A Human Language Technologies Platform for the Dutch language: awareness, management, maintenance and distribution

Catia Cucchiarini, Elisabeth D'Halleweyn and Lisanne Teunissen

Nederlandse Taalunie
Postbus 10595
2501 HN Den Haag
The Netherlands

c.cucchiarini@let.kun.nl
{edhalleweyn, lteunissen}@ntu.nl

Abstract

In this paper we report on two of the four action lines within the project "Dutch Human Language Technologies Platform": 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 reveals that the goals of action lines A and D have been achieved and that there are clear directions for how to proceed in the near future. We hope that the experiences of the Dutch speaking area may be useful to other countries that intend to start similar initiatives.

1. Introduction

At the previous LREC conference in Athens we reported on a Dutch-Flemish project that had just started: The Dutch Human Language Technologies Platform. In that paper we mainly presented the background of the project and the action plan. Now, after two years of hard work, we are able to report on the results of this project.

The Dutch HLT Platform is a supranational initiative aimed at strengthening the position of HLT in the Netherlands and the Flemish part of Belgium. The plan to set up this platform was launched by the NTU, which is an intergovernmental organisation that has the mission of dealing with all issues related to strengthening the position of the Dutch language. 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)

2. The Dutch HLT Platform: action plan

The main purpose of the Dutch HLT Platform is to contribute to the further development of an adequate HLT infrastructure for Dutch. To this end, in 1999 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.

Given the wide scope of the project, it is impossible to deal with all four action lines in one paper. Therefore, action lines B and C will form the topic of a companion paper (Binnenpoorte et al, 2002), while in the present paper we discuss action lines A and D in more detail. Owing to the differences in aims and results between these

two action lines, more space will inevitably be devoted to Action line D than to Action line A.

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. Action line D: results

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.

4.1. The HLT Blueprint: aims

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 the *Blueprint*, which is to provide guidelines for the acquisition, management, maintenance and distribution of HLT materials, a number of specific goals are addressed in this document. In particular, the Blueprint provides:

(1) information and evaluation criteria to be used by policy organisations when assessing research projects aimd at developing HLT materials.
(2) information that policy organisations can use for preparing policy plans concerning the acquisition, management, maintenance and distribution of HLT materials.
(3) practical information on how to acquire, manage, maintain and distribute HLT materials.
(4) answers to questions concerning the (re)usability of HLT materials after the consortia that were set up for their development cease to exist.

4.2. The HLT Blueprint: contents

All the information mentioned above is presented in the Blueprint in nine chapters that, apart from the introductory chapter 1, deal with the following topics:

4.2.1. Chapter 2 Acquisition

In this chapter various scenarios for the acquisition of HLT are first presented. Subsequently, legal and financial issues that concern HLT materials are discussed. In the remainder of the chapter, the selection of digital texts and digital corpora of spoken language is discussed. In this connection attention is paid to the acquisition, production and typology of text databases, but also to technical aspects regarding the collection of digital databases and the realization of digital recordings of spoken language. Finally, some recommendations for the acquisition of HLT materials are presented, while model agreementsare provided as annexes.

4.2.2. Chapter 3 Processing acquired data

This chapter deals with the various ways in which digital written and spoken language data can be processed and administered. Conversion and adaptation procedures are discussed, as well as the importance of well structured directories and transparent administration for data storage. Finally, some conclusions and recommendations for policy organisations are presented.

4.2.3. Chapter 4 Linguistic processing of language material

This chapter deals with the way in which language data can be enriched with linguistic information, a process that is also known as annotation. The language data in question can be corpora of written or spoken language, digital dictionaries and computational lexicons. These different types of language data are addressed in succession. For each data type the annotation procedure is explained in detail while information is also provided on the personnel required for such tasks. This information is intended for those who want to set up projects for developing these resources, but also for organisations that have to evaluate such projects or maintain such resources. The last part of this chapter is devoted to issues such as the choice of hardware platforms, operating systems and programming languages. Recommendations for programming, documentation, assessment and future policy are also provided.

4.2.4. Chapter 5 Management

Management of HLT materials is the central topic of this chapter. HLT materials are language data (speech corpora and corpora of written text) and software for processing such data. Practical issues such as personnel, technical management, storage and conversion are also addressed.

4.2.5. Chapter 6 Maintenance

Once HLT materials have been developed, they need to be maintained in order to ensure that they remain usable. In general, maintenance is often omitted in many project proposals that envisage the development of HLT materials, probably because the importance of maintenance becomes apparent only after the development project has finished. In the long run, however, this practice will inevitably lead to a waste of public money because materials that have been developed with government money are no longer usable. This chapter covers various aspects of maintenance, such as technical maintenance (hardware, software, computer platforms and storage media) content maintenance and legal maintenance (contracts with providers, developers, distributors and users).

4.2.6. Chapter 7 Distribution

This chapter begins with a presentation of the various scenarios according to which the distribution of HLT materials can be realised. The legal requirements to be met when distributing HLT materials are then discussed. Subsequently, attention is paid to the financial and the technical aspects of data distribution. Finally, practical issues such as manuals and documentation are dealt with.

4.2.7. Chapter 8 Support to users

Distribution also implies providing support to the users of the materials distributed. Support can be provided in various ways. In this chapter the following possibilities are described: a website with on-line help information, a helpdesk, mailing lists, the provision of tailor-made software and data, software services and consultancy.

4.2.8. Chapter 9 Recommendations for future policy

On the basis of the information presented in the previous chapters, eight recommendations for future policy are made in this final chapter. These recommenadtions are presented in detail in the following section.

4.3. The HLT Blueprint: recommendations

4.3.1. Recommendation 1: An HLT agency is necessary

In order to prevent that HLT materials that have been developed with governement money outside a permanent infrastructure become obsolete and therefore useless, a legal body such as an HLT agency is required.

4.3.2. Recommendation 2: Organisation form of HLT agency and role of NTU

The permanent infrastructure to be set up could be a Dutch-Flemish consortium of institutions. The agency should not be related to one existing institution in particular, because some HLT materials require various sorts of expertise that are not usually available in one single institution.

In order to guarantee optimal co-ordination among the members of the consortium a co-ordinator could be appointed by Nederlandse Taalunie/Dutch Language Union (NTU) as theNTU can ensure that the interests of the whole HLT field are represented.

4.3.3. Recommendation 3. tasks of the HLT agency.

The following two criteria should be adopted in establishing the primary and secondary tasks of the HLT agency: (a) HLT resources deriving from government-funded, temporary projects for which no permanent infrastructure is available go automatically to the HLT agency under the restriction of recommendation 6.
(b) the distribution of HLT materials should be entrusted to specialised organisations such as ELDA and LDC.
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

4.3.4. Recommendation 4. Costs to be met by the government.

The activities of the HLT agency cannot be carried out by the consortium partners in addition to their daily work, but require extra staff. Moreover additional equipment will probably be necessary. Since these extra costs for personnel and hardware cannot be borne by the users of the HLT agency (see recommendation 5), it follows that additional government funding is required.

4.3.5. Recommendation 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. In the case of particular requests, such as those concerning the realisation of tailor-made products, an amount should be invoiced that at least covers the costs of production.

4.3.6. Recommendation 6. Acceptance of HLT data and software by the HLT agency.

The HLT agency has the right to refuse HLT data and software that do not meet certain quality standards (also concerning documentation) or that are not essential for a wide range of applications. This applies across the bord, irrespective of whether such materials have been developed by a company, by a renowned research institution or within the framework of a project and without a permanent infrastructure.

4.3.7. Recommendation 7. International participation.

To secure the position of the Dutch language in multilingual HLT research and product development, 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. Both on a national and an international level the HLT agency should contribute to the definition of standards and methods for the evaluation and validation of HLT language materials.

4.3.8. Recommendation 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.

5. 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).

5.1. Action line A

From the overview of the results presented in Section 3, it is clear that Action line A has greatly contributed to creating transparency and structure in the HLT field in The Netherlands and Flanders. Since these two important goals have now been achieved, our activities in the future will no longer be directed to creating a cooperative framework, but rather 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.

5.2. The HLT Blueprint: implementation of the recommendations

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.

5.2.1. 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)..

Tasks:

The NTU will look after the interests and demands of the entire HLT field and will function as a kind of 'broker';

- will supervise the activities of the HLT agency (see B) and the various HLT committees (see C);
- will look after legal issues;
- will stimulate the application of international standards;
- will stimulate funding bodies to stipulate that in proposals proper attention is paid to allocating funding for management and maintenance;
- will stimulate funding bodies to stipulate that resources that have been financed with public funding should be made available through the HLT agency;
- can play 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 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.

Tasks:

- 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 validating HLT resources.

For the actual, physical distribution of the resources appeal will be made on the expertise of organisations s.a ELRA and LDC as they have the proper expertise and marketing tools.

C Expertise: Dutch-Flemish steering committees and HLT management committee

In dissolving the managerial and operational infrastructure after the completion of a project, valuable specific knowledge concerning the project may be lost causing difficulties in the exploitation of the results. All the same it would not be realistic to maintain these structures. A solution would be to install a number of Dutch-Flemish **steering committees** and one co-ordinating Dutch-Flemish **HLT management committee**. The tasks of these committees should not be too heavy, but to ensure continuity and effectiveness a strong secretarial support should be provided

Organisational structure: For each completed large scale project the results of which are contributed to the HLT agency, a steering committee should be installed. Each steering committee delegates one representative to a co-ordinating HLT management committee. For small scale projects it has to be examined whether the necessary expertise is already present in the HLT management committee. Probably one expert, responsible for the

combined 'small' projects, will be added to this committee. The various committees should receive the appropriate secretarial support.

Tasks:
The steering committees will be responsible for specific resources and specific domains. They will

- act as a knowledge base for specific 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.

5.2.2. Financing

It is to be expected that 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.
When funding HLT resources, a considerable percentage of the development costs should be allocated to management and maintenance after completion of the projects. By efficiently combining the required infrastructures of different projects the percentage can decrease as doubles can be avoided. This applies as much to the material infrastructure (equipment, data, software, licences, etc…) as to the immaterial infrastructure (experts, personnel etc.). The *Blueprint* describes in detail the costs that can be expected, it gives for example the estimated work load and the type of personnel that is needed for each task of the HLT agency. 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. This costs should be covered with funds allocated to management, maintenance and accessibility at the start of the development of new projects.

5.2.3. 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 that starts as of 2003.

6. 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.

7. Acknowledgements

We would like to thank the steering committee of Action line D and the authors of the Blueprint (P. Van der Kamp, T. Kruyt, and P. Van Sterkenburg) for their invaluable contribution to the work presented in this paper.

8. 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*.