommendations

The survey of Dutch and Flemish HLT resources resulted in an extensive overview of the present state of HLT for the Dutch language. The overview gives a clear picture of the available modules, data, and applications for the Dutch language and where they can be found.

By combining the BLARK with the inventory of components that are available and of sufficient quality, a priority list can be drawn up for the components that need to be developed to complete the BLARK. The prioritisation proposed is 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.

The following priority lists were drawn up (one for language technology and one for speech technology):

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

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 such as newspapers, WWW, 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 open source 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.

4. Dissemination

The results of the survey have been disseminated to the field through a web page, http://www.taalunieversum.org/tst/beleid/platform.html. The priority list and the recommendations will be made available to funding bodies and policy institutions by the NTU. A summary of the report, containing the priority list, the recommendations, and the BLARK will be translated into English to reach a broader public.

5. Conclusion

This paper describes the method employed to conduct a survey for Dutch HLT resources. First a BLARK, which is more or less language universal, was defined. Subsequently, an inventory was made of available Dutch HLT resources. Finally feedback from experts in the field was gathered to complete the overview. Following this method a report was drawn up with an up-to-date inventory of Dutch HLT, a priority list to complete the BLARK for Dutch and some recommendations. Collecting information to complete the overview of existing Dutch HLT resources was rather time consuming although essential for finally defining a priority list for Dutch HLT.

6. Acknowledgement

We would like to thank the steering committee for their invaluable contribution to the work presented in this paper: J. Beeken, G. Bouma, E. D'Halleweyn, E. Dewallef, A. Dirksen, A. Dijkstra, D. Heijlen, F. de Jong, J.P. Martens, A. Nijholt, D. van Compernolle, F. van Eynde, R. Veldhuis and L. Teunissen.

7. References

Cucchiarini, C., D' Halleweyn, E. (2002), A Human Language Technologies Platform for the Dutch language: awareness, management, maintenance and distribution. Submitted to LREC2002, Canary Islands, Spain.

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