The Centre for Dutch Language and Speech Technology (TST Centre)

J. C.T. Beeken, P.H.J. van der Kamp

Institute for Dutch Lexicology (INL)
Matthias de Vrieshof 2-3, 2311 BZ Leiden, NL
PO Box 9515, 2300 RA Leiden, NL
beeken@inl.nl, kamp@inl.nl

Abstract

In September 2003, the Centre for Language and Speech Technology (TST Centre) was created. The TST Centre is responsible for the maintenance, management, distribution, availability and accessibility of basic digital language resources (LRs) of Dutch that are financed by public funding; the TST Centre is located at the Institute for Dutch Lexicology (INL) in Leiden, the Netherlands.
In the first year of its existence, the TST Centre at the INL in Leiden and its dependanceat the University of Antwerp (B) made agreements with various Dutch and Flemish academic as well as industrial subcontractors to assist them in performing the tasks linked to the maintenance, management and accessibility of the LRs. The partners involved are the Max Planck Institute in Nijmegen (NL), the University of Leuven (B), the Free University of Amsterdam (NL) and the company Technologie & Integratie in Ghent (B).
The initial set of LRs entrusted to the Centre are the Spoken Dutch Corpus (CGN), the bilingual and monolingual lexicons of the CLVV (Commissie Lexicografische Vertaalvoorzieningen), and some of the major products of the INL.

1. Language Resources

1.1 Projects

The initial set of LRs has been selected by the TST Centre in co-operation with the Dutch Language Union (Nederlandse Taalunie) and a group of suppliers and users. From June 2004 onwards, this task will be assigned to the management board in co-operation with the TST Centre.
The initial set contains the results of three projects, being

1 the Spoken Dutch Corpus (CGN),
2. the monolingual as well as the bilingual dictionaries and application tools of the CLVV (Commissie Lexicografische Vertaalvoorzieningingen),
3. the monolingual corpora, lexicons and application tools of the INL.

1.1.1 The Spoken Dutch Corpus (CGN)

The Spoken Dutch Corpus is a database of Standard Dutch spoken by adults in Flanders and the Netherlands. The corpus contains ca. 9 million words (approximately 1000 hours of speech). The sixth release of the Corpus appeared in November 2002. The final release, which has been made available by the TST Centre appeared 1st March 2004 (see also Schuurman et al., 2004). All of the speech recordings have been orthographically transcribed; all of the data are enriched with POS tags and lexicological relations. Also made available are the sets of phonetic transcriptions, alignments at word level, prosodic annotations, syntactic annotations, metadata and the CGN lexicon. The exploration tool used is COREX, a tool developed at the Max Planck Institute in Nijmegen. In order to guide users through the available data and exploration software a demo-CD will be realised by the TST Centre before the end of May 2004. More information can be found on the website of the TST Centre (http://www.tst.inl.nl).

1.1.2 The CLVV-products

During the period 1993-2003 the CLVV was the organisation responsible for the development of multifunctional lexical databases which now will all be transferred to the TST Centre. These databases function as fundamental material for both bilingual dictionaries and learner's dictionaries which all have a social, economic and/or political importance but which only could be developed thanks to public funding. Already available are the data for Arabic, Hungarian, Italian, Russian, Danish, Indonesian, Norwegian, Portuguese, Polish and Rumanian, in relation to the corresponding or equivalent data of Dutch as the source and/or target language. From 2005 onwards also the Estonian, Finnish, Korean, New Greek and Turkish data can be accessed at the TST Centre.
The CLVV not only developed and elaborated all of the above mentioned data, it also developed a meta-infrastructure, being the Reference Data for Dutch (ReferentieBestand Nederlands) and the OMBI-tool, i.e. a tool for the inversion of bilingual data.
Users of the CLVV-products are lexicographers, but also translators and language teachers.

1.1.3 The INL-products

The INL-products which will be made available by the TST Centre are:

1. the standard work for Dutch Spelling (Woordenlijst Nederlandse taal (Groene boekje)),
2. some results of the dictionary of Old Dutch (ONW or Oudnederlands Woordenboek),
3. some results of the dictionary of Standard Dutch (ANW or Algmeen Nederlands Woordenboek),
4. some results of the Language Database (TB or Taalbank)

The Woordenlijst Nederlandse taal (Groene boekje) contains the official spelling rules for Dutch and a list of about 111.000 entries, each heading a set of types of information, e.g. the flectional forms of nouns, adjectives

and verbs, grammatical gender of nouns, references to alternative spelling forms, meaning or semantic domain in case of homonyms etc.
The project Dictionary of Old Dutch (ONW) started in 1999. The main goal of the project is to describe the vocabulary of Old Dutch in a standardised lexicographical way and to make all of the information electronically available for a public of both specialists and common users. The project covers the period between ca. 500 and ca. 1200. The project will be finished in 2008. In 2004 about 25% of the enriched and annotated data will become available at the TST Centre.
In 1999 also the project Dictionary of Standard Dutch (ANW) started; it will end in 2019. The main goal is the lexicographical description of Standard Dutch in Flanders and the Netherlands during the years 1970-2019. The ANW will be an on line dictionary, that is a dictionary stored at a local server that can be reached via the internet by both specialists and common users. In 2004 the first results of the task *Neologisms* will be made available by the TST Centre.
Since 1990, the INL has been collecting digital texts offered by Flemish and Dutch authors and distributors. Up until now the project TB delivered 'data on request' asked for by national and international researchers and academics without immediate economic interests. From 2004 onwards, this task will be taken over by the TST Centre. In the past years, the project TB also developed text corpora, which can be consulted over the internet by means of a retrieval system. These corpora are the *5 million words corpus*, the *27 million words corpus* and the *38 million words corpus*. Users of these corpora are researchers, lexicographers, corpus linguists, and common users with an interest in Dutch as their native language (see also http://www.inl.nl/eng/corp/corp.htm). Besides the three corpora mentioned, the TB also developed the PAROLE Distributable Corpus. It contains 3 million words of which 250.000 are linguistically enriched and annotated. The Distributable Corpus is a subset of the Dutch PAROLE Corpus, which contains about 20 million words, all annotated for TEI-code CES-level1, automatically lemmatised and POS-tagged (see also Kamp & Kruyt, 2004). Furthermore, the TB developed the PAROLE lexicon, containing 20.000 entries, all morpho-syntactically and syntactically annotated (for more information, http://www.inl.nl/eng/corp/parole.htm). All of these projects will in the near future be added to the list of products for which the TST Centre is responsible as far as management, maintenance, distribution and support are concerned.

1.2 Overview of Tasks

In the first phase of the project (January 2004 - October 2004), the following tasks are carried out:
-installation of the technical infrastructure;
-installation of the global infrastructure, the division of tasks;
-design of profiles of LRs, its suppliers and users;
-elaboration of the procedures for acquisition, management, maintenance, distribution, support, helpdesk and user-specific services connected with each type of LR (Van Sterkenburg, Kruyt & Van der Kamp, 2002);
-development of the acceptance criteria for LRs;
-selection and acceptance of LRs which are highly ranked on the BLARK-ELARK priority list (Daelemans & Strik, 2002);
-installation, configuration, conversion of the initial set of LRs;
-implementation of security devices and a well-designed access rights system;
-design of a website giving users conditional access to the infrastructure and LRs;
-institutionalisation of the management board and the steering committees;
-evaluation of the first results.

1.3 Procedures

Due to technical considerations, criteria concerning the quality of the content, legal issues and conditions and administrative demands it is necessary that for every type of product a specific procedure for management, maintenance, support etc. will be developed and elaborated. All of these aspects have to be considered in every phase of the lifecycle of an LR: from acquisition to distribution. Moreover, every type-specific procedure will be adapted to the characteristics of every acquired product, be it a corpus, a lexicon, a thesaurus, an application or exploration tool. Consequently, every product will also be uniquely identified in terms of location, team members, required expertise, platform and distribution channels, and, finally, user specifications.
In order to guarantee that the complex activities of management, maintenance, distribution, accessibility, support and services are performed in a correct and efficient way, all of these considerations and demands have to be listed and concretised. The different sets of basic actions, constituting the set of procedures making different types of LRs available to different groups of users, are extensively analysed and described in the Blueprint written by Van Sterkenburg, Kruyt and Van der Kamp (2002).

2. Infrastructural issues

2.1 Introduction

To make all the TST data available to users, a technical infrastructure is a prerequisite. The infrastructure will not be discussed here in detail; only some aspects that are relevant from the user's perspective will be looked at
Put in general, since LRs are subject to copyright restrictions, measures have to be taken to protect the infrastructure from being hacked or threatened by other internet related security hazards. Beside that, a well-designed access rights system has to be implemented in order to give users and/or suppliers access to those language resources (LRs) for which they are contractually authorised.

2.2 User Identification

Only authorised users are allowed to access the LRs. A user can be regarded as authorised when all legal issues are dealt with, i.e. when the contract is signed. He will then be supplied with a userid and (one time) password. If

necessary other techniques will be used like access based on IP-address or digital signatures.

2.3 Services and Access Rights

To implement access rights for the available services, we need to make a distinction between external and internal services. Both can be further decomposed in public and private.
External public means that the service is accessible for non-registered users; external private means that a service is available for registered and accepted users or suppliers only. There will be two external public services: the website and the helpdesk. The website will provide detailed information about available LRs and on how to become a registered user or supplier, and will offer technical information about e.g. accepted data formats. Information about additional services of the TST Centre like e.g. data delivery on demand, can also be found on the website. The helpdesk can be contacted in case the information on the website is not sufficient. With respect to information about available LRs it is foreseen that metadata will play an important role. Currently we have to deal with several types of metadata or no metadata at all. Part of the maintenance task will be to unify metadata so that it complies with appropriate international standards, thus making it accessible for software that supports such standards.
The external private service consists of downloading or uploading LRs, running software and the helpdesk. Depending on the contract and access rights a user can download or upload specific LRs. Uploading will be used by suppliers or by users who maintained a LR and want to upload a new version. If the users' own computer facilities are not sufficient to run TST software they can run the software on the computers of the TST Centre. The helpdesk can be contacted whenever technical problems occur during downloading/uploading or using LRs. The helpdesk will also provide additional information in case the documentation of the maintained LR is not sufficient.
The internal services have to do with system management tasks like implementing security and backup and data maintenance tasks. They will not be further discussed here.
Having external or internal private access rights does not mean that a user has access to all LRs. Depending on the contract or maintenance tasks a user is granted specific access rights. Well-known access rights are read, write, delete and execute. A user who is authorised to download one or more LRs will only be granted read access rights for those LRs. Typical authorised suppliers will be granted write access rights only in order to enable them to upload the LR. The same applies to internal users. A software engineer or computer linguist who wants to maintain a LR will be granted with the applicable access rights to do the job.

2.4 Version Control

Users that are allowed to upload new, updated or maintained LRs place them in a kind of quarantine. We do so because uploads should never replace the original version of a LR. Furthermore, it has to be checked whether the uploaded LR really meets the TST Centre's quality standards. If the LR is okay, authorised personnel will make it available for distribution.
For each LR a directory structure will be implemented, consisting of a top level directory and one or more subdirectories, each subdirectory containing a specific version of the LR. Previous versions of the LR cannot be replaced by newer ones, for research – or other requirements can make it necessary to use an older version.

3. Organisational Structure

The TST Centre is responsible for the acquisition, management, maintenance, accessibility, distribution, help-desk function and service of each of the LRs entrusted to it. The Centre also decides at which location a specific LR is to be stored and has to provide for the various service functions.
The daily management is entrusted to the project manager. At a more structural level, the TST Centre has to report to and work together with several committees.
In the near future, the Dutch Language Union (Nederlandse Taalunie) will set up the different committees, i.e. the management board and the scientific steering committees, each of which will exist of supervisors and users, linked to one or more LRs. In the management board, each steering committee is represented by one of its members, preferably its chairperson. An observer of the TST Centre will be assigned to each steering committee.
The most important task of the steering committees is to advise the management board about the needs and responses of academics and industrials regarding the supervised LRs, suggesting a priority list of (a selection of) further actions and a timetable.
The management board communicates directly to the Centre and the Dutch Language Union, providing them with an overview and a priority list of a selection of further actions together with a short- and long-term timetable. The suggestions made will be presented to the TST-Platform. This Platform consists of Dutch and Flemish policy makers responsible for the development, evaluation and accessibility of language and speech technology; it finances the development and maintenance of data and tools in the domains of these technologies.
Furthermore, an important task of the TST Centre is the establishment of well-functioning communication channels with its subcontractors, the management board, the steering committees and user groups. The TST Centre will also develop and establish the correct relations with its suppliers and distributors as well as with the user groups. The Dependance and the subcontractors perform the same tasks for the LRs entrusted to them, supervised by the TST Centre.

4. Future Work

Concerning the LRs that are acquired, maintained and distributed by the TST Centre, the Electronic Grammar of Dutch (eANS), NL-Translex (the pairs Dutch-English and Dutch-French for machine translation) and the results of various terminological projects for Dutch will be added to the initial list in 2005 or 2006.

In the next phase, much effort will also be put into a set of model contracts which will broaden the groups of users from academics and researchers to industrial users and users with an economic interest. Moreover, a business plan will be developed for each of the products, opening up the use of the material made available by the TST Centre for purposes and goals other than purely scientific ones.

Finally, from an infrastructural point of view security, access rights and accessibility are important for the TST Centre, so the focus is currently on these issues. That implies that for the time being topics like archiving and digital durability don't have the highest priority. Especially the latter cannot be neglected as proprietary or other data formats can change or disappear. "Digital documents are in general dependent on application software to make them accessible and meaningful." (Rothenberg, 1999). This is a threat for the reuse of such data and without proper measures the data can become useless. These topics will be addressed later as well as the development of guidelines.

References

Daelemans, W. et al. (2002). Dutch in Language and Speech Technology: Priorities for a Basic Toolkit (Het Nederlands in taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen).

Kamp, P.H.J. van der & Kruyt, J.G. (2004). Putting the Dutch PAROLE Corpus to Work. In Proceedings LREC 2004.

Rothenberg, J. (1999). Avoiding Technological Quicksand: Finding a viable technical foundation for digital Preservation. *Report to the Council on Library and Information Resources.* Washington D.C.

Schuurman, I. et al. (2004). Linguistic Annotation of the Spoken Dutch Corpus: If we had to do it all over again... In Proceedings LREC 2004.

Sterkenburg, P.G.J. van , Kruyt, J.G., Kamp, P.H.J. van der (2002). Blueprint for the maintenance, management and availability of digital language resources financed with public funding (De Blauwdruk voor onderhoud, beheer en terbeschikkingstelling van door de overheid gefinancierde digitale materialen).