Towards a roadmap for Human Language Technologies: Dutch-Flemish experience

Diana Binnenpoorte[1,2], Catia Cucchiarini[2,3], Elisabeth D'Halleweyn[3], Janienke Sturm[2] and Folkert de Vriend[2]

[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, C.Cucchiarini, F.deVriend, Janienke.Sturm}@let.kun.nl
[3]Nederlandse Taalunie, The Hague, The Netherlands
EdHalleweyn@ntu.nl

Abstract

In this paper we describe how the project "Dutch Human Language Technologies Platform" has contributed to creating the preconditions for establishing a roadmap for Human Language Technologies in the Dutch speaking area. Our overview of the results obtained so far reveals that the goals of all four action lines have been achieved and that there are clear directions for how to proceed in the near future. We hope that our experiences will be useful to other countries that intend to start similar initiatives.

1. Introduction

Establishing a roadmap for Human Language Technologies for a given language requires that first a number of important basic elements be defined, such as:

1. what is minimally required to guarantee an adequate digital language infrastructure for that language?
2. what is the current situation of HLT in that language?
3. what needs to be done to guarantee that at least what is required be available?
4. how can 3 best be achieved ?
5. how can we guarantee that once an adequate HLT infrastructure is available, it also remains so?

It is exactly these questions that were at the core of the activities that in the last two years were carried out within the framework of the Dutch-Flemish project "Dutch Human Language Technologies Platform". The ultimate aim of this project is to further the development and secure the usability of an adequate digital language infrastructure for Dutch, which is required to maximise the outcome of future efforts and to guarantee progress in the field of HLT.

In this paper we will report on our approach and our experiences in carrying out the activities envisaged in this project, because we think that this information can contribute to the aim of this workshop: establishing a roadmap for Human Language Technologies for the next decade.

2. The Dutch HLT Platform: action plan

The plan to set up a Dutch HLT platform was launched by the Dutch Language Union (Nederlandse Taalunie, NTU) which is an intergovernmental organisation established in 1980 on the basis of the Language Union Treaty between Belgium and the Netherlands. The NTU has the mission of dealing with all issues related to strengthening the position of the Dutch language (see also www.taalunie.org). In addition to the NTU, the relevant Flemish and Dutch ministries and organisations are involved in the HLT Platform. The various 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).

To achieve the objectives mentioned above, an *Action plan for Dutch in language and speech technology* was defined, which encompasses various activities organised in four action lines:

2.1. 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 take-up of these results.

2.2. 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.

2.3. 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.

2.4. 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.

Soon after the HLT Platform was set up it was decided that survey (action line B) and evaluation (action line C) be carried out in an integrated way because the actual availability of a product is not determined merely by its existence, but depends heavily on the quality of the product itself.

In the remainder of this paper we analyse the results of each action line in detail and in the final section we consider how this work has paved the way to a roadmap for Dutch HLT.

3. Action line A: results

In setting up HLT projects such as the *Spoken Dutch Corpus* and *NL-Translex,* much time was invested in the search for the appropriate responsible (funding) bodies in the Netherlands and Flanders. Moreover, various studies had indicated that the fragmentation of responsibilities made it difficult to conduct a coherent policy and meant that the field lacked transparency for interested parties. For these reasons the NTU, as the coordinator of the HLT Platform, stimulated the creation of a network aimed at:

disseminating the results of research in the field of HLT;

bringing together demand and supply of knowledge, products and services;

stimulating co-operation between academia and industry in the field of HLT.

After only two years of activity the HLT Platform has already produced important results. The success of Action line A is also partly due to the fact that the NTU acts as the National Focal Point (NFP) in the HOPE (Human Language Technology Opportunity Promotion in Europe) project. HOPE is a multi-country, shared-cost accompanying measure project of the IST-Programme of the European Commission that aims at providing awareness, bridge-building and market-enabling services to boost opportunities for market take-up of 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 Action line A.

At the beginning of the HOPE project an extensive informational website on the HLT sector in The Netherlands and Flanders was established by the NTU. This website provides up-to-date information on all relevant actors in the field of HLT (i.e. researchers, developers, integrators, users and policy makers) on how the HLT sector evolves on a cross-border Dutch/Flemish level, and on HLT related events throughout Europe. All this information is presented in Dutch and English.

The site also includes a calender of HLT events and a form for people who want to be included in the contacts database, as well as links to the HLTCentral website. All information on HLT related programmes and actions of the European Commission is provided on a separate website, established and maintained by subcontractor Senter/EG-Liaison, which is the most knowledgeable party on this subject. These two sites have one entry point from the HOPE point-of-view, via an intermediate site that was developed to provide clarity on where to find which information. This intermediate site (also in Dutch and English) has been placed on http://www.hltcentral.org/euromap/ and should be considered as the common homepage for the two websites. Visitors who do not find answers to their questions on the website can contact the NTU or Senter/EG-Liaison directly (preferably by e-mail) and may expect to receive quick and accurate replies.

Part of the infodesk task is also to conduct mailings to national contacts. These mailings are done on an ad-hoc basis, either at a third party's request (e.g. if an organizing committee wants to announce an event) or on the NFP's own initiative (e.g. if there is important news about an EC programme). From the beginning of the HOPE project, an extensive contacts database has been compiled by the NFP. At present, this database contains almost a thousand persons from over six hundred organisations in The Netherlands and Flanders. It is a valuable backbone for all information activities of the NFP.

The Dutch/Flemish NFP also visits companies with HLT related needs to demonstrate the benefits of HLT, to solicit a clear picture of the company's knowledge state and future plans, and to provide information of cross-linking services where appropriate. The NFP, in collaboration with its partners in The Netherlands and Flanders, has organised various seminars and workshops, which were attended by people from industry, academia, and policy institutions. The aim of such events is to further enhance awareness of recent developments in the HLT sector at the national and international level, such as the dissemination of information on European Commission HLT actions and their relevance to the national situation. Note that the cross-border Flemish/Dutch level should be considered here as the "national" level. The first national seminar took place in March 2001, and was a major event with over 150 participants. The second seminar was held in November 2001 and was directly related to the general survey carried out under action line B and C. Two other events are being organised for 2002. To conclude, we can safely state that in two years time the activities carried out within Action line A have certainly contributed to creating transparency and structure in the HLT field in The Netherlands and Flanders.

4. Results of Action lines B and C

The field survey comprised the following three stages: defining the BLARK for Dutch, making an inventory of available HLT resources, establishing a priority list. These three stages are described in more detail below.

4.1. Defining the BLARK

In defining the BLARK 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.

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 technology 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. Note that only language specific modules and data were considered in this survey.

Based on the data in the matrix the BLARK for Dutch should consist of the following components:

4.1.1. Language technology BLARK

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 written Dutch (a treebank with syntactic, morphological, and semantic structures),
- Benchmarks for evaluation.

4.1.2. Speech technology BLARK

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.

4.2. 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. 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. Given the limited amount of time, only a formal evaluation was carried out by using a checklist with the following items: availability, programming code, platform, documentation compatibility with standard packages, reusability, adaptability and extendibility.

The information on availability, the matrix in Table 1 and the preliminary inventory were submitted to a group of HLT experts from both industry and academia, so that a balanced picture could be obtained.

Based on this information a second matrix was filled in which the availability of the components in the BLARK (cf. Table 2) was described. 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 overview of available HLT resources for Dutch, a priority list of components that need to be developed, and a number of recommendations. This report was considered as being provisional as feedback on this version from a lot of actors in the field was considered desirable.

4.3. Feedback

One of the aims of Action lines B and C was that the majority of the actors in the HLT field would subscribe to the priorities and recommendations for the future. To this end, the provisional report containing the inventory, the priority lists and the recommendations was sent to a total of about 2000 people active in the HLT field who were asked to send their comments by email. After the relevant comments had been incorporated in the report, the same group of people was invited to participate in a workshop in which the results (overview, BLARK, priority lists and recommendations) were officially presented to the public.

Special attention should be paid to the issue of open source policy and its possible effects for companies.

Modules	Data									Applications							
	mono lex	multi lex	thes	ann corp	unann corp	speech corp	multi ling	multi mod	multi media	CALL	access control	speech input	speech output	dialog system	doc prod	info access	transla-tion
Language Technology																	
Grapheme-phoneme conv.	++			++						+			++	++	+	+	
Token detection	++			+	++					+		+		+	+	+	+
Sent boundary detection	+			++	++					+		++	++	+	++	++	++
Name recognition	+	+	+	++	++	++				+		++	++	+	++	++	++
Spelling correction										+							
Lemmatizing	++			++	+					+		+	+	+	+	+	+
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.*

Modules	**Availability**
Grapheme-phoneme conversion	8
Token detection	9
Sentence boundary detection	3
Name recognition	4
Spelling correction	3
Lemmatizing	9
Morphological analysis	
Morphological synthesis	
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

4.4. 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. By combining the BLARK with the inventory of components that are available and of sufficient quality, the following priority for language and speech technology lists were drawn up.

4.4.1. Priority list 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-preprocessing: tokenisation and named entity recognition,
4. Semantic annotations for the treebank mentioned above,
5. Translation equivalents,
6. Benchmarks for evaluation.

4.4.2. Priority list 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 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.

On the basis of the inventory and the reactions from the field the following recommendations were made:

- existing parts of the BLARK should be collected, documented and maintained by a central institution;
- the BLARK should be completed by financing the development of the resources prioritised;
- the BLARK should be made available to industry and academia through open source development;
- benchmarks, test corpora, and methods for evaluation and validation should be developed.
- the training of qualified HLT researchers should be encouraged.

5. Results of Action line D: the HLT Blueprint

In many cases official bodies such as ministries and research organisations are prepared to finance the development of language resources and no longer feel responsible for what should happen to these materials once the project has finished. However, materials that are not maintained quickly lose value. Moreover, unclear intellectual property right arrangements can create difficulties for exploitation. The purpose of action line D was to draw up a blueprint for management, maintenance and distribution of basic language materials that have been developed with government money. This includes, among other things, dealing with intellectual property rights issues, with the acquisition of resources, the adaptation of data and modules to other systems and applications, making documentation available, providing a help desk function, maintaining and updating the material. Finally, this blueprint should provide guidelines for organizing a structural form of co-operation in this respect and should serve as an instrument for field organisations as well as for funding bodies.

The *Blueprint for management, maintenance and distribution of digital materials developed with public money (Blueprint)*, P. van der Kamp, T. Kruyt en P.G.J. van Sterkenburg) was prepared in the period 2000 -2001 by a team of language technology experts of the Institute for Dutch Lexicology, INL. In addition to the general aim of providing guidelines for the acquisition, management, maintenance and distribution of HLT materials, the *Blueprint* aims at providing information to be used by policy organisations when assessing research projects aimd at developing HLT materials, for preparing policy plans concerning the acquisition, management, maintenance and distribution of HLT materials and practical information on how to acquire, manage, maintain and distribute HLT materials, answers to questions concerning the (re)usability of HLT materials after the consortia that were set up for their development cease to exist. All this information is presented in the *Blueprint* in nine chapters that, apart from the introductory chapter 1, deal with the following topics:

- Acquisition of HLT resources (Chapter 2)
- Processing of acquired data (Chapter 3)
- Linguistic processing of HLT resources (Chapter 4)
- Management of HLT resources (Chapter 5)
- Maintenance of HLT resources (Chapter 6)
- Distribution of HLT resources (Chapter 7)
- Support to users (Chapter 8)
- Recommendations for future policy (Chapter 9)

The following eight recommendations for future policy are made in the final chapter:

1. An HLT agency is necessary
   In order to prevent that HLT materials developed with government money outside a permanent infrastructure become obsolete and therefore useless, a legal body such as an HLT agency is required.
2. Organisation form of HLT agency and role of NTU
   This HLT agency could be a Dutch-Flemish consortium of institutions and should not be related to one existing institution in particular, because not all expertise is available in one single institution. A co-ordinator could be appointed by NTU to ensure that the interests of the whole HLT field are represented.
3. Tasks of the HLT agency.
   Primary tasks of an HLT agency:
   Task 1. Management
   Task 2. Guarantee accessibility of data and software
   Task 3. Maintenance
   Secondary tasks of an HLT agency:
   Task 4. User support
   Task 5. Acquisition
   Distribution should be entrusted ELDA and LDC.
4. Costs to be met by the government.
   Since extra costs for personnel and hardware will be incurred, additional government funding is required.
5. Costs to be met by the users of the HLT agency
   Depending on the specific use and user, general conditions must be agreed on that guarantee fair tariffs.
6. Acceptance of HLT data and software by the HLT agency.
   The HLT agency can refuse HLT resources that do not meet certain quality standards or that are not essential for a wide range of applications.
7. International participation.
   The HLT agency should be given the possibility, through government funding, to participate in European and/or global projects that are related to its tasks.
8. Development and maintenance of HLT expertise.
   Given the considerable shortage of language and speech technologists, the government should stimulate policies that are aimed at developing and maintaining expertise in the field of HLT.

6. Future prospects

In the previous sections we have provided an overview of the results obtained within Action lines A and D. This has revealed that the aims identified in the *Action plan for Dutch in language and speech technology* have been achieved, at least for these two action lines. Now it remains to be seen how these results will be used in the future in order to achieve the ultimate aim of the "Dutch Human Language Technologies Platform" project: to further the development and secure the usability of an adequate digital language infrastructure for Dutch. To this end in the following sections we consider our future plans with respect to Action lines A (5.1) and D. (5.2).

6.1. Action line A

Since Action line A has already contributed to creating a co-operative framework in the HLT field in The Netherlands and Flanders, our future activities will be directed to maintaining and enlarging it. This entails among, other things, keeping our databases and websites up to date, ensuring communication between interested partners, gradually enlarging the initial network, identifying and promoting the inclusion of new representatives; increasing the visibility and the strategic impact of relevant results and new initiatives; fostering cooperation; providing a forum for discussing, exchanging and sharing experiences, best practices, information data and tools.

6.2. Action lines B and C: HLT priorities

The future activities of these two action lines will be directed to ensuring that the priorities identified in the survey are realized so that an adequate HLT infrastructure for Dutch is obtained.

6.3. Action line D: implementation of the recommendations in the HLT Blueprint

In the near future a number of Dutch-Flemish digital HLT resources will become available. These development projects, in many cases, do not provide a permanent infrastructure. As projects aimed at the development of digital basic resources mostly result in intermediary products, extra efforts and investments are needed in order to implement them in applications that find their way to the end users. Furthermore, when planning such large scale projects a lot of time is invested in building the necessary structures (often at a supra-institutional level) and finding the right experts. The completion of a project often means that the managerial and operational structures cease to exist. Therefore it is of vital importance that the right measures are timely taken in order to ensure that the resources are stored in such a way that they will be expertly managed and maintained. When establishing an adequate infrastructure for maintenance of digital basic resources, proper attention should be given to a) intellectual rights, overall responsibility and co-ordination, b) actual physical management and maintenance of the resources and c) maintenance of expertise. In the following sections we will describe the facilities that we envisage to implement in the Dutch speaking area in the near future.

6.3.1. Necessary facilities

*A. Intellectual rights, responsibility, co-ordination: **NTU***

A careful transfer of intellectual rights is of crucial importance to the exploitation of resources. Furthermore, after completion of projects a visible policy responsibility is needed, even if the actual management and maintenance is carried out by an HLT agency (see B).

Organisational structure: The NTU (Nederlandse Taalunie/Dutch Language Union), representing a permanent Dutch-Flemish infrastructure, can act as the appropriate legal body handling all legal affairs. A member of the NTU will be appointed as co-ordinator and supervise from a policy point of view management, maintenance and exploitation of HLT basic resources that are contributed to the HLT agency (see B)..

The NTU will look after the interests of the entire HLT field and will function as a kind of ‘broker’ by:

- supervising the activities of the HLT agency (see B) and the various HLT committees (see C);
- looking after legal issues;
- stimulating the application of international standards;
- stimulating funding bodies to stipulate that in proposals proper attention is paid to allocating funding for management and maintenance and that resources financed with public funding be made available through the HLT agency;
- playing an intermediate role in the acquisition of digital data, e.g. from the industry.

B. Management and maintenance of digital resources: HLT agency

The *Blueprint* recommends the co-operation of the institutes in a consortium, an **HLT agency**, as this makes it possible to use dispersed expertise and infrastructure. This construction clearly has a number of advantages:

- efficient use of persons and means can be cost-reducing;
- combining resources and bringing together different kinds of expertise can create surplus value (e.g. extra applications);
- offering resources through one window (one-stop-shop) will create optimal visibility and accessibility;
- in international projects the Dutch language area can act as a strong partner;

Organisational structure: The HLT agency can take the form of a Dutch-Flemish consortium of organisations contributing their resources and expertise in a virtual resource centre. These organisations should strike binding agreements for a determined period of time. One Dutch-Flemish organisation (e.g. the Dutch Institute of Lexicology in Leiden) should be appointed as responsible co-ordinator.

- management: taking the appropriate (mostly technical) measures so as to make sure that data and software remain operational and usable;
- accessibility data and software: facilitating reusability of HLT resources: e.g. technical, legal and administrative settlements so as to optimise the route from developer via HLT agency to the distributor;
- maintenance: taking the appropriate measures to ensure long-term usability of data and software: technical maintenance of formats of HLT data, HLT software, system and application software, equipment; maintenance of legal contracts; content management of the HLT data and annotations;
- service: help desk, service to the users of the HLT data and HLT software (e.g. advising, maintenance of website and mailing lists, supplying tailor made data or software on demand);
- acquisition: active acquisition of HLT data and HLT software developed by the industry or research institutes;
- evaluation and validation: contributing to establishing international standards and methods for evaluating and validats Tw.1(i5(f)88.1(H)32.7(L)26.3(nr10.52.1(-11.4d)-5.4.5(u

- act as a knowledge base for questions concerning the resources contributed to the HLT agency;
- act as intrinsic supervisors on management, maintenance and exploitation of specific resources;
- act as advisors in specific domains s.a. language and speech technology, terminology, lexicology;
- be instrumental in the organisation of ‘major repairs’ of the resources that are put in their custody;
- be instrumental in developing the appropriate infrastructure for new projects or updating of existing results in their domain.

The HLT management committee will be responsible for the co-ordination, overall management, maintenance and distribution of HLT resources. It will

- act as general knowledge base and give advise in the broad field of language and speech technology, terminology, lexicology etc..
- act as general intrinsic supervisor on management, maintenance and exploitation of finished resources;
- be instrumental in developing the appropriate personnel infrastructure for new projects or updating of existing results.

6.3.2. Financing

Since the exploitation of basic resources will not result in considerable revenues, the authorities have expressed their explicit wish to make these resources available as broadly as possible. This results in keen prices: cost price for non-commercial research, a higher but not prohibitive price for commercial organisations. Consequently, the implementation of the above mentioned structures requires extra funding. Since a considerable percentage of the development costs should be allocated to management and maintenance, by combining the infrastructures required for different projects the percentage the costs would decrease. This applies as much to the material infrastructure (equipment, data, software, licences, etc…) as to the immaterial infrastructure (experts, personnel etc.). As is stressed in the recommendations of the *Blueprint*, the activities of the HLT agency cannot be carried out by the consortium partners in addition to their daily work, but require extra staff. Based on the data in the *Blueprint* and on experiences in other projects, a number of persons will be appointed at one or more of the organisations forming the HLT agency (e.g. experts on language and speech technology, IT-specialist, administrative personnel etc.). One overall co-ordinator and at least one secretary of the committees will be appointed at the NTU.

It is to be expected that the costs will increase with the increase of project results contributed to the HLT agency. These costs should be covered with funds allocated to management, maintenance and accessibility at the start of the development of new projects.

6.3.3. Conclusions

After the completion of projects aimed at developing HLT resources, efforts are needed to ensure long-term usability of the results. Timely attention to intellectual property rights, management, maintenance and distribution can guarantee that investments pay off in the future. In this respect, it is recommended, to make optimal use of existing expertise and infrastructure. In concrete this would mean that in the Dutch speaking area:

- the co-ordinating policy responsibility and as much intellectual property rights as possible should be placed in the hands of the NTU;
- the actual exploitation (management, maintenance and distribution) should be entrusted to a Dutch-Flemish HLT agency, that will take the shape of a consortium of institutions but acts as a one-stop-shop of digital HLT resources for the Dutch language
- the existing expertise should be combined as much as possible in a number of Dutch-Flemish steering committees consisting of representatives of projects, the results of which are contributed to the HLT agency and a co-ordinating Dutch-Flemish HLT management committee.

The NTU envisages to implement the above mentioned structures in its new long-term policy plan (2003-2007).

7. General conclusions

In this paper we have reported on the activities that in the last two years have been carried out within the framework of the project "Dutch Human Language Technologies Platform". In particular, we have focussed on two of the four action lines within this project: Action line A, which was aimed at raising awareness of the results of HLT research and promoting communication among interested partners, and Action line D which was concerned with management, maintenance and distribution of HLT resources.

Our overview of the results obtained so far has revealed that a cooperative framework has been created and that there are clear plans to set up a structure that will take care of all HLT resources developed with public funding, so that they will remain available for all interested parties: an HLT agency. In other words, the goals of action lines A and D have been achieved (for the results of B and C, the reader is referred to Binnenpoorte et al. (2002)) and clear directions for how to proceed in the near future have also been outlined. To conclude, it seems that in the Dutch speaking area pioneering work has been carried out from which other countries can probably profit in their attempts to start similar initiatives.

8. Acknowledgements

We are indebted to the steering committees of Action lines B, C, and D and to the authors of the Blueprint (P. Van der Kamp, T. Kruyt, and P. Van Sterkenburg) and of the Report B and C (G. Bouma, W. Daelemans, A. Dirksen, D. Heijlen, F. de Jong, J.-P. Martens, A. Nijholt, H. Strik, D. van Compernolle, F. van Eynde, and R. Veldhuis) for their invaluable contribution to the work presented in this paper.

9. References

Binnenpoorte, D., de Vriend, F., Sturm, J., Daelemans, W., Strik, H., and Cucchiarini, C. (2002). A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch. In *Proceedings of LREC2002*.