Consolidating and Increasing the Availability of Norwegian Human Language Technology Resources

Project group appointed by the Ministry of Culture and Church Affairs

Report delivered October 2002

CONSOLIDATING AND INCREASING THE AVAILABILITY OF NORWEGIAN HUMAN LANGUAGE TECHNOLOGY RESOURCES

Project group appointed by
The Ministry of Culture and Church Affairs
Report, October 2002

Foreword

This document comprises the final report of the project group appointed by the Ministry of Culture and Church Affairs on 19 March 2002 and charged with examining key issues within the sphere of consolidating and increasing the availability of Norwegian human language technology resources.

The project group members are all in agreement regarding the substance of this report. Although all the recommendations and conclusions provided here express the views of the project group alone, we would like to give special thanks to the extended resource group representing the various user environments. The resource group has offered a wide variety of valuable and useful insights throughout the efforts leading up to this document.

While the conditions for the Sami language were not included in this project group's mandate, we are aware that efforts have been initiated to establish Sami language technology resources.

Torbjørn Svendsen
Chair (until June 2002)

Torbjørn Nordgård
Deputy Chair (Chair from June 2002)

Leiv Hartly Andreassen

Jon Trygve Berg

Knut Kvale

Tron Espeli

Stig Johansson

Torbjørg Breivik
Project Secretary

Translation: Carol B. Eckmann, Information services and language consultancy, Solfallsvn. 31, NO-1430 Ås, Norway

MEMBERSHIP, MANDATE AND WORK PROCESS

The following project group was appointed by the Ministry of Culture and Church Affairs in a letter dated 19 March 2002:

Torbjørn Svendsen, Professor, Norwegian University of Science and Technology (NTNU) (chair)
Torbjørn Nordgård, Professor, NTNU (deputy for Svendsen and deputy chair of the project group)
Tron Espeli, Adviser, Programme Coordinator (ICT), Research Council of Norway
Leiv Hartly Andreassen, Managing Director, SAIL Port Northern Europe AS (SPNE)
Knut Kvale, Professor, Senior Research Scientist, Telenor ASA
Jon Trygve Berg, Chief Technology Officer, Nordisk språkteknologi AS (NST)
Stig Johansson, Professor of Modern English Language, University of Oslo

All members had personal deputies who have also taken active part in the activities of the project group:
Bernt-Erik Heid, Senior Adviser Research Council of Norway (for Espeli)
Anja Hilt, Head of Telecom Technology Department, (for Andreassen)
Robert Engels, Manager R&D Projects, CognIT a.s. (for Kvale)
Helge Dyvik, Professor of General Linguistics, University of Bergen (for Johansson)

Bente Maegaard, Director of the Centre for Language Technology, Copenhagen, and Lars Ahrenberg, Professor of Computational Linguistics at the University of Linköping, Sweden, have served as observers and external resource personnel.

On behalf of the Ministry of Culture and Church Affairs, the Norwegian Language Council appointed the following resource group to assist the project group:

Jan Olav Fretland, Associate Professor, Norwegian Language Council (chair)
Grete Knudsen, Adviser, GBM-Partners, Bergen
Per Morten Hoff, Secretary General, ICT-Norway, Oslo
Bjørn Seljebotn, Managing Director, Nynodata a.s, Bø in Telemark
Trond Andreassen, Secretary General, Norwegian Non-Fiction Writers' and Translators' Association (NFF), Oslo
Øyvind Haaland, Country Manager, Berlitz GlobalNET, Bergen
Ove Nyland, Manager, Noreg.no, Leikanger
Petter Korseth, Assistant Director, Norwegian Board of Education, Oslo
Kristin Bech, Language Technology Coordinator, HIT-Centre/University of Bergen
Bjørn Norman Hansen, Executive Vice President (and Chief Credit Officer), Norwegian Industrial and Regional Development Fund, Oslo
Ruth Vatvedt Fjeld, Professor, University of Oslo

Secretariat:
Adviser Torbjørg Breivik served as the secretary for the activities of both groups.

MANDATE

“To consolidate and increase the availability of Norwegian human language technology resources”

OBJECTIVE

To study and clarify the framework for consolidating and increasing the accessibility of Norwegian human language technology resources.

The project group was asked to provide an updated, realistic analysis of the need to increase the availability of Norwegian language resources and to estimate the investment required for this process. Previous need assessments were to be re-evaluated and the final report was to specify a minimum volume (as well as recommended volumes) of the types and amounts of resources needed for research and industrial purposes, respectively. Legal issues related to these efforts were assessed in a separate report (only available in Norwegian).

This report also outlines a potential model for how to organize the activities, including proposals for operations based on the principles of an independent structure (independent legal entity), staffing, and requirements regarding expertise within the organization.

In addition, a financial plan has been designed specifying the cost associated with establishing a minimum resource collection contra a collection of the recommended magnitude. Financing schemes are based on combined private and public funding.

WORK PROCESS

The project group has held six meetings, while the resource group has held three. Between meetings, the members of both groups have submitted input and proposals. The project group has weighted the various points of the mandate according to its own views and assessment capabilities. The project group has participated in two study trips, one to ELRA/ELDA* in Paris, France and one to NST** and SPNE*** in Voss, Norway.
The efforts of the project group have been focused on identifying the various resources that are needed, designing a financial plan and devising recommendations for how to implement the proposals contained in this report.

*European Language Resource Association / European Language Distribution Agency
**Nordisk språkteknologi AS
***S.AI.L. Port Northern Europe AS

CONTENTS

CHAPTER 1 CONCLUSIONS

The project group has been charged with examining the framework for consolidating and increasing the availability of Norwegian language technology resources. The group has chosen to utilize the term "språkbank" (Norwegian for "language bank") to designate a collection of such resources. In the English version, however, the collection is referred to as a Norwegian human language technology resource collection. A "language bank" is responsible for administering its capital, which in this case comprises national language resources. Like financial banks, which do not keep all their capital in a single vault, a language bank does not need to be confined to a single location.

It is the unanimous conclusion of the project group that a collection of Norwegian language technology resources must be created as soon as possible. This view also represents the consensus of the resource group as well as a unified Norwegian research and industrial community within the field of human language technologies (HLT). A resource collection of this type will be vital to efforts to:

- help to fulfil the objective of ensuring that Norwegian – spoken and written – remains the dominant language of use in Norwegian society;
- ensure that Norwegian language technology promotes participation in society and enhances cultural identity by utilizing the overall Norwegian linguistic culture;
- strengthen the Norwegian language (*Bokmål*, *Nynorsk*, and the Norwegian dialects)[1] and counteract loss of domain, i.e. prevent English from gradually emerging as the language of use in an increasing number of areas;
- encourage the Norwegian ICT industry to invest in language technology solutions for Norwegian as well as other languages;
- fully exploit the potential for increased productivity inherent in a link between top expertise in ICT and linguistic fields;
- increase the incentives for foreign suppliers to create Norwegian language products.

A number of countries in Europe are in the process of establishing human language technology infrastructures similar to that proposed here, primarily as a means of strengthening the position of national languages against the encroaching influence of English. These countries have a far better starting point for their efforts than Norway because key resources were allocated as far back as the 1990s. In every case, these initiatives have been based on substantial public funding, and in the Netherlands, including the Flemish part of Belgium, a survey and the collection process itself have been given 100 per cent government funding.

A Norwegian HLT resource collection should be organized as a publicly-owned foundation. Its board of directors should comprise broad representation from research and industry, and the number of administrative personnel should be kept to a minimum. The compilation, operation, maintenance and distribution of resources should be outsourced to external actors with relevant expertise and experience.

Norwegian HLT resources should be owned by a foundation under public administration in order to ensure clear ownership and user rights. A combination of private and public ownership would be unconstructive both for legal reasons and for reasons related to market competition. Realistically speaking, nearly all financing will have to come from public

[1] Norway has two official written languages, *Bokmål* and *Nynorsk*.

funding. This has been the case in the countries that have launched similar projects. The related Norwegian industry does not have the financial capacity to provide any substantial degree of funding. This has also been the situation internationally.

The resource collection must contain material that will be useful to the HLT industry as well as the research community, material that is representative of different areas of linguistic use (text, speech, dialects, both written versions of Norwegian, etc.), and all material must be well-documented and encoded in conformance with international standards. The material must be subject to quality assurance controls (validation) and all user rights must be clarified. A national HLT resource collection is necessary to ensure that the Norwegian language is adequately represented in ICT solutions that employ natural language. Without good-quality Norwegian-language products and services, English will increasingly come to dominate the commercial, educational and public sectors.

CHAPTER 2 WHY DO WE NEED A COLLECTION OF NORWEGIAN LANGUAGE RESOURCES?

Language and speech technology help to simplify and enhance communication. While a large number of products and services have been designed for the English language, only a small proportion of these are available for Norwegian. The ability to maintain a viable Norwegian language, culture and identity is contingent upon access to Norwegian language technology products and services. Such products must be based on a collection of language data of adequate size and quality.

Language is fundamental to our society. All communication between people is based on language, oral as well as written. And in today's world, it has also become more and more common to utilize natural language in communication between people and machines.

2.1 Why are Human Language Technologies Necessary?

The role of information and communications technology (ICT) is gaining increasing importance within society. The use of computers and need for Internet access are not limited to private individuals and the workplace; ICT is being incorporated into virtually all types of technology, from household appliances, consumer electronics and automobiles to professional and industrial equipment and systems. In an expanding number of contexts, users are no longer contending with a lack of information or built-in features, but rather with a surfeit of possibilities from which they must choose.

The exchange of information often implies the use of an electronic medium for storage and transfer. The effectiveness of information exchange is contingent upon functional tools for document generation and editing.

Human language technologies (HLT) involve simplifying and enhancing communication between people and facilitating the man-machine interface. Such technologies make it easier to utilize modern information technology because they allow users to communicate in the mode they know best – their own oral and written language. This in turn lowers the threshold for utilizing information technology, enabling more people to access information, services and products. Examples of human language technologies include automatic speech recognition (computer generated text from speech input) and machine generated speech, machine translation and applications for document production and information retrieval.

Human language technologies can be used to rationalize many kinds of work processes. This can be illustrated by means of a few examples, starting with the hospital sector. More than 3,000 man-years are devoted to transcribing dictation into written journals, and wage costs exceed NOK 300,000 per man-year. This represents a total annual cost of approximately NOK one billion. Experience from Philips AS indicates that clerical personnel can produce reports up to 40 per cent faster with the help of automatic speech recognition tools. A 10 per cent gain in effectivity in relation to ordinary hospital dictation – which is a conservative estimate by any measure – would be sufficient to provide the funds needed to establish a minimum database volume for a Norwegian HLT resource collection in the course of a single operational year. There is good reason to expect that a great deal more will be saved as dictation tools are gradually incorporated into many other public sector areas, assuming that the linguistic data needed to train the systems is available. And these estimates do not take into account the additional potential savings within the private sector.

Another example is retrieval of information stored in electronic form. It is a common problem that an increase in the volume of information reduces accessibility. Human language technologies can curtail this problem by means of automated indexing and retrieval of information ("digital librarians").

A third example involves machine translation between English and Norwegian, or machine-assisted translation between these two languages. The need for text translation within the public and private arena is vast, and many resources have been invested in manual translation. This work must be carried out by highly qualified personnel to ensure that technical and juridical information, for example, are correctly translated. Under normal circumstances, machine-assisted translation can reduce the time needed for translation by 20 to 40 per cent. Assuming that some two thousand man-years are used annually for Norwegian-English translation, the implementation of relevant tools would lead to large savings in the public as well as the private sector. Again, the investment in Norwegian HLT resources would quickly be recouped.

Many different HLT products and services can already be found on the international market. Unfortunately, however, most of these are not available in Norwegian for either *Bokmål* or *Nynorsk*. One of the reasons for this is a lack of language resources. English has become the dominant language. In order to truly reap the benefits of human language technologies, users must be able to utilize their native language. Only then will such technologies be accessible to all.

The population of Norway is low compared with countries such as Germany, France and Great Britain. As a result, the market for Norwegian language products is modest, and far smaller than for the key European languages. This will limit how much national industry can be expected to contribute to the costs of establishing the recommended Norwegian HLT resource collection, especially since the expense of developing HLT products remains basically the same regardless of the specific language. In this context, it should also be pointed out that the content of an HLT resource database would be of great benefit to the language research community at large. Only if there is inexpensive access to basic linguistic resources will Norwegian and international players will be encouraged to view Norway as an interesting market.

Like all other commercial activities, the goal of the HLT industry is to develop profitable products, i.e. products that the market both needs and wants. Today, HLT products and applications employ spoken commands and speech recognition to carry out household tasks, enhance the efficiency, safety and flow of traffic (navigational systems), translate between different languages (machine translation), and help to simplify daily tasks in many other areas, particularly for people with various kinds of disabilities.

Nonetheless, the fact remains that these products are by and large only available for English or other major languages. The HLT industry is capable of adapting these products for use by all people living in Norway, regardless of an individual's dialect or the environment within which the product is utilized. However, achieving this is contingent upon access to Norwegian HLT resources. Such resources would make it possible to use Norwegian on a par with English within a technological sphere that will become increasingly important to users in coming years.

It is an overall political objective to uphold Norwegian as the language of use in all contexts in Norwegian society. This entails devising a framework that best enables Norwegians to continue using Norwegian to communicate with each other. It is through our native language that we best can express ourselves and understand one another. English is our second language. Tools that are available for English but not Norwegian may well reinforce the tendency to select English instead of Norwegian, for instance as an internal corporate language.

Language is an inherent part of culture and identity. This is the case in every social context – at home, in school, at work, and during recreation. Increasingly, computers are becoming a new "partner" in communication. Allowing the Norwegian language to be suppressed by English because we cannot afford to establish a framework for Norwegian-language technology would represent a serious cultural policy setback.

2.2 Linguistic Data

The development of human language technologies requires a combination of technological know-how, linguistic expertise and digital language resources. With few exceptions, all modern HLT utilizes various forms of statistical models. For example, a dictation system that automatically converts speech to text uses statistical modelling of the annotated part of speech sounds, and of the links between words. These models need to be trained by means of examples of speech and text from large-scale databases. The training of statistical models requires a much larger textual basis than that needed to produce a traditional grammar or dictionary, for example. The training phase is the most vulnerable phase of the system creation process. If the input foundation, i.e. the training data, is too small or of unsatisfactory quality then the end product will be inferior. The correlation between the input data and end-product quality applies not only to dictation systems, but also to machine translation, speech synthesis, proofreading software, and more.

The lack of Norwegian linguistic data represents the greatest obstacle to enabling all inhabitants to gain equal accessibility to, and have equal ability to utilize, the new technology. The language industry cannot develop *Norwegian language* products without sufficient Norwegian linguistic data. Requirements regarding the volume and quality of linguistic data are basically the same independent of language, which means that the outlay for establishing HLT resources for Norwegian would be roughly the same as for English. Due to the special situation in Norway, where there are not only two written languages with a significant degree of free word choice but also a high tolerance for use of dialects, the costs are likely to be somewhat higher than for other European languages it would be natural to use for comparison.

2.3 The Situation Internationally

Many other European countries have understood the imminent risks to their national languages if they do not themselves take the initiative to establish the resources needed to ensure that their inhabitants will have access to the new services in their own languages. The EU has given priority to linguistic diversity, and a number of database compilation projects received support during the 1990s through the Framework Programmes, either as projects specifically designed for data collection or in connection with research projects.

The European Language Resource Association / European Language Distribution Agency, (ELRA/ELDA) in Paris, is an organization that distributes language resources. Activities take place primarily within the EU countries, although the organization also cooperates with

similar organizations, such as the Linguistic Data Consortium (LDC) in the USA. ELRA is a membership organization. As such, it cannot sell any products, but has founded ELDA as a distribution agency to market the available resources and services via an Internet-based catalogue (http://www.elda.fr/cata/tabtext.html). ELRA has a contract with the EU Commission stating that all institutions receiving EU funding for projects involving collection of linguistic material must make these resources available to ELRA/ELDA. The intellectual property rights to and ownership of the language resources distributed by ELDA remain with the providers. ELDA is responsible for independent validation of the resources and channels royalties back to the rightsholders after sale. The Association has been established as a non-profit organization.

The ELDA catalogue contains text corpora, speech corpora and lexical data for various languages. The different resources are not necessarily available in all relevant languages. Many of the resources are multilingual, and the catalogue also contains non-European languages such as Japanese and Chinese. The Norwegian company Telenor AS has participated in one of the EU-funded projects, and as a result Norwegian is represented in a minor part of a large-scale multilingual speech corpus (SpeechDat). The largest corpora include the British National Corpus (BNC), with 100 million words and the European Corpus Initiative (ECI), which is multilingual and contains 98 million words.

In most of the countries where compilation of language resources has been initiated, efforts are the result of cooperation between several sectors, including national authorities, and research and (language) industry circles. Of the current international initiatives we would like to mention Italy, France and The Netherlands/Belgium. Italy and France have both based their efforts on approximately 50 per cent public financing. In both countries, this involved expanding existing language data collections by producing new text and speech corpora within an overall cost framework of some eight million Euro. There are more than 20 million Dutch speakers in The Netherlands and Flanders combined. Speech data is being collected for nearly five million Euro, with 100 per cent government financing. This data collection corresponds well with the needs specified for Norwegian speech data. An important underlying motive for the full government funding of the Dutch project was to ensure that the intellectual property rights to the information would remain part of the public sector.

None of the Nordic countries possesses an overall national language resource database for use in both research and industrial development activities. A number of institutions have independently collected, refined and stored electronic language resources for their own purposes. For the most part, this comprises university and research circles working in the fields of HLT and computational linguistics. The closest thing to a national resource database for language research in the Nordic countries is probably the collection amassed at the University of Gothenburg, which consists of a Swedish speech corpus and a Swedish HLT resource collection (text data, lexical resources).

2.4 Human Language Technologies in a Political Context

According to European statistics from recent years, Norway is ranked towards the top of the list with regard to utilization of ICT products and services. This is in part due to the commitment of the Norwegian Government, which has both formulated plans and allocated funding for their implementation in this area. Several such plans are tied to developments in Norwegian as well as European R & D activities. In one of the most recent plans of action,[2]

[2] Cf. *eNorway 2005*, available in English at www.enorge.org/

three overarching visions for Norwegian IT policy are set out: creating value in industry, enhancing efficiency and quality within the public sector and promoting involvement and identity. When presenting the plan of action in the spring of 2002, Prime Minister Kjell Magne Bondevik stressed these points, elaborating on them by pointing to the importance of maintaining a national language, culture and identity.

The above plan of action represents the latest in a series of official documents dealing with the basis for Norwegian ICT policy. Other documents include a plan of action for the Norwegian language and ICT submitted by the Norwegian Language Council in 2001 (available in Norwegian only).

There is political consensus that everyone should have equal access to and an equal opportunity to utilize the technology that forms the basis for these efforts. This is also specified in the *Strategy for Electronic Content 2002–2004*,[3] presented by the Norwegian Government in April 2002, in which one of the stated targets is to ensure "...good access to high-quality electronic content produced in Norway or localized to Norwegian conditions." Moreover, the document asserts that "(g)ood access to technical terms and jargon will cut costs in the development of new products" (Chapter 1). The present report represents the first concrete document to emerge from the several action lines proposed in the strategy plan.

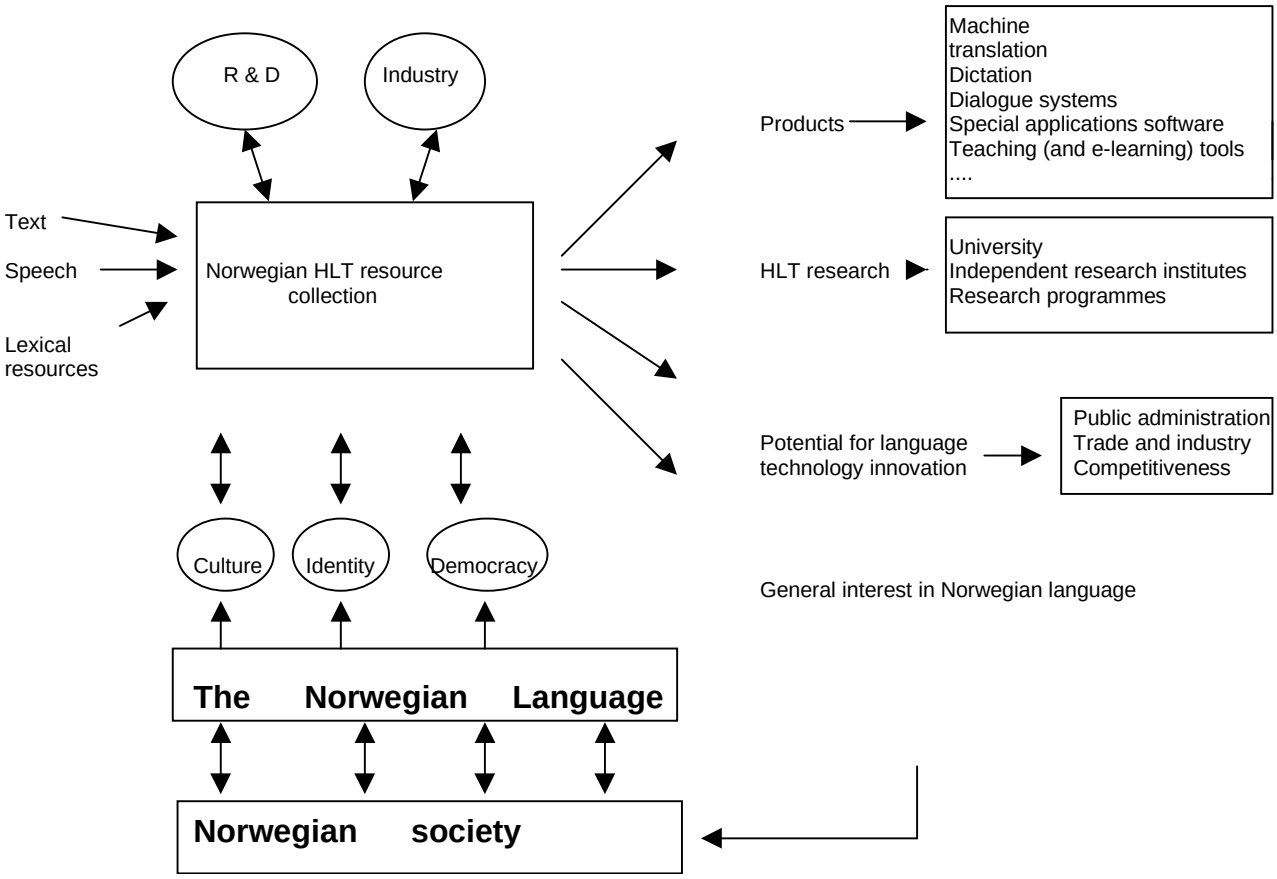
Many EU countries began compiling language data early in the 1990s on the basis of the same motivation: to safeguard their national languages as the language of use in all areas of society, to maintain their cultures and identities, and at the same time, to attempt to stem the rapidly growing influence of English, particularly within the ICT sector. Some of the initiatives launched have been discussed above (2.3), and the language resources distributed by ELRA/ELDA are the results of these early compilation activities.

2.5 A Norwegian Human Language Resources Collection

A collection of Norwegian HLT resources must be compiled in order to generate language technology products and services in Norwegian. This material must be accessible to all players on the Norwegian market. This is a national responsibility in respect of both cultural and industrial policy. Given the limited market size, however, public funding will be required to cover most of the outlay. Together with the focus on language technology initiated by the Research Council of Norway, the establishment of a Norwegian human language resource collection will create a foundation for a Norwegian HLT industry as well as enhance the availability of Norwegian-language products and services. The language resource collection, combined with R&D activities and HLT industry, will provide the foundation needed to furnish Norwegian society with the Norwegian-language products and services that it needs.

[3] See www.enorge.org/

Figure 1 (below) shows how the HLT resource collection can function in relation to relevant players.



CHAPTER 3 ORGANIZATION

The HLT resource collection should be organized as a publicly-owned foundation. Its board of directors should comprise broad representation from research and industry, and the number of administrative personnel should be kept to a minimum. The compilation, operation, maintenance and distribution of resources should be outsourced to external actors with relevant expertise and experience.

3.1 Introduction

The tasks of a Norwegian HLT resource collection agency will be twofold. On the one hand, it will be responsible for ensuring that the resources are of a correct type and amount. On the other hand, it will be responsible for managing the investment in resources and ensuring that the database is utilized for relevant industrial as well as research purposes. While the former task is limited to the period during which the collection is being compiled, the latter is a more permanent endeavour. During the establishment phase, the responsibilities of the resource collection agency shall include:

- purchasing the rights to use existing material and creating a framework for further distribution;
- identifying material that can be distributed (without purchase of rights);
- organizing the production of new material.

The HLT resource collection agency must also see to it that existing material is adapted according to satisfactory technical specifications.

Another important task is to maintain and further develop the resources. Material that is not adequately maintained and augmented loses its value over time. Permanent tasks will include:

- distribution;
- responsibility for resource enhancement measures;
- responsibility for proposing production of new resources, e.g. by proposing how production can be funded once the collection has reached its basic level;
- responsibility for quality control in connection with refinement and new production of resources;

It is important to avoid devoting greater effort than necessary to the administration of the language resources.

3.2 Should the HLT Resource Collection Agency be Owner or Distributor?

A role as merely distributor of the language resources implies that the HLT resource collection agency is only responsible for making resources available for research and industrial purposes, while all rights to the material remain with the institutions responsible for their original production. Those who place their material at the disposal of the agency would thus be responsible for clarifying the legal rights to it. In the event of misuse, all liability would lie with the institution or individual who made the material available, and not with the resource collection agency itself.

One ramification of the distributor role is that the resource collection agency would be unable to refine, select or restructure resources in new constellations. This makes it difficult when

situations arise in which customers in need of specific resources prefer to have others assist them in choosing and compiling the resources they require. According to ELRA/ELDA, the role of distributor tends to diminish the levels of re-use of data resources. This indicates there may be greater advantages in ownership or an extended right of use.

However, a HLT resource collection agency that owns all its own resources will have to provide compensation to the holders of ownership and intellectual property rights associated with all existing material that is incorporated into the collection. The agency would then acquire all rights to refine, adapt, and restructure the material. The decision to purchase existing or produce new material will have to be considered in each individual case on the basis of the quality, re-usability and price of a collection in relation to the costs of new production. It may prove just as expensive and resource-intensive to reconfigure existing material for re-use as it is to produce new material from scratch.

A legal study requisitioned by the project group revealed potential difficulties as regards incorporating some of the existing language resources currently in the possession of various institutions into a national database. This is due to a lack of rights to use the material in contexts other than those for which it was produced. In particular, this applies to material that is made available for specially-defined purposes, which may in effect disallow all re-use. In cases such as this there are two choices: either to renegotiate existing agreements with each institution that has supplied material or simply to compile new material.

The most flexible solution is one that combines the roles of owner and distributor. Previously collected resources that can be made publicly available could then be distributed without incurring additional costs (e.g. material produced with public funding at the universities), while it also remains possible to initiate the production of new resources. One disadvantage would be that the lack of ownership or right of use will leave it up to the rightsholder to determine whether resources should be refined or further developed. Furthermore, this solution will limit the ability of the HLT resource collection agency to administer the resources strategically. Ideally, it would be best to have ownership or complete user rights to as much of the resources as possible.

3.3 Compilation of Resources

The organization of the resource collection agency should be streamlined, efficient and flexible. It would not be beneficial to establish a large organization to compile language resources only to have to dismantle most of it after the consolidation phase is completed. Instead, the agency should be responsible for commissioning the compilation efforts and for organizing these activities within a flexible, cost-effective project framework. Existing resources will need to be assessed in terms of whether they fulfil the needs specified for the HLT resource collection, whether they adequately satisfy quality requirements and whether the acquisition of ownership and/or user rights is economical compared to the cost of new production. New production of language resources should be based on tenders within a specified portfolio of resources given priority by the agency board.

All data to be incorporated into the Norwegian HLT resource collection, existing as well as newly produced, must be validated (subjected to quality control measures) by an independent institution separate from the production source. This is critical to ensuring that the quality of the language resources conforms to the stipulated requirements, and that the substance corresponds to the documentation.

3.4 Distribution

Wherever possible, existing distribution channels should be utilized to avoid unnecessary administration. At the same time, safeguards must be in place to ensure that the parties paying for the resource compilation have ownership and user rights according to signed contracts.

Resources for distribution may remain stored in their present locations if this proves to be most practical. Other resources that have been purchased or newly produced may be stored at various sites. What is important is that the resources can be delivered quickly, and that sufficient expertise is available to deal with needs in relation to administration, operation and tasks in connection with refinement, maintenance, reconfiguring, copying and transfer.

In the context of Europe, ELRA/ELDA is the key body for distribution of language resources. Norwegian language resources must be made accessible to international users through the ELRA/ELDA framework. The Norwegian HLT resource collection agency must cooperate closely with ELRA/ELDA as regards development, application of standards and quality requirements.

3.5 Type of Organization

In the view of the project group, two potential organizational models are relevant for the management of the Norwegian HLT resource collection: a limited company and a foundation. The discussion of the type of organization must take into account the relationship to the resources. If the agency's task will be solely to distribute the resources, then it will merely serve as a link between owner and user. This does not place any great constraints on the choice of organizational model. If, however, the agency is destined to be the owner of, or possess extended user rights to, the resources, the situation is a different one. In this case, the agency must function as a legal person (an independent legal entity).

In choosing between these two types of organization it is important to ensure that the agency will be flexible, will be able to take rapid action, and is authorized to outsource projects. It is also crucial that the membership of the board of directors represents a broad spectrum of relevant interests within industry and research. The ownership interests must also be adequately represented on the board. The board will be responsible for formulating strategies and assigning priorities. This will require in-depth insight into the needs of the HLT industry, and HLT-related R & D activities at the national as well as the international level.

In legal terms, a limited company and a foundation are on equal footing when it comes to latitude to take action, ability to increase capital during operation, public auditing of accounts and, for the most part, discontinuation of an organization. The difference lies in their ownership; a limited company has owners, while a foundation does not. In a foundation, the founder(s) provide the capital to the foundation. The two types of organization also differ when it comes to discontinuation or bankruptcy: with a foundation, the resources are returned to the founders, while the estate of a bankruptcy from a limited company can be purchased by anyone.

In a limited company, the owners are responsible for appointing or electing the board. However, if the founders of a foundation wish to control the manner in which the investment capital is administered, they must incorporate this into the statutes.

In the opinion of the project group, organization as a limited company poses certain problems because a company can go bankrupt, and the consolidated resources could thus be lost to society as a whole.

Establishing a company regulated by individual legislation would be one way of solving this, but it takes time to draft and adopt a law of this nature. Should this path be chosen, the work to establish the agency must be carried out in parallel with the compilation of the resource collection, as time is very much of the essence. It is critically important that the effort to compile Norwegian HLT resources be launched as soon as possible.

The project group considers the best alternative to be the establishment of a foundation with the relevant ministries as the founders. This will accentuate the role of the HLT resource collection agency as a purveyor of shared national resources. It will also create a more stable framework for the agency than organization as a limited company, thus eliminating the risk that these resources can be lost through bankruptcy. The foundation will need primary capital for independent use upon establishment. It would be natural for the ministries involved to provide the initial funding needed as a kind of "endowment," but that it also be possible for others to contribute to the financing, cf. the chapter on financing.

The statutes for the agency foundation must clearly state the purpose of the HLT resource collection and must define an appropriate structure for decision-making and advisory bodies. The board should consist of major players in relation to funding (ministries) as well as representatives of key user interests (HLT industry, research institutions, public and private user groups).

The foundation needs to have an independent position vis-à-vis the ministries, who will safeguard their own individual interests through membership of the board. For practical reasons, the foundation should be located together with an existing institution, for example the Norwegian Language Council, in order to minimize administrative costs (offices, technical infrastructure, general administrative support services) - especially in connection with the initial phase of establishment.

There must be no doubt as to where the ownership and right of use to the resource collection lies. Specific contracts must be designed to regulate the relationship between ownership and user rights in cases where the resource collection agency does not have formal ownership of the material. Distribution of existing resources will in no way affect ownership rights.

3.6 Operation and Maintenance

There are various ways to organize the operation of the Norwegian HLT resource collection agency. The project group envisions a scenario in which the Norwegian HLT resource collection agency foundation (which owns and administers the resource collection) establishes an operational enterprise, for example a limited company, to organize the compilation process and devise a framework for distribution and other services. Alternatively, these tasks can be outsourced to existing organizations. As a rule, the job of compiling the resources should preferably be awarded on basis of tenders and carried out externally by existing institutions and circles of qualified personnel with sufficient expertise in and experience with the production and distribution of linguistic data. The Centre for Humanities Information Technologies (HIT Centre) is closely tied to the University of Bergen and is an example of one such institution. Other examples of potential institutions for outsourcing are Nordisk språkteknologi (NST) and Sail Port Northern Europe AS (SPNE) in Voss, as well as a number

of private or university groups. This model will also save on costs, as several of these institutions are willing to invest a certain amount of their own resources in the activities. A set of framework agreements could be designed for cooperation with institutions and expert groups who are qualified to conduct compilation activities. The individual projects could then be negotiated in greater detail, or – in some cases – assigned after a round of tenders.

The agency board must ensure that there are clear guidelines stipulating how the operational enterprise is to follow up the priorities set by the board. The board will be responsible for assigning overall priority to the various types of language resources as well as the time-frame for and progression of the compilation process. This can be achieved with basic allocations accompanied by clear stipulations that nonetheless allow the company the flexibility needed to enter into favourable ad hoc agreements, e.g. in specific situations involving the purchase of rights to a substantial resource collection.

The tasks of maintaining and further developing the language resources can be administered in a similar manner and can also be carried out by the operational enterprise owned by the HLT resource collection agency. In this case, however, it may be more appropriate to establish a set of agreements for cooperation with institutions specially qualified to deal with these activities. One possibility would be to allocate funding earmarked for maintenance and further refinement of language resources over the budgets of the universities. The responsibility for determining which tasks should be given priority would remain with the agency board. A solution along these lines will ensure that maintenance activities are carried out by qualified experts. Moreover, this model would help to strengthen the university environment in HLT-related disciplines, which would lead to more research and greater expertise. At the same time, it provides a basis for increased educational capacity for highly-qualified personnel, which will directly benefit the industry.

The organization of the Norwegian HLT resource collection agency as envisioned by the project group is presented in Figure 2, below.

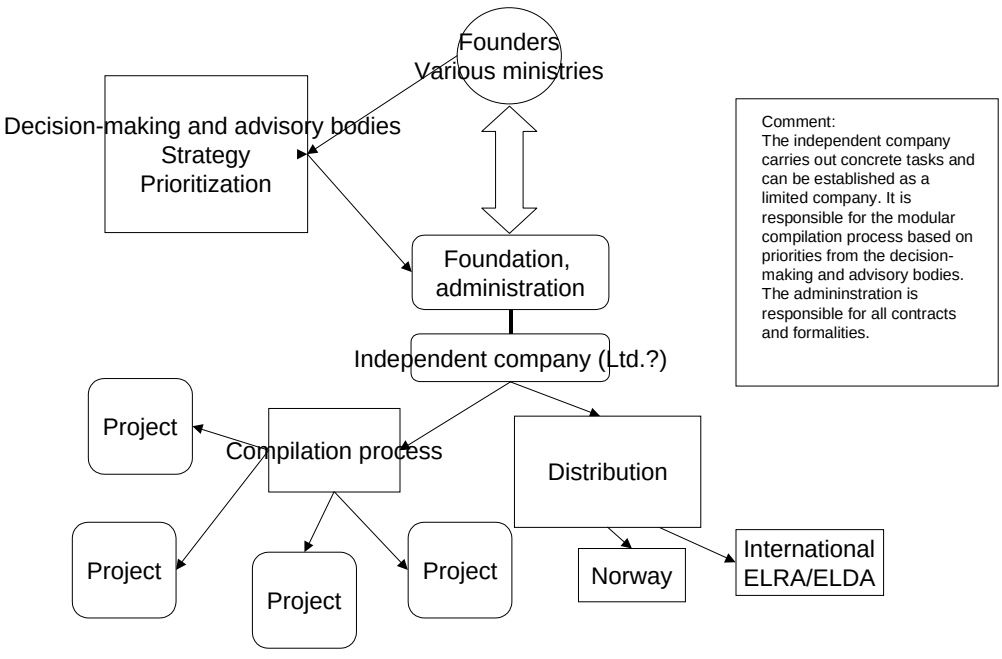


Figure 2: Organizational structure

The project group recommends that the Norwegian HLT resource collection agency have a small, permanent administration with one administrative officer in a full-time position. In addition, administrative resources will be necessary to support prioritization and planning activities as well as to formulate and negotiate agreements and contracts. It would be possible to obtain access to such administrative resources by means of a contract with another institution, such as the Norwegian Language Council. The operational enterprise will have more comprehensive tasks, as it will serve as the hub for the outsourcing and follow-up of concrete projects. Within the foundation (all decision-making and advisory bodies plus the administration), and particularly within the operational enterprise, expertise in the field of HLT will be a crucial requirement, while the operational enterprise will also need insight into and experience with the administration, follow-up and implementation of projects within a defined financial framework.

CHAPTER 4 THE CONTENT OF A NORWEGIAN HLT RESOURCE COLLECTION

The content of a national HLT resource collection should satisfy the fundamental needs of HLT-related research and product development, and should facilitate cost reductions in relation to the development and adaptation of Norwegian HLT products and services. All issues pertaining to rights of use must be clarified.

4.1 Introduction

The content of a Norwegian HLT resource collection must satisfy the following overall requirements:

- The material must be of use to both the Norwegian HLT industry and the research community at large.
- The material must be representative of different areas of linguistic use (text, speech, dialects, both written languages), and must to the greatest possible extent reflect current usage.
- The material must be well documented.
- All encoding and mark-up of the material must conform to international standards.
- All user rights must be clarified.

In a report discussing a Norwegian HLT resource collection as a national corpus (Svendsen, 1999), the content of a Norwegian language resource database was classified into broad categories such as speech data (sound recordings of various types), text data (various textual corpora) and lexical resources (word lists, terminology databases). In the view of the project group, this broad classification is still applicable.

Much of the discussion in the 1999 report continues to be relevant in the present context, and has been incorporated into the assessments of the project group. The argumentation underlying the classification into resource types, for example, remains useful here. The project group has found no reason to reiterate the discussion, but refers readers to the 1999 report for further details (report is available in Norwegian only).

A large-scale project has recently been carried out in The Netherlands and Belgium to specify the types of data and basic technology applications that are essential to an all-purpose HLT resource collection for Dutch. These specifications (referred to as BLARK[4]) can be generalized to apply to other languages, and have been employed as a starting point for the compilation of new language resources for French. The Dutch specification corresponds closely to the recommendations contained in the present report, as well as those found in the 1999 Norwegian report.

The content of a national HLT resource collection should satisfy the fundamental needs of HLT-related research and development, and should facilitate cost reductions in relation to the development and adaptation of Norwegian HLT products and services. The project group has evaluated the proposals recommended in 1999 in relation to the projected needs today and in the future. There is consensus within the group that the main components of the previous

[4] Basic Language Resources Kit

proposal can remain as they are, but that some modification and additional elements are needed.

The projects group's estimates of the resources needed are somewhat higher than those in the Svendsen report. This is due to the addition of the following extra parameters:

- A much higher volume of spontaneous speech – this contributes greatly to the increase in the speech data figures.
- Multilingual texts have been added to support activities relating to machine translation.
- More emphasis is placed on coordinating lexical data – many sources will require intensive harmonization activity.
- A plan to establish databases for concept descriptions and semantic networks for Norwegian has been included.

4.2 Types of Linguistic Data:

The resource collection consists of three main components:

- Speech data (electronically-stored recordings of speech)
- Text data (collections of text with and without mark-up)
- Lexical data (collections of general language vocabulary and language for special purposes, e.g. terminology lists)

These components correspond to the proposals of the Svendsen report and the data compiled for other languages.

Speech data is used in speech recognition (speech-to-text, speech-to-phonetic transcription, speech-to-concept) and speech synthesis (machine-generated speech). A distinction is usually made between telephony data recorded over telephone lines and data recorded in an office environment, as each has specific and different quality and noise sources. These two data types are seldom interchanged.

Text data is necessary both in order to generate language models, for example for speech recognition purposes, and to analyze how language is actually used. Most human language technology applications, such as speech recognition, spell checks, grammar checks and translation software, must be based on a very large text volume in order to function satisfactorily.

Adequate lexical data is essential to all HLT applications. A relatively large volume of Norwegian language lexical resources has already been generated using public funding, either as basic allocations through the National Budget or via the Research Council of Norway. Altogether, these constitute a substantial material base.

4.3 Principles for Prioritization

The objectives of establishing an HLT resource collection are to:

- help to fulfil the objective of ensuring that Norwegian – spoken and written – remains the dominant language of use in Norwegian society;
- better utilize the potential for increased productivity inherent in human language technologies, not least within the public sector;

- make it possible to conduct HLT research utilizing Norwegian as the central empirical component;
- encourage the Norwegian ICT industry to invest in HLT solutions to prevent Norway from lagging behind international industry in this sphere.

As regards prioritization between the various types of resources to be incorporated into the collection, the following principles should be emphasized:

- The type of resource should be relevant to key areas of use.
- The specific resource does not currently exist, is not accessible, or is of sub-standard quality.
- The specific resource must be able to be compiled and structured for the resource collection in the course of a delimited time-frame.
- There is a concrete demand for the specific resource from HLT industry or research circles.
- The resource is vital to strategic research programmes that have been implemented.

The fundamental principle for prioritization is to ensure that the objectives above can be reached as rapidly and cost-effectively as possible.

Commercial and research-related activities relevant to the objectives above have been initiated in Norway. These include:

- dictation of electronic journals in the health care sector;
- automated telephone information services;
- proof-reading, grammar check and translation software;
- tools for persons with disabilities or problems with reading and writing.

There is budding commercial activity among Norwegian companies. However, the relevant language data available for such activities is substantially less than for English, and is also less than that available for Swedish and Danish.

The establishment of the HLT resource collection coincides with the launching of a long-term programme on knowledge development for Norwegian language technology (KUNSTI) by the Research Council of Norway. Thus, it would be natural to relate certain aspects of the resource collection's priorities to the needs of that research initiative. The Research Council's underlying documentation for the KUNSTI programme stipulates that projects under the auspices of KUNSTI must be based on the fact that language resources are available, which is not the case as regards speech data, text data or lexical data. Priorities concerning the content of the resource collection should be designed to support the needs of central research projects such as those emerging from the KUNSTI programme.

Furthermore, the database priorities must ensure that useful products found for other languages can be adapted to Norwegian within a reasonable time-frame. The details regarding identification of such products should be compiled in collaboration between the industry, user groups and researchers, and submitted to the resource collection agency board for review and prioritization. Activities in relation to dictation systems within the health care sector, for instance, are a natural candidate. Machine translation is another under-developed area in Norwegian HLT product development. It is useful to keep in mind the priorities utilized by

other countries in the consolidation of language resources. In this context the Dutch-language initiative is particularly interesting.

4.4 Minimum Content And Recommended Content

The project group has chosen to focus its efforts on the minimum volume of content that must be present in order for the database to be serviceable to the target groups. The recommended content would be larger, but it is difficult to calculate this precisely because there can never be too much data for training statistically based HLT products. In general one could say that the content of the resource collection should well exceed the minimum proposed by the project group. In the tables below the figures for minimum content have simply been doubled to determine the recommended content.

The presentation below is in no way meant to serve as an absolute. The resource collection agency board will naturally be expected to adapt the content to changing needs over time.

4.4.1 Speech

Speech data form the core of all technology involving recognition and production (synthetic or artificial) of speech. Speech recognition technology requires recordings from many speakers representing different age groups and different dialects. The recordings should be linked to realistic user situations to the greatest extent possible. This applies, for example, to speech recognition in automobiles and via mobile telephones, recognition of spontaneous speech, dictation by health personnel, legal personnel, etc. Recordings of this type are costly to obtain. Developers must often utilize data compiled from controlled situations to train a first-generation application. Once this has been achieved, the application itself (e.g. a traffic route information hotline) can be used to compile further data (assuming that the appropriate information has been provided to informants in compliance with applicable legal guidelines).

Development of an application for artificial speech requires input of large amounts of text incorporating a planned vocabulary, read aloud by a single speaker using natural prosody. The vocabulary should as far as possible encompass words and phrases likely to appear in the texts that the machine is being designed to read.

A minimum input would comprise a collection of digitalized speech recordings corresponding to 1,700 spoken hours (close to 17 million words), distributed between reading and spontaneous speech. Read speech from a manuscript would form the basis of this collection, and would supply the variation in material needed for the basic speech technology. Representation of various types of noise is also important, but this can be at least partially compensated for using simulation. Spontaneous speech would provide the dominant input for use in speech technology. Wide-ranging representation of this type of speech is therefore essential. Prosodic tagging of natural, spontaneous speech is extremely useful for improving the quality of synthetic speech as well as the quality of the next generation of speech recognition technology. A large volume of transcribed spontaneous speech would form a good foundation for enhanced modelling of structural phenomena in this type of speech. Similarly, speech from human dialogue is important in the context of training speech recognizers for dialogue systems.

Approximately 900 hours should be read speech from manuscripts, as this will form the basis for acoustic models for speech recognition. The distribution between *Bokmål* and *Nynorsk* should be equal to ensure the quality of the speech technology models. This part will also include speech data for development and experimentation with synthetic speech. The rest

should comprise various types of spontaneous speech: dictation, human-machine dialogues, human-human dialogues and conversations between several people. The material must be divided into high-quality recordings and recordings from fixed and mobile telephones. Representative coverage of different dialects, age groups, sociolects and genders is crucial to the value of the database for users. These recordings should at the very least be marked up (tagged) at the orthographic level, while a smaller portion must be tagged at a detailed phonetic and linguistic level.

In addition, it would be desirable to have collections that contain a substantial speech component, such as:

- multimodal corpora, i.e. databases containing speech and data from other modalities such as pointing, nodding, keystrokes, etc.;
- multilingual speech databases that can be utilized to find connections between various spoken languages;
- multimedia corpora that, in addition to speech from radio and television broadcasts, also contain information from other media such as texts and figures from the Internet, newspapers, magazines, etc.

In this context, the project group has chosen to give priority to data that is necessary to ongoing or planned technological development.

Table 4.1: Speech Data, Needs (modified from Svendsen 1999):

			Minimum			**Recommended**		
Type	Speaking style	Purpose	Hours of speech, minimum	Man-years, researchers	Man-years, others	Hours of speech, rec.	Man-years researchers	Man-years, others
Quiet room	Spontaneous	Dictation, dialogues	500	6.25	18.75	1000	12.50	37.50
Quiet room	Manuscript	Dictation, models	500	4.17	12.50	1000	8.33	25.00
Telephone	Manuscript	Models	120	0.60	1.80	240	1.20	3.60
Mobile phone	Manuscript	Models	120	0.75	2.25	240	1.50	4.50
Telephone in car	Manuscript	Multipurpose	120	3.00	9.00	240	6.00	18.00
Telephone	Spontaneous	Dialogues	100	1.25	3.75	200	2.50	7.50
Telephone	Spontaneous	Dictation	100	1.25	3.75	200	2.50	7.50
Quiet room	Manuscript	Diphone database	2	1.00		4	2.00	
Quiet room	Manuscript	Prosody / speech corpus	20	1.00	0.50	40	2.00	1.00
Telephone	Manuscript	Topic detection in multimedia databases	20	1.00	0.50	40	2.00	1.00
Audio	Spontaneous	Topic detection	100	1.25	3.75	200	2.50	7.50
Quiet room	Spontaneous	Multimodal user interfaces				100	1.25	3.75
Quiet room	Spontaneous	Models, multilingual applications, transcription				100	1.25	3.75
Sound-proof room	Manuscript	Concatenative speech synthesis	40	0.25	0.75	80	0.50	1.50
TOTAL			**1742**	**21.77**	**57.30**	**3684**	**46.03**	**122.1**

4.4.2 Text

A major portion of the text material must consist of Norwegian texts that are automatically tagged for word class – a minimum of close to 100 million words each for *Bokmål* and *Nynorsk*. The texts should comprise non-fiction prose, miscellaneous small publications, unpublished texts, newspapers and other printed media in addition to works of fiction. A small portion should be devoted to data for training and validation of statistical language analysis programs that provide a relatively shallow text analysis. Comprehensive text databases serve as a primary source for developing lexical resources and statistical language models for speech recognition. The structure of the text collections must take this into account. The proposed size of the text base is an absolute minimum.

The text databases should be tagged in accordance with the recommendations of the *Text Encoding Initiative* (TEI). This is in keeping with the proposal in the 1999 Svendsen report.

In addition there is a need for multilingual parallel corpora. Such collections are crucial for machine translation purposes, regardless of whether these involve translation between *Bokmål* and *Nynorsk* or between Norwegian and a foreign language. Parallel corpora also provide a source of information for the construction of semantically structured lexical databases, which are valuable for activities such as information searches and text summarizing. At the very least, a parallel corpus for Norwegian-English should be included in the Norwegian HLT resource collection, and it would be preferable if other language pairs were represented as well. Most of this material should be structured so that the original text and translated text are linked sentence by sentence, with a smaller portion linked word for word.

Above and beyond this, there will be a need for tree banks for storage of correct sentence structures. Material of this nature is necessary for activities such as the development of statistical models for syntactic analysis, for example to provide training data for programs that are designed to learn grammatical structures from text and apply this knowledge in parsing. This computer-assisted language learning (CALL) approach is especially important in creating tools for (semi-)automatic annotation of corpora.

A special collection of data for training within medical dictation has also been included, as this area entails a great potential for enhanced efficiency in the public sector.

Table 4.2: Text Data, Needs

		Minimum			Recommended		
Text types	Encoding, processing, purpose	Size	Man-years, researchers	Man-years, others	Size	Man-years, researchers	Man-years, others
Bilingual texts (Norwegian - English)	Basic preparation	2 500 000	0.71	2.14	5 000 000	1.43	4.29
Bilingual texts (English - Norwegian)	Basic preparation	2 500 000	0.71	2.14	5 000 000	1.43	4.29
Bilingual texts (English - Norwegian and Norwegian - English)	Extended preparation	500 000	0.71	2.14	1 000 000	1.43	4.29
Non-fiction, miscellaneous small publications, unpublished texts	Basic preparation	50 000 000	1.43	4.29	100 000 000	2.86	8.57
Newspapers, media, fiction, etc.	Basic preparation	50 000 000	1.43	4.29	100 000 000	2.86	8.57
Non-fiction, miscellaneous small publications, unpublished texts	Extended text encoding, manually controlled part-of-speech tagging	500 000	0.24	0.71	1 000 000	0.48	1.43
Newspapers, media, fiction, etc.	Extended text encoding, manually controlled part-of-speech tagging	500 000	0.24	0.71	1 000 000	0.48	1.43
Newspapers	Tree bank	200 000	0.29	0.86	1 000 000	1.43	4.29
Anonymized medical journal texts	Training data for medical dictation	0			200 000	1.00	3.00
TOTAL			11.52	34.57		26.76	80.29

4.4.3 Lexical Resources

This part of the resource collection will contain lexicons and thesauri. In this context, a lexicon is understood to be an electronic word list with information on the vocabulary of a language at various linguistic levels. Thesauri are lexicons containing semantic and associative relations between the words. This also encompasses subject-defined terminology resources (e.g. for discipline-specific terminology).

Lexical resources will comprise general dictionaries (general language vocabulary, terms, names, etc.) with the ability to generate inflected forms, pronunciation lexicons, dialect descriptions, spelling descriptions and subject-defined thesauri (list of synonyms and discipline-specific semantic dictionaries).

Comprehensive dictionaries for Norwegian (*Bokmål* and *Nynorsk*) have been developed at the universities. These contain lemmatizers, inflection generators, phonetic descriptors and lists of inflection forms. Nordisk språkteknologi has collected material totalling 1.5 million words.

A large amount of applicable material has already been collected and can be utilized. The minimum need for lexical data (dictionaries) is 500 000 words each for *Bokmål* and *Nynorsk*. All material entered into the resource collection must be checked for quality and standardized in terms of grammatical information, orthographic conventions and pronunciation standards. More detailed specifications regarding the pronunciation of names, foreign-language words and neologisms must be prepared and must be adapted to the individual dialect areas.

Industrial and research circles have clearly indicated that they need a Norwegian version of the English "Wordnet," which has existed for approximately ten years. This type of resource can be used in applications such as information retrieval and translation programs, and it has therefore been included as one of the resources the collection should be able to provide. Finally, the project group has included a concept description in which a conceptual database for Norwegian is linked to a corresponding database for English. This is a Norwegian version of existing EU resources.

Table 4.3: Lexical Resources, Needs

	Minimum				Recommended			
Activity	No. of words (full forms)	Man-years, re-search.	Man-years, others	Acquisi-tion costs	No. of words (full forms	Man-years, re-search.	Man-years, others	Acquisi-tion costs
Word lists, *Bokmål*	500 000			1 666 667	1 000 000			3 333 333
Word lists, *Nynorsk*	500 000			1 666 667	1 000 000			3 333 333
Incorporation of word lists from various sources		2.00	1.00			2.00	1.00	
Spelling variants / basic dialect variants		1.00	1.00			2.00	2.00	
Pronunciations for names, foreign words and new words		1.00	3.00			0.50	3.00	
Quality control of existing word lists (*Bokmål* and *Nynorsk*)		2.00				2.00		
Pronunciations for dialect regions		1.00	3.00			1.00	4.00	
Wordnet (Norwegian)	50000	0.71	2.14		100 000	1.00	3.00	
Concept descriptions – SIMPLE	50000	0.71	2.14		100 000	1.00	3.00	
TOTAL		8.42	12.28	3 333 334		9.50	16.00	6 666 666

4.5 Standards

To the greatest extent possible, the material compiled must be adapted to EU standards such as the Expert Advisory Group for Language Engineering Standards (EAGLES) and TEI. It is an absolute requirement that all existing international standards be stringently applied to all material collected for the resource collection. In the event that applicable standards are not available, the collecting agency shall conform to the best practice utilized internationally (cf. ELDA/ELRA).

4.6 Tools

Tools that are developed or acquired in connection with the collection and refinement of data for the HLT resource collection must also be viewed as resources and made available to other users. This may involve software for reading and recording speech, transcription and annotation, analysis, conversion between different data formats, etc.

The taggers for automatic text mark-up developed at the University of Oslo in cooperation with the HIT Centre in Bergen will be made available to the resource collection agency. These taggers have a close to 95 per cent level of accuracy, which is satisfactory for automatic text mark-up.

CHAPTER 5 COST ESTIMATES

5.1 Introduction

Most of the costs related to compiling the basic resources that should be included in a Norwegian HLT resource collection are linked to the amount of labour involved. It is possible to purchase the rights to existing collections of data, but in many cases the acquisition price may be nearly equal to the cost of new resource production. For this reason, the project group has based its cost estimates on the amount of labour required. This also provides a means of measuring the worth of existing material. The project group has not reviewed how much of the existing material can be used, nor how much must be compiled from scratch to satisfy the needs of the resource collection. This will be a task for the strategic planners of the resource collection when that time comes.

To the degree possible, the amount of labour needed has been defined in terms of man-years. For text and lexical data, the total man-years have been broken down into number of words per hour, while for speech data they have been broken down into hours of speech per man-year. Otherwise, the cost estimates incorporate the following elements:

- 25 % of all labour must be led by a researcher or other qualified personnel with adequate experience and expertise in the relevant area.
- Validation of all data is included in the model.
- Personnel costs are based on the Norwegian State Industrial and Regional Development Fund's salary estimates for highly qualified personnel and lower-level executive officers.
- All equipment and relevant tools are presumed covered within the framework of the man-hours.

It has been difficult to arrive at exact estimates. The figures presented in this report are based on those of the previous report, the experience of the project group members and comparisons with similar projects in other countries. Few sources have wished to be cited as regards their precise figures, which has been problematic for the project group. The group has nonetheless done its utmost to reach realistic estimates. In most cases, the estimates presented here are based on the direct experience of the members of the project group in relation to compilation activities at NST, Telenor, the University of Oslo and NTNU.

Table 5.1: Parameter Values for Calculating Costs

Salary, researcher per year (parameter values)	800 000
Salary, assistant per year (parameter values)	500 000
Number of hours of manuscript speech annotation per man-year	30
Number of hours of spontaneous speech annotation per man-year	20
Number of words per hour in basic prepared text corpora	5 000
Number of words per hour in bilingual corpora, one direction	500
Number of words per hour in manually controlled POS corpora	300
Exact word by word alignment per hour in bilingual corpora	100
Number of hours recorded speech, fixed telephone, per man-year	50

Number of hours recorded speech, mobile telephone, per man-year	40
Number of hours recorded speech in car per man-year	10
Concatenative speech synthesis: hours read per man-year	40
Lexical data - number of words transcribed and grammar-controlled per hour	100
Tree banks - words per hour	100
SIMPLE, Wordnet – number of sense items per hour (="words/concepts")	10

5.2 Speech Data

The cost of compiling the speech portion of the Dutch corpus is calculated to be 5 million Euro, which is approximately NOK 38 million. This corpus encompasses 1 000 hours of speech, which is somewhat less than has been recommended for Norwegian. The reason for this difference is twofold: first, in the case of Norwegian it is necessary to compile speech read from manuscripts in both *Bokmål* and *Nynorsk*, and second, a certain amount of speech data has been previously compiled in The Netherlands. A large portion of the activity is being carried out under the auspices of universities in The Netherlands and in Belgium, which also helps to keep costs to a minimum.

The Svendsen report from 1999 provided cost estimates for compilation of data resources. These estimates correspond closely to the findings of Nordisk språkteknologi in connection with its internal compilation of material. Compilation activities in Norway can employ tools and know-how, e.g. from The Netherlands, which may help to reduce the amount of labour needed, and thus the cost per hour, for speech data. Therefore, the project group has based its estimate on the costs associated with the Dutch speech corpus. The speech portion of the Norwegian HLT resource collection is estimated to require slightly in excess of 57 man-years.

Table 5.2: Speech Data, Costs

Type	Speaking style	Purpose	Expenditures, minimum	Expenditures, recommended
Quiet room	Spontaneous	Dictation, dialogues	14 375 000	28 750 000
Quiet room	Manuscript	Dictation, models	9 583 333	19 166 667
Telephone	Manuscript	Models	1 380 000	2 760 000
Mobile phone	Manuscript	Models	1 725 000	3 450 000
Telephone in car	Manuscript	Multipurpose	6 900 000	13 800 000
Telephone	Spontaneous	Dialogues	2 875 000	5 750 000
Telephone	Spontaneous	Dictation	2 875 000	5 750 000
Quiet room	Manuscript	Diphone database	800 000	1 600 000
Quiet room	Manuscript	Prosody/speech corpus	1 050 000	2 100 000
Telephone	Manuscript	Topic detection in multimedia databases	1 050 000	2 100 000
Audio	Spontaneous	Topic detection	2 875 000	5 750 000
Quiet room	Spontaneous	Multimodal user interfaces	0	2 875 000
Quiet room	Spontaneous	Models, multilingual applications, transcription	0	2 875 000
Sound-proof room	Manuscript	Concatenative speech synthesis	575 000	1 150 000
TOTAL			**46 063 333**	**97 876 667**

5.3 Text Data

In the previous report, the compilation of text data was estimated to require close to 40 man-years. This figure must be assumed to be somewhat higher if one or more parallel corpora are added to the collection, so the estimate is now 45 man-years. This is reflected in the number of man-years in the final row of figures above the total. Note that no resources have been allocated for purchase of the right to use or compensation for texts included in the database. The total costs will rise if such compensation is necessary. See also section 7.13.

Table 5.3: Text Data, Costs

Text types	Encoding, processing, purpose	Expenditures, minimum (*Bokmål* and *Nynorsk*)	Expenditures, recommended (*Bokmål* and *Nynorsk*)
Bilingual texts (Norwegian - English)	Basic preparation	3 285 714	6 571 429
Bilingual texts (English - Norwegian)	Basic preparation	3 285 714	6 571 429
Bilingual texts (English - Norwegian and Norwegian - English)	Extended preparation	3 285 714	6 571 429
Non-fiction, miscellaneous small publications, unpublished texts	Basic preparation	6 571 429	13 142 857
Newspapers, media, fiction	Basic preparation	6 571 429	13 142 857
Non-fiction, miscellaneous small publications, unpublished texts	Extended text encoding, manually controlled part-of-speech tagging	1 095 238	2 190 476
Newspapers, media, fiction	Extended text encoding, manually controlled part-of-speech tagging	1 095 238	2 190 476
Newspapers	Tree bank	1 314 286	6 571 429
Anonymized medical journal texts	Training data for medical dictation	0	4 600 000
TOTAL		**26 504 762**	**61 552 382**

5.4 Lexical Resources

The discussion of the lexical resource database in the 1999 Svendsen report was based on the assumption that much of the basic resources were already in place for *Bokmål* and *Nynorsk* alike. Most of the activity for this portion of the resource collection was expected to comprise adding spelling variants, pronunciation and devising pronunciation descriptions for names. Semantic thesauri and subject-defined terminology resources (such as synonym word lists within various fields) were not given priority. Multilingual lexicons were not included.

The previous report estimated that 10 man-years would be needed for the lexical resources. Since the project group has proposed that the content in this part of the collection be expanded, the 1999 estimate has been adjusted up to 21 man-years.

Table 5.4: Lexical Data, Costs

Activity	Expenditures, minimum	Expenditures, recommended
Procurement, word lists, *Bokmål*	1 666 667	3 333 333
Procurement, word lists, *Nynorsk*	1 666 667	3 333 333
Incorporation of word lists from various sources	2 100 000	2 100 000
Spelling variants / basic dialect variants	1 300 000	2 600 000
Pronunciations for names, foreign words and new words	2 300 000	1 900 000
Quality control of existing word lists (*Bokmål* and *Nynorsk*)	1 600 000	1 600 000
Pronunciations for dialect regions	2 300 000	2 800 000
Wordnet (Norwegian)	1 642 857	2 300 000
Concept descriptions - SIMPLE	1 642 857	2 300 000
TOTAL	16 219 048	22 266 666

5.5 Administrative Costs

The administrative costs associated with the compilation efforts are based on the experience of NST and SPNE as well as general practice from other distribution activities. The costs are highest during the first year, which is when major planning activities take place and a framework for tenders is drawn up. After that, the administrative costs are expected to decline to approximately NOK 1.5 million annually.

Table 5.5: Administrative Costs

Administrative Costs

	Year 1		**Year 2**		**Year 3**		**Year 4**		**Year 5**		
Tasks	**Man-years**	**Expendi-tures**	**Man-years**	**Expendi-tures**	**Man-years**	**Expendi-tures**	**Man-years**	**Expendi-tures**	**Man-years**	**Expendi-tures**	**Total**
Adm.manager, foundation	1.0	800 000	1.0	800 000	1.0	800 000	1.0	800 000	1.0	800 000	4 000 000
Operational costs, foundation, infrastructure		250 000		250 000		150 000		150 000		150 000	950 000
International contacts, foundation		75 000		75 000		75 000		75 000		75 000	375 000
Legal assistance, foundation		200 000		50 000		50 000		50 000		50 000	400 000
Adm. manager, operational enterprise	1.0	800 000	1.0	800 000	1.0	800 000	1.0	800 000	1.0	800 000	4 000 000
Executive officer, operational enterprise	1.0	500 000	1.0	500 000	1.0	500 000	1.0	500 000	1.0	500 000	2 500 000
SUM		**2 625 000**		**2 475 000**		**2 375 000**		**2 375 000**		**2 375 000**	**12 225 000**

5.6 Overall Outlay

Overall, activities relating to the compilation and configuration of data will require some 150 man-years. Approximately one-third of this will consist of highly qualified personnel, while people with less expertise can be used for the remainder. The basis for personnel expenditures is an average cost of NOK 800 000 per man-year for the highest qualified staff, and NOK 500 000 per man-year for the assisting personnel. Given 150 man-years, the total outlay will run to nearly NOK 90 million. In addition come expenditures for administration, infrastructure,

informant remuneration, travel expenses, etc., totalling approximately NOK 10 million. Thus, the total cost of establishing a Norwegian HLT resource collection will be close to NOK 100 million.

The figure below shows the distribution of the costs related to the various categories.

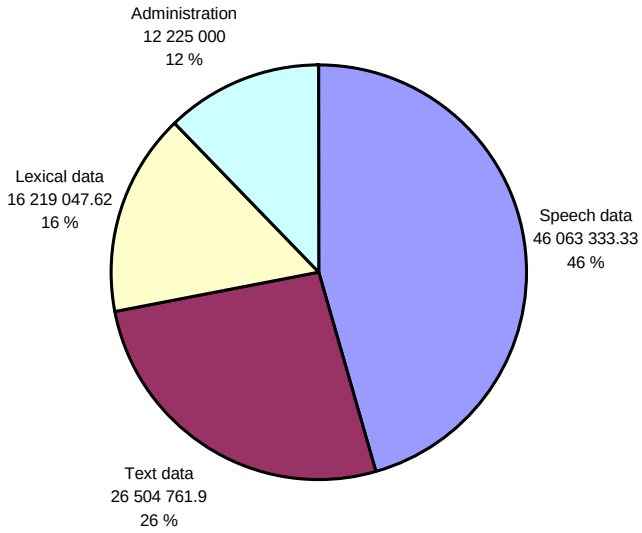


Figure 3: Costs distributed by category

CHAPTER 6 FINANCING

Realistically speaking, this endeavour must be supported by public funding. The Norwegian HLT industry does not have the ability to finance these activities to any great degree. Public funding is being provided for compilation of resources in European countries and among language communities that are larger than Norway, despite the fact that these countries or language communities often have a stronger HLT industry and a greater volume of previously compiled resources.

6.1 Introduction

The mandate of the project group includes the task of specifying a financial plan outlining all costs associated with establishing, operating and refining the resource collection, from minimum to recommended size, based on combined private and public funding schemes.

Cost estimates have been dealt with in Chapter 5.

This chapter discusses the principles for, prerequisites to and potential sources of funding for a Norwegian HLT resource collection. The discussion emphasizes which principles and prerequisites are crucial to financing of the initial phase, i.e. the compilation and configuration activities needed to generate a resource collection in conformance with the specified minimum requirements. Additionally, the project group has assessed matters pertaining to funding of the maintenance, operation and further refinement of the resource collection.

6.2 Prerequisites and Principles for Funding

The creation of a Norwegian HLT resource collection is a means of establishing a national infrastructure to serve cultural, socio-economic and industrial purposes. Examples from other European countries indicate that such infrastructure should and must be primarily supported by public funding, since the endeavour involves compiling and managing resources that constitute a shared national asset with obvious socio-economic benefits if they are administered under a cohesive framework.

Since the number of speakers of Norwegian is low compared to other countries in Europe, a Norwegian-language HLT resource collection carries particular significance within the cultural policy sphere. The market in Norway is so small that it would not be feasible, economically or otherwise, for individual players in the public or private sector to finance the compilation and configuration of language resources for anything other than specifically defined and clearly delimited purposes.

The work of the project group has revealed that this is an area in which great advantage is to be gained from large-scale efforts – HLT resource collections need to have roughly the same volume regardless of the geographical range of the language or its number of users. These assessments have led the project group to conclude that the establishment of a Norwegian HLT resource collection is dependent on earmarked funding of infrastructure investment outlay through the national budget. This conclusion is reinforced by the fact that Norway’s plans to compile a national HLT resource collection are emerging rather late in the day. Time is truly of the essence, and implementation based on private funding would be unrealistic and only lead to further delay.

The arguments advocating the establishment of a Norwegian HLT resource collection indicate that the main responsibility for coverage of the infrastructure investment outlay lies with three ministries: The Ministry of Culture and Church Affairs (MCC), the Ministry of Trade and Industry (MTI) and the Ministry of Education and Research (MER). The MCC is responsible for maintaining and strengthening Norwegian language and culture, and as such is naturally involved vis-à-vis the cultural policy sphere. The MTI has a central part to play in respect of its responsibilities regarding establishment of new industrial activity, industrial development and its role as coordinator for Norwegian ICT policy. Human language technology is one of several relevant areas that has been designated as a significant ICT policy instrument, as well as an industry in which product development can lead to enormous efficiency benefits, particularly for the public sector. This is clearly illustrated in the eNorway 2005 strategy plan, which incorporates HLT activities as part of a clear strategy for Norwegian digital content.

A language resource collection of this nature would be invaluable as part of the infrastructure for linguistic and HLT-related research, which makes the MER a third candidate to participate in basic investment financing.

The cultural-policy, ICT and industrial-policy and research-policy dimensions of this infrastructure should be reflected in the relative distribution of financing responsibility. In the opinion of the project group, a reasonable distribution would be 3 : 3 : 2 between the MCC, MTI and MER. It is essential that the three ministries are made jointly responsible for procuring the necessary funding. Given the purpose and nature of the material designated for use in a Norwegian HLT resource collection, it would not only be unnatural but also unproductive to attempt to assign any one of these three ministries funding responsibilities for a given portion of the resources.

Based on the assumption that the resource collection agency would be established as an independent organization, there are practical considerations that speak in favour of channelling infrastructure investments from the ministries involved directly to that organization. Alternatively, the ministries could channel funds as earmarked (ad hoc) allocations over subordinate agencies that have been asked to safeguard significant use-related and user interests. The Norwegian Language Council (vis-à-vis MCC) and the Research Council of Norway (vis-à-vis MTI and MER) comprise the obvious choices. This arrangement would give these institutions a clearer academic and operative role in the establishment of Norwegian HLT resources. The creation of a Norwegian HLT resource collection would provide both of these institutions with better opportunity and reason to commit more strongly to tasks that could benefit from Norwegian HLT resources. Funding to establish the resource collection must therefore be allocated specifically, not as earmarked funding on the ordinary budget. The latter alternative would exert pressure on the capacity of these institutions to finance their daily activities, thus diminishing their ability to participate actively in the effort to make the best possible use of the emerging infrastructure.

As the areas of application of language technology continue to expand, most ministries will find themselves responsible for activities that represent relevant user interests for a Norwegian HLT resource collection. This becomes even more significant in the light of increased efficiency of public administration and modernization of public services. As is the case with private actors, however, it is not reasonable to assume that user interests in relation to a national infrastructure can be linked to participation in funding of the necessary investments to set up the infrastructure required. An overly complex financing framework would be highly impractical. It would only create confusion and decrease the viability of

safeguards designed to ensure that funding for infrastructure investment is in place. It could also lead to unnecessary extra administrative activity in relation to establishing the resource collection.

Although the project group is recommending that financing of the infrastructure investment outlay for the resource collection be channelled through a limited number of ministerial funding sources, it is assumed that other private and public actors will provide substantial contributions to the (basic) resource collection. These will entail assisting in and contributing to the academic and operative tasks involved in compiling the collection, partly by making language resources available (according to agreed-upon terms), and partly by entering into cooperation with the resource collection agency on institution-initiated compilation of language resources and (project-)specific tasks. To ensure that the organization and priorities laid down for the establishment process remain coherent, it would be productive to expect this type of contribution from the most substantial ownership and user interests. In realistic terms, the amount of direct financial contributions towards the infrastructure investment will be minimal, and such funding could well result in more ad hoc establishment of the resource collection than desirable, with less planning and priorities based on specific situations instead of long-term goals. Mobilizing key user interests in the compilation of the collection will provide useful input regarding cooperation models for its day-to-day operations and maintenance.

The project group considers Norway's public financing institutions for industrial development to be key players as regards funding R & D projects based on utilization of the type of language resources contained in the collection. However, it is the view of the project group that these institutions do not, and should not, have a natural role to play in funding the establishment this type of national infrastructure. Instead, their participation should in this context be directed towards strengthening the financial foundation for industrial activities aimed at creating Norwegian HLT products and services.

6.3 Funding Alternatives

The cost estimates underlying the funding model outlined above will require allocations of roughly NOK 100 million across the national budget over a five-year period, i.e. approximately NOK 20 million annually. The actual distribution of funding between the ministries is of less practical importance, and the proportional distribution is open to discussion between the parties involved. However, to ensure an effective process, a cohesive plan for the establishment phase of the resource collection must be devised. It is crucial to this process that the relevant ministries participate in the overall funding package outlined in the recommendations of the project group.

The project group believes that a public funding guarantee would provide the best foundation for cooperation with key public and private actors who, in turn, can each help to ensure that the resource collection is compiled successfully and cost-effectively. Large- and small-scale language resource collections that may be eligible for incorporation into the collection are found at a number of public research and private institutions. Many of the potential participants have expressed a clear willingness to cooperate on this as well as future compilation of resources. This will make it possible to generate the resource collection at a substantially lower price than would be the case if the entire production process were contracted out as independent projects.

A brief look at the potential contribution of the University of Oslo can help to illustrate this. The university has wide-ranging experience with the preparation of electronic texts. The relevant institutes possess the academic expertise needed and have an interest in participating in the effort to compile a Norwegian HLT resource collection. The university also has a great deal of experience in organizing and conducting large-scale projects. The administration of the Faculty of Arts views participation in a national project in a very positive light, and has stated that it will be possible to provide personnel resources as a means of partial financing to any major sub-projects assigned to the faculty.

The same applies to the University of Bergen, where work has been carried out in human language technologies since the early 1970s. The Faculty of Arts there also has staff with top-notch expertise and in-depth insight into the organization of contract research projects as well as other externally funded R & D activities.

In 2002, the Research Council of Norway launched a long-term research programme in language technology, called KUNSTI. This initiative has sprung out of the conviction that language technology is rapidly emerging as such an important field that a Norwegian HLT resource collection must be established. The research programme is aimed at strengthening the national research community in order to ensure the optimal utilization of the potential for Norwegian HLT research and development inherent in a resource collection. The KUNSTI programme has a wide-ranging researcher interface, and will play an active role in achieving broad-based national cooperation on academic as well as operational aspects of the resource collection project.

The establishment of a Norwegian HLT resource collection represents the development of an infrastructure for research and industry alike. In many of the larger European countries, public and EU-based funding has been used to finance the necessary infrastructure investment outlay for simple HLT products, while hi-tech companies themselves have carried the costs associated with product development. In some cases, the priorities regarding what type of material to compile first have been altered after the industry has provided funding for specific projects.

Currently, the Norwegian HLT industry is very small compared to international companies such as Nuance, SpeechWorks, the language technology division at Philips, IBM, Siemens, etc. The largest Norwegian player is Nordisk språkteknologi (NST) in Voss. At Telenor R&D there is a group of researchers who have been working with speech technology for over 20 years. In addition, there are a few small companies scattered throughout the country. These cannot be expected to have the financial means to fund any major portion of the resources recommended in this report. What they do need, however, is the content, which would enable them to develop Norwegian-language products. Moreover, a Norwegian HLT resource collection would make it possible for Norwegian industry to create new products that may then be adapted to other languages. As with other industrial sectors in general, it is important in this context to initiate activities on the domestic market before expanding internationally. A Norwegian HLT resource collection will be attractive to foreign producers, such as those mentioned above, who can use it to generate Norwegian-language products, e.g. dictation and machine translation tools. Once the resource collection is available, international companies will be able to utilize the resources to improve or create new speech recognition products for the Norwegian market. This may lead to contracts for Norwegian industry, for instance in connection with systems integration of speech technology in new products and services.

It is important to note that EU countries such as France, Italy, The Netherlands and Belgium are now creating an opportunity for a targeted public effort to establish an HLT infrastructure that can, among other things, withstand the pressure from English. Germany is considering implementing similar initiatives. These efforts are taking place in countries *that from the outset have a much better developed HLT infrastructure*, and that represent a much larger language community than Norway. The language technology industry of these countries is no better equipped to carry the costs of this infrastructure than the Norwegian industry, so it is as a result of cultural and industrial policy priorities that large-scale public resources have been allocated in these countries for the creation of national language databases. A small language community like Norway cannot count on having a language industry with sufficient revenues to finance the cost of resource compilation over time. It is at the national level that Norway must invest in the infrastructure needed to allow the creation of HLT products and services for Norwegian featuring the same quality as products for other languages.

Once the resources have been compiled, they must be maintained, kept operational and further refined. The users, i.e. research and industry, will have to pay to utilize the resources, and user fees should in time be sufficient to cover some of the costs of operation. However, if prices are set too high, the resource collection will lose its appeal as a source for the development of Norwegian-language HLT products and services.

The project group has identified two alternatives for funding of the HLT resource collection:

a) Complete public funding
b) Public funding for the main activities, potential suppliers can submit data for compensation in the form of either access to other data or cash payment, and users pay a small fee to utilize the material in the collection.

There should be room for the industry to take part in determining priorities if it is willing to commit funds to the compilation process. In the event that industrial concerns offer to provide financial or input resources, special provisions for restrictions on use for a limited time period may be considered. The problem with co-financing of this type is that it makes ownership somewhat more complicated. Complicated ownership issues and restrictions on use may enable a specific actor to reduce or inhibit market competition, which would be unconstructive.

There must be an absolute requirement that all resources will be subject to quality validation by a neutral institution and made available to research and industry alike.

In the light of ongoing projects in other countries, approximately NOK 100 million will be required to compile a resource collection of a nature and quality that will be serviceable to a modern HLT industry and research community. These funds shall be employed to purchase the rights to use existing material of satisfactory quality, for example from Nordisk språkteknologi, to identify material or data collections that can be distributed, and to finance production of new data. Activities in connection with the purchase of rights and distribution will take the least amount of time, and should be given priority together with the new compilation of spontaneous speech. The need for international contacts, for example through participation in the EU *ENABLER* network, must also be given priority. Such participation will provide useful insights into activities in this field in Europe, and will make it possible to exploit international expertise by applying international standards and best practice. It would also enhance access to the use or localization of appropriate software that could facilitate the

compilation and distribution of language data. This may in turn lead to higher quality and lower costs.

CHAPTER 7 A PLAN FOR IMPLEMENTATION

The following plan for implementation is only an outline. The project group has not had time to discuss the individual proposals adequately with either the institutions involved or the relevant industrial organizations. Therefore, these proposals are preliminary, and must be viewed as a basis for possible models.

7.1 Time-frame

The compilation of resources should be initiated as soon as possible. This process will probably extend over a period of five years because a major portion of the material must be compiled from scratch, which is time-consuming. As regards funding, it is also an advantage to distribute costs over several years. This time-frame corresponds well with similar projects in other countries.

7.2 Establishment Costs and Costs to Users

Setting up the resource collection entails either compiling new material or purchasing the rights to use existing material, after which the resources will be made available to user groups within research and industrial circles. While user groups will be expected to pay to utilize the data, user fees will be far lower than the costs of establishing the resource collection (otherwise the whole point of this endeavour is lost). Fee structures for commercial users could be, for example, five per cent of the calculated costs of the material in the resource collection, while fees for research institutions should be roughly half this amount.

Rights to use the material can be purchased in several ways:

1. The supplier can receive payment upon submission of data to the resource collection agency.
2. The supplier can receive payment-in-value of resources for an amount based on the submission value of the data resources, i.e. there is a “trade” for more data than was submitted.
3. The supplier receives payment as data are utilized.
4. A payment scheme combining 1 and 2.
5. A payment scheme combining 1 and 3.

Alternatives 2, 3, 4 and 5 may provide a means of cost-sharing, cf. the table over cost distribution. A pricing structure of this type would enable those who submit data according to alternative 2 to get back data many times the value of the original submission, which will enhance the appeal of providing resources.

Issues pertaining to intellectual property rights and purchase of user rights to material must be considered by the resource collection agency board in each individual case. Some of the general problems have been examined in the legal report commissioned by the project group. See also section 7.13 below.

7.3 Existing Material

Information has been compiled regarding relevant institutions in possession of resources that may be of interest to the resource collection. An overview of actors and resources was provided in an attachment to the Norwegian version of this document. The overview clearly indicates that a great deal of existing material can be incorporated into the resource collection.

It should be noted that a large volume of relevant resources has already been compiled. While some of this can be incorporated into the database without further refinement, much of the material will need to be monitored for quality as well as reviewed to determine whether or not it can be used in the light of established intellectual property rights.

The value of existing material must be based on estimates and known costs for new production of similar resources with the tools of today. In cases where new production is not substantially more costly than the price of purchase, the material should be compiled from scratch, assuming there is a reasonable time-frame.

The relevant commissioning agencies must each take responsibility for assessing the degree to which resources funded over government budgets can be made available to the resource collection, given that conditions pertaining to intellectual property rights are satisfied. This applies in particular to material compiled at the universities. This point will be discussed in more detail later in this chapter.

7.3.1 Speech Data

The overview compiled of existing material (attachment to Norwegian version only) shows that a relatively large volume of manuscript-read speech is already available from Nordisk språkteknologi. An extensive collection of recordings from telephone speech is also available from NST and Telenor. There is virtually no spontaneous speech included in the existing collections. Assuming that agreement can be reached regarding compensation schemes, it should be possible to purchase the rights to read speech material and thereby rapidly obtain a good portion of the speech data needed in the resource collection. The compilation of spontaneous speech will take place over several years.

7.3.2 Text Data

It appears that a certain amount of text data will be available, but much of this material is reserved for use by researchers. New data compiled in this area should to the greatest extent possible be commercially available, and the text resources must achieve a better balance with regard to types of text. Newspapers and non-fiction texts are currently over-represented.

7.3.3 Lexical Data

The existing lexical data is generally acceptable, with a few exceptions. However, a great deal of effort will be required to refine and systematize the data, e.g. to standardize grammar mark-up and pronunciation information. Regardless, standardization activity of this type will require much less labour than compilation of new material with corresponding mark-up and generation of inflected forms.

As concerns pronunciation descriptions, the material must also be standardized and reviewed, with regard to annotation conventions (SAMPA, XSAMPA, etc.), tagging of syllable boundaries, accent, tonality and dialectal origin.

7.3.4 Recommendation

Whenever possible, use should be made of existing resources that demonstrate adequate quality. All matters pertaining to intellectual property rights must be clarified. All relevant data must be validated by neutral experts. Compensation schemes must be discussed with each individual supplier.

7.4 Resources Financed By Public Allocations

These resources comprise the most important sources at the University of Bergen (UiB), the University of Oslo (UiO), as well as at the University of Science and Technology in Trondheim (NTNU). In Bergen, the Centre for Humanities Information Technology (HIT Centre) has text and word list material available. The University of Oslo also has text and word list material of value to the resource collection. The following resources have been identified and given a value in keeping with the principles set out previously in this document (it is presumed that the data satisfy quality requirements adequately).

Table 7.1: Relevant Resources under University Ownership

Institution	*Type of resource*	*Total value per institution*
UiB:	Text, 0.5 mill.	0.5
NTNU:	Speech, lexical data: 0.8 mill	0.8
UiO:	Text: 1.9 mill., lexical data: 6.6 mill., speech data: 1 mill	9.5 mill. (does not include NorKompLeks[5] from NTNU due to overlapping, assuming that there are equal parts *Bokmål* and *Nynorsk*)

The Norwegian Board of Education under the Ministry of Research and Education has an audiobook database containing good quality digital read speech. The project group became aware of this resource too late in its efforts to assess the material in any detail, but such an assessment should certainly be carried out. The Board of Education also distributes DAISY-disks (DAISY=Digital Accessible Information System). This encompasses CD-ROM disks with up to 50 hours of audio. These are normally used as textbooks for the hearing and sight-impaired, and can be played on regular PCs with special equipment. DAISY is in the process of becoming an international standard, and the next version will be very close to a digital, multifunctional format that it is technically possible to provide over the Internet.

The resource collection agency should negotiate with the relevant suppliers, particularly the University of Oslo, regarding terms for incorporation of these language resources into the database.

7.5 Resources Financed by State-owned Enterprises

In this context, Telenor is the most relevant supplier. Telenor's resources encompass speech data and transcribed word lists (the estimates are unverified).

Table 7.2: Relevant resources owned by Telenor

Word lists	0.2 mill. (estimated 50 000 Norwegian words from Onomastica[6])
Speech data	2 mill. (probably more that cannot be quantified through existing information)

[5] Nowegian computational lexicon.

[6] An EU project involving the phonetic transcription of expected pronunciation of first names, surnames, place names, etc. in 11 European countries, totalling 8.5 million names.

It would be unrealistic to expect Telenor to offer its data without compensation, but the project group proposes that the government authorities try to find a solution that would serve the short-term needs of the resource collection, for example by offering Telenor access to other resources it might need, or by paying for the data over time as the resource collection enters into ordinary operations.

7.6 Resources Financed (Wholly/Partially) by the Research Council of Norway

This encompasses primarily resources that are currently being compiled under the auspices of the KUNSTI Research Programme because the necessary language resources are not otherwise available. Since many aspects of the programme had not been clarified during finalization of this document, no conclusions can be drawn regarding use of this data. However, any contracts that are signed should include provisions stating that compiled data must be turned over to the HLT resource collection. In addition come HLT tools developed with funding from the Research Council, for example software for automatic word class annotation at the University of Oslo and the HIT Centre. These tools can be utilized by the resource collection agency without compensation, but the labour costs of the annotation activity must be covered.

7.7 Resources Financed by the Public Funding Institutions for Industry

The data from Nordisk språkteknologi are the most relevant in this context. Nordisk språkteknologi has received, and continues to receive, a good deal of support from public funding sources, but has not received grants for compilation of the language data itself. The project group considers this data to be of value to the resource collection.

Nordisk språkteknologi cannot be expected to offer its data to the HLT resource collection for no compensation at all. Nonetheless, the government authorities should assess whether it would be reasonable to ask for a discount on the material from this institution.

Table 7.3: Relevant supplementary resources

Lexical data outside of the material from the Norwegian Word Bank at UiO	Approx. 1 mill.
Speech data	12.1 mill.

7.8 Other Data that Could Be Incorporated into the Resource Collection

Text material from BerlitzGlobalNet and Oracle may be of relevance to the resource collection. These actors have indicated their interest in other language resources as a form of compensation. The value of the material has been estimated at NOK 3.6 million. Publishing houses also have dictionaries and text material that could be useful. The project group has been in contact with Kunnskapsforlaget and Det Norske Samlaget, both of which are willing to help once the necessary agreements and contracts are in place.

7.9 Cost-sharing During the Compilation Process

The universities of Bergen and Oslo have both indicated their willingness to help with personnel resources during the compilation process. This is particularly important for the academic aspects of production of new data. The extent of the resources contributed by the universities will depend on which compilation projects are assigned to them, and the degree to which the necessary expertise is available.

The efforts of the University of Oslo should be focused on lexical resources. Signals from the university have not yet been clear enough to ascertain the actual size of their intended commitment, but the project group assumes that at least one position per year for four years will be allocated. This is predicated upon a task assignment that is targeted to the institution's expertise. The value for the resource collection will be NOK 2.5-3 million, and perhaps more if more personnel resources are made available.

The University of Bergen has indicated interest in activities involving text data compilation, and has stated that approximately 2 positions would be allocated for these efforts. Cooperation with the University of Oslo would be natural in this context. This means that roughly two positions at the level of researcher or project manager can be calculated in for the entire five-year compilation period. The University of Bergen has also stipulated that projects must be targeted for the available expertise. The value of these positions is between NOK 6.5 and 8 million.

Another type of cost-sharing for production of new material involves contributions in the wake of prioritized projects through the Norwegian Regional and Industrial Development Fund.

Some of the Development Fund's grant schemes are designed to strengthen the competitiveness of Norwegian industry both nationally and internationally by means of cooperation with a demanding public sector client. Grants are intended to promote better quality and/or reduce the costs of public services through access to new technology or new solutions.

Other grant schemes aim at encouraging R & D cooperation between client and supplier companies to develop new processes, methods or services that can be utilized by one or more companies. These schemes are intended to lead to competitively viable products with export potential, preferably in cooperation with a foreign client company.

The project being carried out in cooperation between Nordisk språkteknologi, St. Olav's Hospital and the Norwegian Regional and Industrial Development Fund in the field of medical dictation illustrates how crucial it is to have language data accessible. This project would not have been possible without the language resources compiled beforehand by Nordisk språkteknologi. The results of this project may form the basis for other dictation-related projects, new applications and tools. This data, together with access to even more language data from a national resource collection, could be utilized in other projects involving public agencies or private companies, assuming it were made available to the HLT resource collection.

For all projects funded by such schemes, the parties must agree that the data developed will be made available to the HLT resource collection at a price equal to the grant from the Development Fund, or special provision for this must be included in the contract.

It may be difficult to measure the actual value of this kind of contribution, but based on existing contracts, the project groups calculates a value of NOK 2 million per year for the duration of the time in which the Development Fund's grant schemes give priority to HLT activities.

7.10 Funding Models

The tables below contain proposed models for funding of the HLT resource collection, based on purchase of rights to use existing material. After language resources have been integrated into the HLT resource collection, they can be obtained by users for a maximum of ten percent of the original cost of incorporating them into the database. Ten percent of the actual procurement price is relatively high, and would make Norwegian resources considerably more expensive than ELRA resources. This applies particularly to applications that need a large volume of resources, for example dictation systems. As indicated in other parts of this report, there is no point in establishing a resource collection of this type if it is too expensive to use. In keeping with established ELRA/ELDA practice, research institutions should pay less than commercial actors to utilize the resources.

The *Total costs* column contains the estimated value of the data, divided between the three main categories speech data, text data and lexical data. *Validation* of the data is included in the total costs, cf. the chapter on the content of the resource collection. The *Existing data* column applies to data that is available and can in all likelihood be incorporated into the database, cf. the discussion on available resources (section 7.3). This data can be exchanged for access to other data as it is integrated into the database. However, if a supplier provides more resources to the database than they will need in exchange, some remuneration will probably be required to compensate for the excess resources. The *User fees* column represents the fees customers will be paying to utilize the data resources. It is realistic to assume that there will be a relatively large degree of overlap between those who provide data to and those who wish to acquire data from the resource collection, and there is reason to believe that most suppliers will prefer cash compensation. The potential for compiling existing data has therefore been reduced by 50 percent. The column for *Cost-sharing – new production* represents the stipulated contributions from the universities in connection with the compilation of new resources.

Table 7.4 presents an optimistic estimate:

Table 7.4: Cost Estimate I

Type	Total costs	Existing data	User fees	Cost-sharing – new production	Net outlay
Speech data	46 mill.	7 mill.	2 mill.		
Text data	30 mill.	3 mill.	2 mill.	8 mill.	
Lexical data	16 mill.	3 mill.	1 mill.	3 mill.	
Administration	7 mill.				
Total	99 mill.	13 mill.	5 mill.	11 mill.	70 mill.

The project group calculates that five percent of the value of the material in the database will come from fees. This figure will be achieved over time, possible over the entire duration of the compilation process. The government authorities and suppliers may need to wait for revenues to be generated. This has been the case in other resource collections, cf. ELRA. The project group has not taken a position as to whether revenues are generated as membership fees alone (cf. LDC) or as a combination of membership and purchase of relevant resources (cf. ELRA).

The realization of existing data for a value of NOK 13 million in an exchange model is an optimistic estimate, and this parameter is thus uncertain in the above model. A figure at half this level is probably more realistic.

In the view of the project group, this model can function effectively even despite a worst case scenario in which the price of the resources makes it impossible to develop some of the products for the Norwegian language. However, if the entire contents of the database are available free of charge it is unlikely that suppliers with relevant material will take the trouble to make their input available. A link between the value of the material supplied and the return on investment serves as an incentive for those providing input to the resource collection. A supplier who has provided input for a certain value will be eligible for a deduction when purchasing resources from the collection. An institution that provides resources at a value of NOK two million will receive resources for at least NOK 20 million in return.

Too high a price on the resources will prevent industry from investing in this sector and inhibit the development of HLT resources for Norwegian. The value of the database components will be realized over time, and will in practice provide the financial basis for the operation and maintenance of the resource collection agency. While the balance defined above may be somewhat over-optimistic, the table below incorporates some of the reservations mentioned here, and may be more realistic.

Table 7.5: Cost Estimate II

Type	Total costs	Existing data	User fees	Cost-sharing – new production	Net outlay
Speech data	46 mill.	3 mill.	1 mill.		
Text data	30 mill.	1 mill.	1 mill.	8 mill.	
Lexical data	16 mill.	1 mill.	1 mill.	3 mill.	
Administration	7 mill.				
Total	99 mill.	5 mill.	3 mill.	11 mill.	80 mill.

7.11 Budget

The preliminary budget that has been circulated in a prior report proposes that the total costs for the five-year compilation effort be distributed as following: 10% the first year, 30% the second year and 20% for each of the subsequent three years. This is based on the assumption that material that can be incorporated into the database “as is” should be validated and included during the first two years. All relevant material must be thoroughly evaluated. The evaluation process should be carried out during the first year of operation with the integration process initiated the following year.

All compilation activities include funding for researchers and assistants. Researchers will be responsible for designing, heading and validating compilation projects. The researchers in charge of quality control will not be the same as those involved in the other tasks, but both types of researcher responsibilities are included in the calculations for the compilation activities. This model assumes that research institutions, independent research institutes, relevant companies or foreign institutions (e.g. SPEX in The Netherlands) divide up the efforts to compile and validate the material (foreign institutions will only be permitted to participate in quality control efforts). Research activities are estimated to comprise an average of 25 % of the compilation costs.

As regards purchase of the rights to use existing material, this will be valued in relation to the costs of new production. Validation of data is included in the cost estimates.

7.12 Administration

7.12.1 Administration During the Initial Phase

The administrative activities during the first couple of years will require the greatest concentration of resources. It is during this phase that the existing material will be evaluated for possible acquisition (in technical, substantive and legal terms), detailed specification for compilation of new material will need to be prepared and the framework for tenders for the compilation projects will be organized.

7.12.2 Post-compilation Administration

As the collection of validated resources grows, the material can be distributed by, for instance, ELRA on behalf of the Norwegian resource collection agency. In the view of the project group, it would be preferable for either the collection agency itself or the proposed operational enterprise to be in charge of distribution activities within Norway. In its submission to the fiscal budget proposal for 2003, the Ministry of Trade and Industry states the following in the results report for 2001: “In 2001, SPNE and the municipality of Voss founded EDDA Språkressurser AS. This company is to be developed in the direction of a Nordic HLT resource collection, and will be able to provide fledgling companies under the SPNE incubator framework with useful services and expertise in the field of human language technologies.” This constellation should be incorporated into assessment efforts regarding the assignment of tasks to institutions within the existing circles of expertise.

The organizational model encompasses an operational enterprise that operates on behalf of the board of the resource collection agency foundation. This enterprise should be self-financed by means of fees obtained for all material delivered to users.

Permanent operational costs should be shared with existing departments at the universities. The University of Oslo is in the process of launching its Norwegian word bank, which could perhaps be given special responsibility for the operation and maintenance of lexical resources once the compilation period is completed. The HIT Centre at the University of Bergen has a great deal of experience in the distribution of text collections, and this expertise could be utilized by giving the Centre responsibility for maintenance of text resources, perhaps in cooperation with the Text Laboratory at the University of Oslo. Both universities have clearly signalled their interest in solutions of this type.

Speech data could be managed by either the University of Oslo or the University of Bergen.

7.13 Legal Deposit of Material

The project group assumes that most of the works to be incorporated into the resource collection are copyright-protected under intellectual property legislation. The ability to use such works for reproduction or for other forms of publication will depend on the consent of the relevant rightsholders. Electronic storage of such works is also dependent on such consent in accordance with Norwegian legislation on intellectual property rights. Intellectual property rights and user rights may be transferred by means of an agreement with LINO, the independent copyright organization established for such purposes.

In addition to this comes all public information in the form of studies, reports, legislation, etc. Naturally, material that is incorporated into the resource collection is not meant for reproduction by clients using the data, nor is the collection intended to serve as a text archive for retrieval of information. By stipulating legal deposit of material, the resource collection

will avoid having to purchase much of its text data, thereby greatly reducing its outlay. A scheme for legal deposit of material could be administered by the National Library in collaboration with the operational enterprise for the resource collection.

This does not entail that all written material encompassed by a legal deposit scheme will automatically be incorporated into the resource collection. However, it provides the resource collection agency with an opportunity to ensure balanced corpora with regard to *Bokmål* and *Nynorsk*, genre, time-period, author, etc., which often poses a great problem affecting nearly all language-technology based text resource collections.

Compilation activities for text data are presumed administered under a scheme stipulating legal deposit of material or similar solutions that minimize acquisition costs.

Input from the Norwegian Non-fiction Writers' and Translators' Association indicates that a legal deposit scheme for text material could be implemented without too much trouble:

Rights clearance in connection with storage and subsequent use of copyright-protected material entails a two-step process, in which independent agreements must be signed for the storage and use of material, respectively, alternatively a collective agreement through LINO. However, the administrative work in connection with storage rights could be reduced by establishing legal deposit of a clearly-delimited group of copyrighted works. It would be constructive to link such a scheme to existing legislation for legal deposit of material, as the practical implementation could then be regulated by the provisions of the applicable regulations.

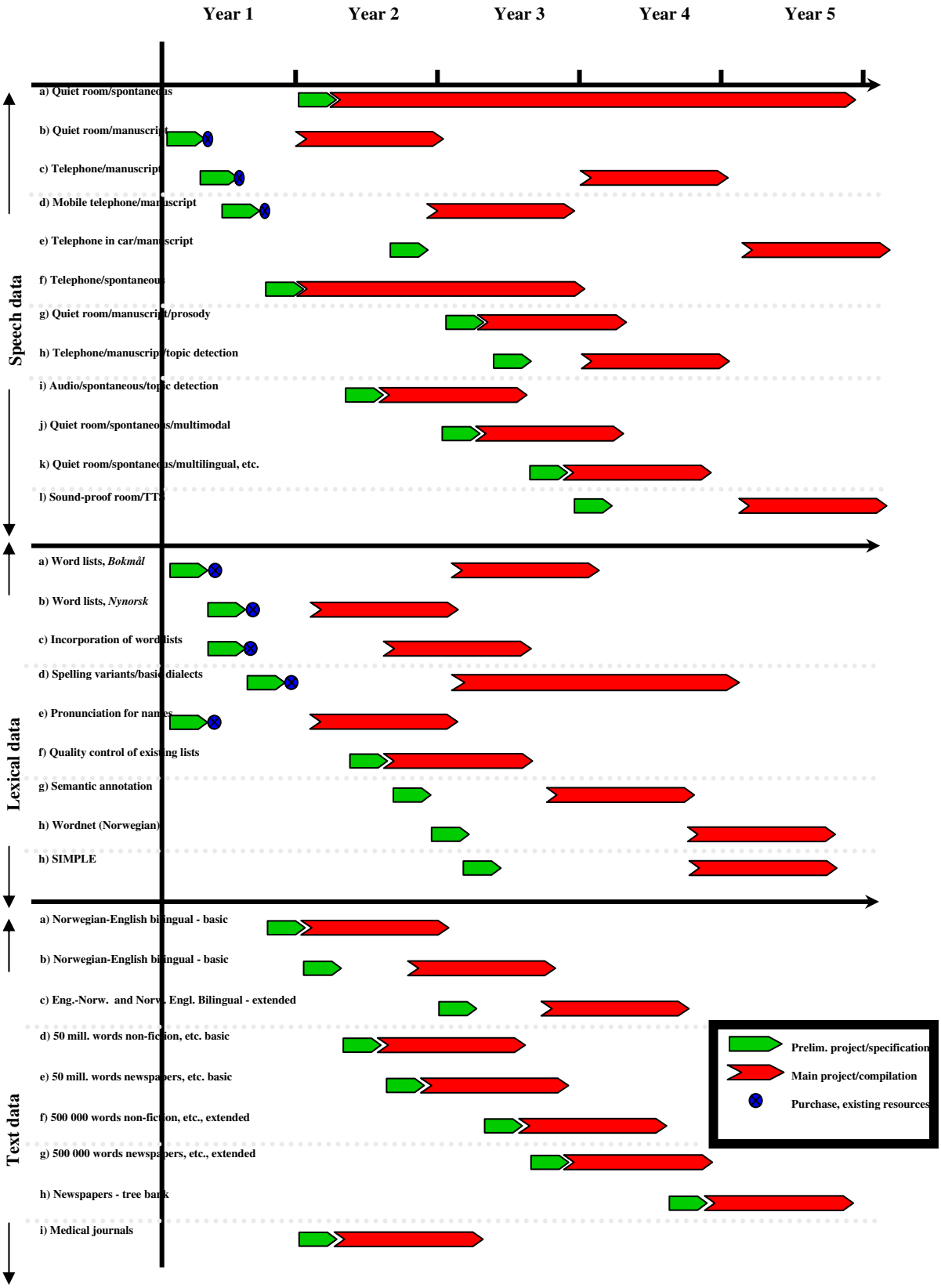
The purpose underlying the Legal Deposit Act is a culturally based need to preserve and document the works created within society. Considerations pertaining to the rightsholder's sole right to storage (reproduction) of the work are not contradicted in this context. If the legally deposited works are subsequently to be made available to the public, this will be regulated by the Copyright Act and applicable agreements, cf. section 6.1 of this document.

Formerly, the lack of technical standards posed an obstacle to the administration of the legal deposit initiative. Thus, it would be appropriate for the individual copyrighted works to be deposited using the same digital platform. In this connection it is important to ensure that the platform has the same digital format as that utilized for the HLT resource collection. Any additional expenses incurred by the depositing body in this context should be wholly or partially covered by the recipient institution (the resource collection agency).

It will be up to the board and administration of the resource collection agency to specify the details regarding a legal deposit scheme for the resource collection.

7.14 Time-frame for the Compilation Activities

The following figures present a time-frame for the compilation activities. Resources for a pilot study, implementation and quality control of data (validation is included in the main projects) have been allocated for all modules. This plan is preliminary and will need to be adapted to allocations to the resource collection and other conditions that can affect priorities. Such decisions will be left to the board of the resource collection agency.


Year 1
Year 2
Year 3
Year 4
Year 5
Speech data
a) Quiet room/spontaneous
b) Quiet room/manuscript
c) Telephone/manuscript
d) Mobile telephone/manuscript
e) Telephone in car/manuscript
f) Telephone/spontaneous
g) Quiet room/manuscript/prosody
h) Telephone/manuscript/topic detection
i) Audio/spontaneous/topic detection
j) Quiet room/spontaneous/multimodal
k) Quiet room/spontaneous/multilingual, etc.
l) Sound-proof room/TTS
Lexical data
a) Word lists, *Bokmål*
b) Word lists, *Nynorsk*
c) Incorporation of word lists
d) Spelling variants/basic dialects
e) Pronunciation for names
f) Quality control of existing lists
g) Semantic annotation
h) Wordnet (Norwegian)
h) SIMPLE
Text data
a) Norwegian-English bilingual - basic
b) Norwegian-English bilingual - basic
c) Eng.-Norw. and Norw. Engl. Bilingual - extended
d) 50 mill. words non-fiction, etc. basic
e) 50 mill. words newspapers, etc. basic
f) 500 000 words non-fiction, etc., extended
g) 500 000 words newspapers, etc., extended
h) Newspapers - tree bank
i) Medical journals
Prelim. project/specification
Main project/compilation
Purchase, existing resources

Key Documentation Underlying the Norwegian Version of this Report

Anbefalinger fra arbejdsgruppen IT på dansk, Ministeriet for Videnskab, Teknologi og Udvikling, Danmark 2001

BLARK: In D. Binnenpoorte o.a.: *A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch*, address at LREC 2002, (Third International Conference on Language Resources and Evaluation, Las Palmas, Grand Canary, Spain, May-June 2002

eContent: EU Council Decision of 22 December 2000 to adopt the Programme for 2001-2004 aimed at stimulating the development and use of European digital content on the Internet and to promote the linguistic diversity of European web-sites in the Information Society.

eEurope *An Information Society for All* – EU Commission Action Plan, 1999 and later (corresponds to eNorway)

eNorway, action plans, versions 1.0, 2.0, 3.0, Ministry of Trade and Industry (the Norwegian Government counterpart to eEurope)

eNorway 2005, Ministry of Trade and Industry, 2002

Handlingsplan for norsk språk og IKT, revidert utgåve, Norsk språkråd 2001

INFO2000 The EU Council Decision 96/339/EC adopting a multi-annual programme to stimulate the development of a European multimedia content industry and to encourage the use of multimedia content in the emerging information society

KUNSTI – Knowledge Development for Norwegian Language Technology (Long-term programme, Research Council of Norway, 2001

MLIS The Council Decision 96/644/EC adopting a multi-annual programme to promote the linguistic diversity of the Community in the information society

Mål i mun, Förslag till handlingsprogram för svenska språket, SOU 2002:27, Stockholm 2002

Norge – en utkant i forkant. Næringsrettet IT-plan 1998-2001, NHD februar 1998

Norsk språkbank. Utredning om et nasjonalt korpus for språkteknologi, Svendsen o.a. 1999

Plan for styrking av norsk, Norsk språkråd 2001

Satsing på informasjonsteknologi for funksjonshemmede (IT-FUNK), 1998-2001, NFR 1998

"Si@!". Elektronisk samhandling i helse- og sosialsektoren, statlig tiltaksplan 2001-2003, SHD 2001

Språkteknologi i Norge – eksisterende og påkrevet forskning, rapport, NFR 2000

St.meld. nr. 9 (2001-2002) Målbruk i offentleg teneste

St.meld. nr. 13 (1997-98) *Målbruk i offentleg teneste*

St.prp. nr. 1 (2002-2003), oktober 2002, for KKD og NHD

Strategy for electronic content 2002 – 2004, Ministry of Trade and Industry 2002

Strategi for eksport og internasjonalisering av IKT-næringen, NHD 2001

Strategi- og handlingsplan for IKT-forskningen i Forskningsrådet, NFR 2000

Strategisk plan for Norsk språkråd 2000-2003, Norsk språkråd 2000

Taleforbedring for funksjonshemmede, sluttrapport, SINTEF 1999

More about Human Language Technology

Washing machines already speak Hindi. A personal digital assistant (PDA) capable of translating between languages might be your new interpreter next time you travel. Your future car may come equipped with an electronic map and a voice to give you directions. While digital librarians and secretaries already excist, most of them only understand English!

What is human language technology? How important is it to incorporate the Norwegian language into this sphere of technological development? What are the consequences of not doing so? What role can a collection of Norwegian language resources play in the further development of this technology?

The Norwegian Language Council
C.J. Hambros plass 5
P.O. Box 8107 Dep
NO-0032 Oslo
Telephone: (+47) 24 14 03 50