Functioning of the Centre for Dutch Language and Speech Technology

Michel Boekestein, Griet Depoorter, Remco van Veenendaal

TST Centre
Institute for Dutch Lexicology (INL)
Matthias de Vrieshof 2-3
NL-2311 BZ Leiden, NL

TST Centre
University of Antwerp
Universiteitsplein 1, CDE, A1.24
B-2610 Wilrijk, BE

boekestein@inl.nl, depoorter@inl.nl, veenendaal@inl.nl

Abstract

The TST Centre manages a broad collection of Dutch digital language resources. It is an initiative of the Dutch Language Union (Nederlandse Taalunie), and is meant to reinforce research in the area of language and speech technology. It does this by stimulating the reuse of these language resources. The TST Centre keeps these resources up to date, facilitates their availability, and offers services such as providing information, documentation, online access, offering catalogues, custom-made data, etc. Also, the TST Centre strives for a uniformised, if not standardised, treatment of language resources of the same nature.

A well-thought, structured administration system is needed to manage the various language resources, their updates, derived products, IPR, user administration, etc.

We will discuss the organisation, tasks and services of the TST Centre, and the language resources it maintains. Also, we will look into practical data management solutions, IPR issues, and our activities in standardisation and linking language resources.

1. TST Centre

The Centre for Language and Speech Technology (TST Centre) (1) was set up in September 2003 and is fully functional since June 2004. The centre manages a broad collection of basic Dutch digital language resources (LRs): audio recordings, digitized texts, annotated corpora, computational mono- and multilingual lexica, POS taggers, parsers, etc.

The TST Centre is an initiative of the Dutch Language Union (Nederlandse Taalunie) (2) and is located at the Institute for Dutch Lexicology (INL) in Leiden, the Netherlands and in an auxiliary branch at the University of Antwerp (UA) in Belgium. The TST Centre was set up to reinforce the research in the field of Dutch language and speech technology. Its premier goal is to stimulate the reuse of Dutch digital materials, financed with public funding. In practice, this amounts to making the LRs available for scientific research and maintaining them by keeping them up-to-date and conform to ruling standards. The unique strengths of the TST Centre lie in the synergy of the combination of LRs. They can be compared to each other in order to track down bugs, and the LRs can be integrated or simultaneously and uniformly consulted. Also, one LR can be used to enrich another one.

1.1. Products

The TST Centre manages different kinds of LRs:

- CGN: The Spoken Dutch Corpus (3) is a speech corpus of standard Dutch spoken by adults in the Netherlands and Flanders. The corpus contains approximately 9 million words.
- ALVV products: monolingual and bilingual dictionaries and application tools of the Adviescommissie Lexicografische VertaalVoorzieningen. Monolingual dictionaries in this product line include the Referentiebestand Belgisch Nederlands (RBBN), and the Referentiebestand Nederlands (RBN). Also, several bilingual dictionaries are, or will be available via the TST Centre, such as the Dutch-Indonesian dictionary, Dutch-Arabic, etc. The set of ALVV products also includes the dictionary editing tool OMBI. This tool is able to reverse a bilingual dictionary.
- INL products: products from the Institute for Dutch Lexicology of which the TST Centre is a department. The institute has composed several written corpora (e.g. the Parole Corpus (4)), and lexical data, like the Neologismenlijst from the ANW project (Algemeen Nederlands Woordenboek).
- e-Lex: a large lexical database containing more than 220 thousand entries and more than 600 thousand inflected forms. The entries are provided with morphological, syntactic, phonologial and (partly) semantic information. This lexicon was especially designed for use in language and speech applications such as ASR systems, etc.
- Master copy of the Woordenlijst Nederlandse Taal (Groene Boekje). This is the official spelling reference for the Dutch-speaking regions. It contains the official spelling rules of Dutch, and a list of more than 100 thousand words illustrating these spelling rules.

In the future a number of products, or at least their peripheral activities (licence administration, user access, etc.) are intended to be transferred to the TST Centre. They include:

- terminology products, compiled under supervision of CoTerm (Commissie Terminologie);
- the results of the STEVIN projects (4);
- diachronic INL products;
- e-ANS (6) (Algemene Nederlandse Spraakkunst), the standard reference of Dutch grammar;
- NL-Translex: an automatic translation system for Dutch, funded by the Dutch Language Union, the European Union, and Systran.

1.2. Tasks

The tasks of the TST Centre can be grouped around five keywords: acquisition, management, maintenance, availability and service.

1.2.1. Acquisition

The TST Centre inventories relevant resources which are not (yet) managed by the centre and it puts efforts in acquiring those digital materials. Prior to successful incorporation into the TST collection, quality control is carried out by an independent third party and the Intellectual Property Rights (IPR) are sorted out. The Dutch Language Union receives all IPR.

1.2.2. Management

Management of the digital LRs comprises, amongst other things, storage, back-up and version management. The latter is done by means of CVS (Concurrent Versions System).

Also indispensable is good knowledge management: documentation and know-how with reference to TST products have to be centralised and stored in a logical way. It is the task of the TST Centre to try and collect missing information.

Knowledge management within the TST Centre was dealt with by installing a set of online collaboration tools. The open source software package eGroupware (7) contains all necessary items: an electronic agenda, project management software, a file manager, and a wiki, in which useful information is shared between colleagues. This is a particularly efficient solution in an environment such as the TST Centre, where not everybody works in the same building.

1.2.3. Maintenance

Maintaining data comprises improving resources by inventorying, assessing and handling bugs and improving and updating software. For more information about bug handling, see section 1.2.5. Maintenance also means publishing new updates through various media like print, diskettes, DVDs, tapes, internet, etc.

1.2.4. Availability

The most pragmatic way to make a newly received resource available is publishing it as it is, i.e., without making any changes to the resource. Particularly when the resource had already been taken in use before the transfer to the TST Centre. People using the resource probably already got used to the user interface and exploitation tools. Another reason is that the manufacturer of the resource is the best person to know how to build and exploit his own resource.

However, converting LRs into standard formats have a number of benefits:

- standardisation facilitates publishing new LRs online, once the needed software appropriate for the standard format is up and running;
- it enables (or at least facilitates) combining LRs. End users would then be able, for example, to search in various lexicons with just one single search action;
- it enables (or at least facilitates) comparing, enriching, or evaluating other LRs;
- it can turn out to be very user friendly: no different user environments for the same type of LRs – a single signon is enough to browse various LRs;
- it makes the LRs more manageable: requirements in terms of platform and software package versions are minimized;
- it increases stability: LR software using data represented in a certain standard is likely to be kept compatible with that standard.

So, on the one hand we try to fulfill our five main tasks for the orginal product. On the other hand, we strive for standardisation, so that LRs of the same nature are available in one and the same way.

However, in case of LRs which are still in development, or only in planning stage, we point at the advantages of using a standard format, and show which existing tools the developers could probably reuse.

In the STEVIN programme, for example, several projects will deliver annotated corpora. As the already existing CGN also is an annotated corpus, and the corpus exploitation software that is included can handle these annotations, we generally suggest the STEVIN project members to comply with the annotation format used in the CGN annotation layers. This way, we are trying to achieve two goals: uniformity in data formats, and reuse of existing tools (in this case, CGN's corpus exploitation software called Corex).

Also, our internal project aiming at linking the lexicons available in the TST Centre, makes use of the representation structure proposed in the LMF (Lexical Markup Framework) standard. LMF is an ISO standard which is still in development.

As far as the standardisation in corpora is concerned, the TST Centre participates in the DAM-LR (8) (Distributed Access Management for Lexical Resources) project. Here, the IMDI standard is used to code the corpus metadata. We intend to make the TST Centre's corpora available through this standard: the prototype of our IMDI portal is planned to be up and running in July of this year. This portal will be connected to other IMDI portal nodes in Europe.

The TST Centre makes various resources available to the public, but in doing so it has to protect the data sufficiently. Therefore a licensing system has been set up, so users can only consult the data or gain access to it after signing a licence, or, in case of an online resource, accepting the licence online.

The TST Centre is primarily meant for researchers, students, educational or government institutions, and other non-profit institutions and users. They are offered free access to the data via the internet. However, when delivering data on CD, DVD or tapes is required or when specific data or information is asked and complex and time-consuming actions have to be undertaken by the TST Centre, a compensation is demanded. The resources can be used for scientific research, education, integration in own projects, etc.

Users or institutions with a profit motive who wish to consult or use the TST resources, are invited to contact the project leader of the TST Centre, after which a trilateral conference can be started between the centre, the supplier of the resources and the user. Furthermore, the TST Centre wants to play an intermediating role in the contacts between the companies amongst one another: the possibility to exchange materials by means of shareware or file sharing will be looked into.

1.2.5. Service

The TST Centre has recently renewed its website (1) which offers (potential) users of the TST data information with reference to the LRs: documentation, manuals,

known bugs, news concerning workshops or releases, evaluation reports, etc. If users need additional information, they can turn to various LR specific helpdesks, either via e-mail or via feedback forms on the website. If needed, selected parts of the data can be delivered. Various catalogues are offered: catalogues of the available LRs, bug reports, custom-made data, demos, etc.

Part of the website is a forum, on which the users can interact with each other. The TST Centre employees act as moderators. A tool by which bug reports can be conveyed is also offered: Bugzilla, which acts as a bug reporting and bug administration tool.

The website is not the only sevice offered by the TST Centre. Another service is the custom-made data we offer on request. Depending on the amount of work involved and the customer's purposes, the customer pays either nothing, cost price, or a commercial price.

Also, the TST Centre was and is responsible for the technical part of the Dutch respelling project. Recently, a number of official updates in the Dutch spelling rules took place, and dictionary publishers could have their word lists respelled through the online respelling software that was (and still is) running at the TST Centre's website. This software uses data from the earlier mentioned Woordenlijst Nederlandse Taal.

We are planning to organise educational activities for interested students at Dutch universities, based on the Spoken Dutch Corpus.

Furthermore, we are actively involved in the earlier mentioned STEVIN programme:

- the TST Centre takes part in the selection of appropriate project candidates;
- we help the project members clearing their IPR issues (see also section 1.4);
- the STEVIN-projects' intermediate and final results will be made available by the TST Centre;
- we closely follow the plannings and progress of the various projects, and advise them about the use of methods, tools, and/or formats. As we have knowledge about other, similar projects and products, we can advise people to use a certain standard format, and encourage them to reuse already existing resources or tools.

Last but not least, we regularly organise workshops (e.g. about IPR, the Dutch Spoken Corpus, knowledge sharing and management, etc.); each six months we organise our own "TST Day" with a relevant theme (we had a Lexicon Day, and a Corpus Day, for example), and we construct demo DVDs about the practical use of (some of) our products.

1.3. Production Street

Although the resources hosted at the TST Centre are not really different in nature (they are either lexicons , corpora or tools), their content and structure generally differ in many ways from each other:

- modality: a lexicon or corpus may consist of written, audio and/or video data;
- format: used to store the resource data: database formats, XML, LMF, IMDI, etc.;
- internal structure: two resources that are similar in nature can be completely different in internal structure.

Additionally, some products have their proper user interface and/or user access management system incorporated.

In this section, a solution is presented to:

- make both the original LRs and their standardised versions available to the end user;
- manage and maintain the LRs, their updates and conversions;
- keep track of version and update information, as well as the way the conversions have been carried out;
- provide an interface to the end user, and garantee data protection.

In order to meet these requirements, we have developed the concept of a production street. See figure 1 for its schematic representation. It consists of three layers; looking upward from the bottom, they are:

- the TST Centre's repository, containing a data storage structure, together with process definitions of converting the original LR into a standard, or uniformised, format;
- the interface & access part; which functions as a bridge between the TST Centre's LRs and the end user;
- the end user's environment.

At the bottom left in the scheme, the LR supplier transfers his LRs to the TST Centre. Translated to our five main tasks, this is the acquisition part. The TST Centre is prepared to handle the different media on which the LRs can be delivered: CD, DVD, electronically, etc. Every time a new LR comes in, the original data is stored in the especially allocated disk space, and we update the administration. The location where we store the orignals is regarded as a "treasure"; no data is allowed to be modified here.

As mentioned before, the original LR should in principle stay available as-is after the transfer to the TST Centre. Hence the arrow upward from *originals*, via the interface to the end user. In case the product has its own user interface (e.g., a text corpus with its proper online search interface), this interface will be used. This also holds for the user access management system and the licensing system (if any).

Updates can be produces by the LR supplier (or any other external party), by the TST Centre, or both. Updates done by an external party can be seen and treated as a new original product. TST-internal updates are part of the update category (see scheme). The version management system CVS takes care of the version administration on the originals and updates. A backup strategy for these resources is taken care of: a backup is made regularly.

The parts *conversion/integration* and TST-database are designed for the standardisation and linking issues discussed above. Each conversion from original/update to the TST-database needs to be reproducable, so that lost data can be recovered, and bugs easily traceable. This requires a conversion environment, where the data, software, technical documentation, and other relevant information are stored. Moreover, this environment is to be used as a space for preparing custom-made data. After all, generating custom-made data can be seen as a kind of conversion.

In the layer above, the interface & access part, the applications which make the resources online available, can be found. However, in cases where the TST Centre

gets a request for a resource on DVD, the DVD writer is the "interface". The user access part is sometimes incorporated in the interface. For example, if a user would like to consult a lexicon online, the interface checks if he is logged in, and if he has access rights to this specific resource.

new products and projects are coming our way. The proposed production street is thought to be an adequate solution to serve both the users of the original products and the (potential) users of the standardised versions.

As for standardisation: we are actively involved in development and application of new standards relevant to

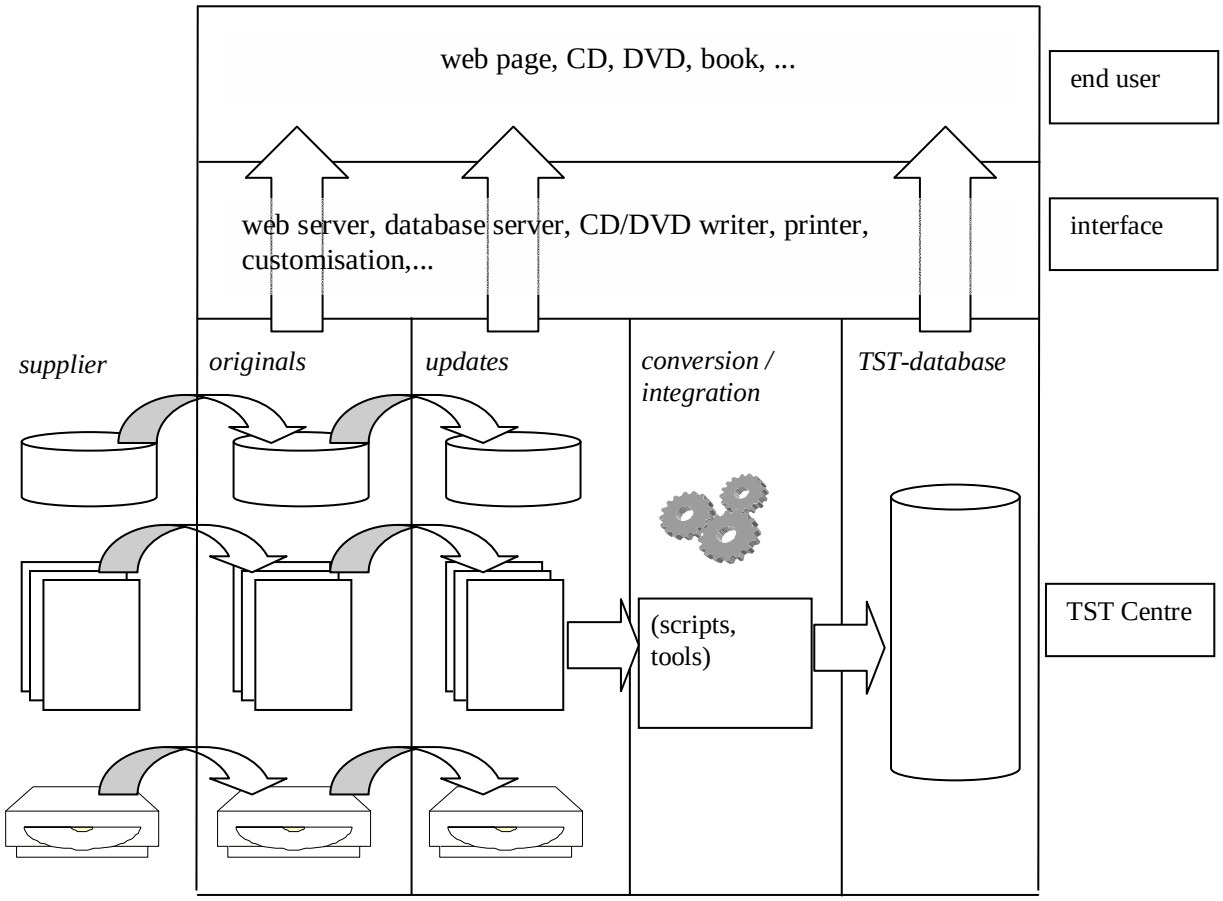

Figure 1: TST Centre's Production Street

1.4. IPR, Contracts, and Licences

The TST Centre's licence agreements can roughly be divided into two types: licence agreements for non-commercial use, and the ones for commercial use. Licences exist for the online LRs, their offline versions

(this generally means the complete database), and the custom-made products. Most frequently used are the online licences for non-commercial use.

The online licences are based on a standard licence. They can be obtained after having registered at the TST Centre's website: they navigate to the LR they would like to gain online access to, the licence is shown on the screen and the user accepts it by simply clicking the OK-button. The software behind the centre's website stores the licence administration. It is known which user has required which licences, and until what date the licences remain valid.

1.5. Conclusion and Future Work

In the two years of its active existence, the TST Centre manages a significant number of high quality LRs, and the LRs we manage.

We are playing a significant role in helping the STEVIN project members clearing their IPR issues, and we will continue to do so. Also, the licence administration system behind the website, which is already meeting its current requirements, will be improved.

The website was renewed this year - its design has become more pragmatic. Users quickly and easily get to consult the LRs we have put online. The new server we acquired is being installed at the time of writing this paper; the online LR consultation software will be replaced by a more powerful and more flexible solution.

2. References

(1) http://www.tst.inl.nl
(2) http://www.ntu.nl
(3) http://lands.let.kun.nl/cgn/ehome.htm
(4) http://parole.inl.nl/html/index.html
(5) http://www.nwo.nl/stevin
(6) http://www.ru.nl/e-ans/
(7) http://www.egroupware.org/
(8) http://www.mpi.nl/DAM-LR/