



# CLARIN-NL

## SUBPROJECTS OVERVIEW 2011

Persistent and stable Research Infrastructure  
for the Humanities

Easy access to language resources  
and technology

Under formal and semantic interoperability





---

## Contents

Introduction	4
IIP	5
Metadata Project	6
S&D	8
AAM-LR	10
Adelheid	12
ADEPT	14
DUELME-LMF	16
INTER-VIEWS	18
MIMORE	20
SignLinC	22
TDS Curator	24
TICCLops	26
TQE	28
WFT-GTB	30
CKCC	32
TTNWW	34
Call 2 Projects	36



# Introduction

---

The CLARIN-NL project aims to create a research infrastructure for humanities researchers that work with language data and tools. It forms the Dutch national counterpart of the CLARIN enterprise on the European level (CLARIN-EU). The project started in April 2009 and will run until April 2015

An infrastructure for language data and tools must first of all provide facilities for storing data and tools, for making them visible and accessible to researchers via browsing, and searching, and for applying tools to data.

A number of projects have been set up to design and implement such facilities in the CLARIN centres in the Netherlands. They include the *Infrastructure Implementation Project (IIP)*, the *Metadata project*, and the *Search&Develop (S&D)* project.

Second, the infrastructure must be filled with data and tools and their metadata. In addition, these data and tools must comply with CLARIN standards, especially standards to guarantee formal and semantic interoperability. A range of data curation and demonstrator projects has been set up to achieve this for existing data and tools after the first open call. The projects are *AAM-LR*, *Adelheid*, *ADEPT*, *DUELME-LMF*, *INTERVIEWS*, *MIMORE*, *SignLinc*, *TDS Curator*, *TICCLops*, *TQE*, and *WFT-GTB*.

Third, we are happy that the existing and independently financed *CKCC* project expressed a desire to work in a CLARIN-compatible manner. Doing that is good for the CLARIN infrastructure, but –and that is even more important– it will also increase the research opportunities for *CKCC* since by working in a CLARIN-compatible manner it becomes much easier to make use of state-of-the-art natural language processing tools.

Many (though certainly not all) of the data and tools from the Netherlands that will be incorporated in the CLARIN-infrastructure will involve the Dutch language. Since the Netherlands and Flanders share the Dutch language, it is natural to cooperate closely with Flanders on such data and tools. This cooperation has been realized in the project *TTNWW*, that focuses on making existing text and speech data and tools that have been developed in earlier cooperation projects between the Netherlands and Flanders (esp. the *STEVIN* programme) compatible with CLARIN requirements.

For all the projects mentioned above, this booklet provides an overview of their plans and achievements so far. Finally, it also provides a very brief overview of plans of the subprojects that have been awarded funding in the second open call for data curation and demonstrator projects and that will start up in 2011.



*Project coordinator: ir. Daan Broeder*

CLARIN's official mission is that it wants to provide a new infrastructure that makes Language Resources and Technology available and readably usable for Language and SSH (Social Sciences and Humanities) researchers. This new infrastructure should offer highly available and robust services and step away from the still currently dominating download-first paradigm. Emphasis is placed on robustness and persistency of services and resource access that allow consistent long-term research procedures and verifiability.

We plan to achieve this by following a few major principles:

- Building a network of centers that offer such services and that take care of data persistence and curation.
- Build a common and flexible framework of high-quality metadata to allow easy discovery of resources and services.
- Setting up a federation that allows researchers to use a single-identity to access services and resources with single sign-on in the federation and create virtual collections.
- Provide a framework for web services that allows researchers to include their algorithms and orchestrate them in an easy way.

All needs to be based on standards and best practices to achieve a high degree of integration of and interoperability between all types of language resources and tools. In this subproject we work to implement (a first version of) such an infrastructure with MPI (Max Planck Institute for Psycholinguistics), INL (Institute for Dutch Lexicology), Meertens Institute and DANS as initial CLARIN centres.

The work in the IIP project is based on and cooperates with what was already achieved and planned within the European CLARIN project's preparatory phase. In particular the work done on CMDI (CLARIN Metadata Infrastructure): a flexible component based metadata schema design together with supporting tools and registries. Also the work on establishing a European CLARIN Service Provider federation will serve as the basis for the AAI (Authentication and Authorization Infrastructure work) within the IIP project.



# Metadata Project

## Creating and testing CLARIN Metadata Components

Project coordinator: ir. Daan Broeder

In the CLARIN-EU project a design was made for the CLARIN Metadata Infrastructure (CMDI) that should create a single interoperable domain of metadata descriptions for all resources housed at CLARIN centers. At the basis of this design is the concept of component metadata; these are bundles of metadata elements that describe related aspects of a resource. All metadata elements need to be linked to a concept in the ISO Data Category Registry (ISO-DCR) for semantic interoperability. Metadata modelers can reuse existing components or create their own and gather these into a metadata schema that they think best describes a specific resource type. The CMDI architecture also builds supporting tools for the creation and exploitation of CLARIN metadata: a GUI metadata component registry and editor and a metadata editor “ARBIL” that allows creation of metadata records. The metadata exploitation tools are currently (2010) in the “prototype” stage.

The CLARIN-NL metadata project’s purpose was to:

- Test the viability of this approach by describing a subset of the resources currently housed at two prospective CLARIN centers: INL and Meertens Institute.
- Provide a ready set of metadata components that can be used to seed the future CLARIN-EU metadata component registry.
- Provide best practice guidance for future users of the CMDI.

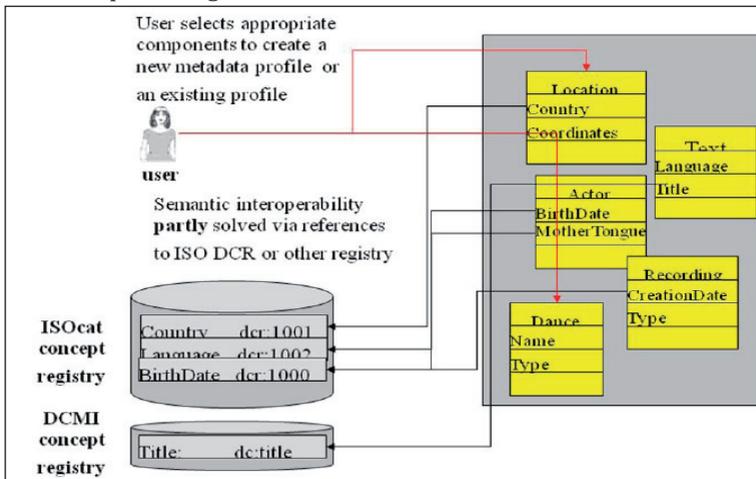


Figure 1. The CMDI component registry showing components with their references to the ISO-DCR.



---

A question to be answered was if the CMDI approach would be applicable to resources of especially non-multi-modal/multimedia type, since that particular type of resources had already been under consideration by the CMDI designers. Therefore the lexicon type of resources from INL and the cultural heritage type data bases from Meertens Institute were taken as the set of resource types to try model with component metadata.

The metadata project started when no sophisticated component editor and registry were yet available and a preliminary rough XML-Schema and style-sheet toolkit was used. The resulting metadata components however, were imported in the current metadata component registry and can be used for creating new metadata schemas.

The results of the metadata project including a best practice guide are available at <http://www.clarin.nl/node/133>

More information about the CLARIN Metadata Infrastructure can be found at <http://www.clarin.eu/cmdl>



*Project coordinator: prof. dr. Hans Bennis*

In the CLARIN infrastructure the backbone is formed by a federation of centres that will offer well-described services to the research community, such as access to data resources, execution of web applications and services, support of infrastructure type of services and long-term persistency services. They need to support researchers in the way they want to use data and tools, i.e. they need to adopt the “business” model that is typical for humanities researchers and offer resources and services along the usage and expectation pattern in research.

Currently, four domain oriented centers have positioned themselves as main centres at the international CLARIN-EU level and followed the relevant assessment procedures:

- Institute for Dutch Lexicology (Leiden)
- Meertens Institute (Amsterdam)
- Max Planck Institute for Psycholinguistics (Nijmegen)
- Data Archiving Network Services (the Hague)

In addition, Huygens Institute (the Hague) has shown interest in becoming a CLARIN-NL centre in the near future as well.

Although there is no prescribed division of functions, each of the centers will have certain foci:

- Institute for Dutch Lexicology will focus on resources and services about the Dutch standard language;
- Meertens Institute will focus on dialectal and technological (a.o. popular songs/folk tales) resources and services for, mainly, the Dutch area;
- Max Planck Institute for Psycholinguistics will focus on multimedia/multimodal resources and services about minority, sign and similar languages;
- DANS will offer its EASY repository that can be used for resources not taken by the other centres and other infrastructure related services.

The CLARIN-NL Search & Develop project intends to provide a generic search mechanism on metadata and content in line with the efforts within the larger European framework.



---

This project addresses a number of issues:

- It will allow researchers with specific research questions to identify resources of interest, within the scope of resources made available within the project, through both metadata and content characteristics.
- It will allow researchers to search across resources located at multiple centers using a single user interface, rather than using the idiosyncratic interfaces developed for each individual resource.
- It will allow researchers to locate other, but related, resources by taking advantage of the semantic interoperability principles developed within CLARIN.

With respect to metadata the project makes use of the component based metadata framework and infrastructure already worked out in the European context. This will facilitate the metadata search, e.g. *'Give me all Dutch resources'*, limiting the potential number of candidates that will be passed to the relevant content search engines. Each of these content search engines is then further approached with specific content related questions, e.g. *'Give me all resources that contain the word "koe"'*. The examples provided here are simple, but will become more complex depending upon the types of resource and research questions. The results of the queries are combined, ranked and presented back to the user who will be able to inspect them or pass them on to further processing services such as for example developed in the TTNWW project.



*Project coordinator: prof. dr. Lou Boves*

Nowadays, there is a rapidly growing pile of recorded audio and video data that is difficult to browse without any tool to automatically create appropriate meta-data. Also potentially interesting old recordings are now hardly used because of the prohibitive manual effort to annotate these data. The lack of efficient tools for processing the data is especially awkward. However, analyzing large amounts of field work recordings is usually prohibitively time consuming and expensive.

AAM-LR addresses this problem. The algorithms developed in AAM-LR allow automatic massive data analysis and annotation of acoustic field data. The tools developed in this project are aimed at providing a semi-automatic analysis of field work corpora. Thereby they reduce the time needed for making linguistic analysis, bring effective usage of the data within the realm of what is feasible, and support an environment for linguists working in areas of language contact.

The AAM-LR project aimed at two goals: a proposal for a labeling system for audio data (with focus on data with very poor documentation), and the design of software that facilitate interpreting and annotating audio recordings. The software developed in AMM-LR provides a global phonetic annotation, based on alignments with acoustic phone models and by using phonetic features. The output can be fed into the ELAN/ANNEX editor, to facilitate further manual annotation. All annotations are conform ISOcat.

The AAM-LR deliverables are:

- Proposal of an extension for ISOCAT. Includes proposal for a new Thematic Domain Group.
- The content: algorithms for free and forced alignment and phonological features. This includes the interface of the algorithms with ELAN

### ***Proposal extension ISOCAT***

In the AAM-LR project we made a proposal (Peia Prawiro-Atmodjo et al., 2010) for entering and documenting the labels for audio (speech) recordings in the ISOcat Datacat Registry (<http://isocat.org/>). AAM-LR proposes a hierarchy of tiers, tier names and contents, specifically for audio and especially for speech, documented and illustrated with examples. It deals with high-level distinctions between, e.g., foreground speakers versus background speakers, speech versus background noise, speech versus singing, and phonetic annotation. The add-on label set covers the introduction of three main tiers (foreground speaker 1, foreground speaker 2, and background), each subdivided in pragmatically relevant sub-tiers (such as word, phone transcription, features, type of background sounds).



In order to test its usefulness in the ‘field’, the proposed AAM-LR scheme has been tested on real recordings from three databases in the DOBES archive (the database with audio and video recordings of endangered languages, kept at the Max Planck Institute of Psycholinguistics in Nijmegen). We focussed on audio recordings of endangered languages (Teop, Movima, and Kilivila/Trobiand).

### Software

The software developed in AAM-LR exists as sets of procedures in C and MATLAB, connected by Perl and C-shell scripts. The MATLAB scripts can also run by the public-domain Octave. For the speech decoding steps, HTK was used (but this can be replaced by an engine with less restricted public use). The input of the algorithms is an acoustic sdf file (wav file); the output of the algorithms are utf-8 text files. The output files can be imported into ELAN/ANNEX.

Here we provide an example of the data processing on Trobiand recordings. The Trobiand recordings are very noisy, available from cassette tape (early 80-ies). In one of the recordings, a singing group of women is clearly audible with many sounds of varying character in the background. Figure 2 illustrates an example of a phone decoding with the phones aligned with the acoustic signal. The tools actually provide a richer and more detailed output, not illustrated here.

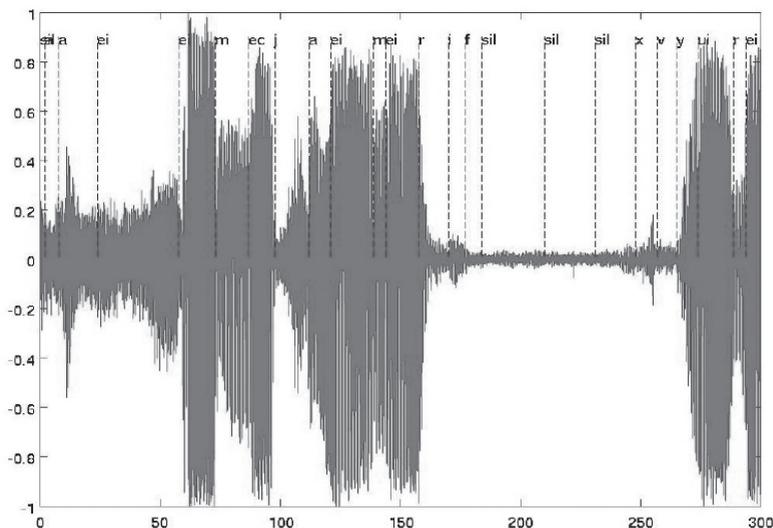


Figure 2. Alignment of a recording.

### Reference

Peia Prawiro-Atmodjo, Louis ten Bosch, 2010. Report on the AMM-LR project. Internal report RU Nijmegen. (Includes AAM-LR ISOCAT extension proposal.)



# Adelheid

## *A Distributed Lemmatizer for Historical Dutch*

---

*Project coordinator: dr. Hans van Halteren*

The Adelheid project provides a web-application with which end users can have historical Dutch text tokenized, lemmatized and part-of-speech tagged, after which they can inspect and where necessary manually correct the annotation. In the long run, the application is intended to allow the user to select the best resources for processing a specific text from those available in CLARIN, wherever they might reside. The project itself provides resources for 14<sup>th</sup> century charters and some test texts to demonstrate the potential of the system.

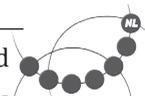
The need for such a system was clearly felt, since research on transcribed manuscripts from the early centuries of the use of the Dutch language, say up to the middle of the 19th century, is seriously hampered by a lack of standardized orthography. As an example, suppose a historical economist would like to know how prices for goods developed in the 14<sup>th</sup> to 16<sup>th</sup> centuries? This question would be very hard to answer when having only unlemmatized text available. In charters in the corpus Van Reenen-Mulder (van Reenen and Mulder, 1993) alone, the lemma *penning* shows 56 different spellings, all of which a researcher would have to guess in order to find the corresponding text passages. Clearly, the accessibility of the material is enhanced enormously by adding annotation layers showing normalized representations for word boundaries, lemmas and preferably also part-of-speech tags. The Adelheid system can provide these.

The starting point of the project was a tagger-lemmatizer system which was built by Hans van Halteren on the basis of the manually created tag and lemma annotations in the Corpus van Reenen-Mulder. On this type of text, the system reaches an accuracy of 94-95%. Inside the system, an input text is first split into tokens by a tokenizing component, then all possible lemma-tag combinations for each token are derived in a lexical component, after which the likelihood of each lemma-tag combination is estimated by a contextual component. Apart from the Adelheid system as a whole, each component can also be accessed directly from other systems through the internet as webservices.

The use of the system is shown by way of a demonstrator scenario in which the user can process example texts taken from the Corpus of the Count's scribes (Rem, 2003). Documentation is also provided on how to prepare other texts for processing by the system, as well as how to prepare additional lexica, e.g. containing lists of speci-

---

fic terms or proper names. The preparation of additional tokenization and contextual resources is rather more complicated, but by no means impossible, and interested parties are invited to contact Hans van Halteren if they are willing to build such resources. The Adelheid system is not meant to be a tool in isolation; it is at least as much intended as a catalyst for the scholarly community dealing with historical Dutch texts to provide not only texts in a standard format but also mutually compatible text analysis tools for Clarin.





*Project coordinator: dr. Charlotte Gooskens*

The study of linguistic variation — especially dialectal (geographical) variation, but also social variation of different sorts — has held a central position in linguistics for well over a century. The last two decades have witnessed enormous progress in the quantitative analysis, i.e., the automatic measurement of linguistic differences (dialectometry), which yields reliable and valid characterizations, e.g. when a hundred or so words are sampled at a few dozen or more sites (Goebel 2006, Nerbonne 2009, Nerbonne & Heeringa 2009). But dialectometry has not been widely used due to its demanding technical threshold, requiring special software installations, some of which have their own pre-requisites.

Gabmap<sup>1</sup> has been developed to make dialect analysis tools available to working dialectologists and other students of linguistic variation in an easy-to-use web application. In addition to dialectometric analyses, Gabmap generates various data summaries, supporting error detection in input data, providing researchers with useful overviews, and enabling the creation of distribution maps of any number of linguistic variables — words, morphological realizations, and also phonetic characters or patterns, depending on the user's data. In this respect Gabmap goes well beyond dialectometry, supporting the exploration of a large number of user-defined variables in different ways.

Gabmap allows linguists to upload their variationist data in different formats, but in particular, in the form of spreadsheets, which are popular systems for linguistic data collection and organization. Various overviews of the data are created automatically in order to support users who wish to explore freely. Tools are made available to support the creation of maps from Google Maps®, and also to convert different character encodings into Unicode IPA (UTF-8 or UTF-16), the “native” format in Gabmap.

The heart of Gabmap is the measurement of differences, which may be categorical (different lexical realizations of one concept or different forms of one affix), numerical (sets of formant frequencies for vowels), or stringbased (phonetic transcriptions). Although various options are supported, we attempted to identify sensible defaults for inexperienced users throughout.

---

<sup>1</sup> The development of Gabmap was supported by a grant from the CLARIN-NL program (<http://www.clarin.nl/>) to the ADEPT project (Assaying Differences using Edit Distance of Pronunciation Transcriptions), which we acknowledge gratefully.

---

Differences in linguistic items are then aggregated to obtain a robust characterization of the relations among the sites (or other groups of speakers), and these are analyzed and projected onto various sorts of maps to support scholarly investigation.



Because traditional dialectology emphasized dialect areas, i.e. areas of relative linguistic uniformity as the most important organizing element in dialectology, particular attention is paid to techniques for identifying natural groups (of sites) in data, examining them critically, and extracting the most representative and distinctive variables in them.

Gabmap has been presented as an active tutorial to groups at the University of Uppsala on Aug. 20, 2010; at the University of Amsterdam in Sept., 2010; at SprachRäume, in the series *Tagung des Forums Sprachvariation*, an Erlangen meeting of the *Gesellschaft für Angewandte Linguistik (GAL)* on Oct. 15, 2010; in Tvärminne (Finland), *Föreningen för nordisk filologis höstsymposium: Nya perspektiv på dialekter* on Nov. 26 2010; at the *Dialect dictionary of Finnish dialects and Dictionary of Swedish dialects in Finland*, at The Research Institute for the Languages of Finland in Helsinki on Nov. 30, 2010; and at the computer department at The Research Institute for the Languages of Finland, Helsinki on Dec. 1, 2010. Further tutorials are planned at the University of Potsdam on Dec. 7, 2010; as a workshop at the international conference *Methods in Dialectology XIV* in London, Ontario in Aug. 2011; and at a workshop on dialect geography arranged by the Society of Swedish literature in Finland and *Kungliga Gustaf Adolfs Akademien för svensk folkkultur*, Nov. 2011.

Gabmap is ready for use at <http://www.gabmap.nl/>.

#### References

- Goebel, Hans. 2006. "Recent Advances in Salzburg Dialectometry." *Literary and Linguistic Computing* 21(4):411–436. Spec. Iss., *Progress in Dialectometry: Toward Explanation* ed. J. Nerbonne & W. Kretzschmar, Jr.
- Nerbonne, John. 2009. "Data-Driven Dialectology." *Language and Linguistics Compass* 3(1):175–198.
- Nerbonne, John & Wilbert Heeringa. 2009. *Measuring Dialect Differences*. In *Theories and Methods*, ed. Jürgen Erich Schmidt & Peter Auer. *Language and Space* Berlin: Mouton De Gruyter pp. 550–567.



*Project coordinator: prof. dr. Jan Odijk*

The *En Garde* project, which was originally called *DUELME-LMF*, is a cooperation project between Utrecht University and the Institute for Dutch Lexicology (INL).

Its goal was to curate the DUELME database, i.e. to make an LMF-compatible version of it. The DUELME database is a database of app. 5000 multi-word expressions of Dutch, enriched with morphosyntactic and syntactic information, in accordance with the Equivalence Class Method for the lexical representation of multi-word expression, originally developed by Odijk (2004) and further elaborated by Gregoire (2010). It comes with a web interface that allows searching, browsing and editing in the database. The DUELME database was developed in the STEVIN programme and is available via the Dutch HLT Agency (“TST-centrale”)<sup>1</sup>.

In order to make the LMF-compliant version of this database, first an XML schema for the database compatible with LMF was developed. It turned out that LMF had several properties that made this difficult. Therefore we proposed some extensions of LMF to accommodate this database. When the XML schema had been defined, a converter was written to convert the original DUELME database into this XML format. Originally, it was the plan to also write a converter from the XML format to the original format, in order to ensure that the web interface to the DUELME database remained usable. Instead, however, we decided to develop a new web interface that interacts directly with the data in XML format (stored in an XML database).

A mapping has been defined between DUELME-specific data categories and existing or newly created ISOCAT data categories to ensure semantic interoperability of the curated resource with other resources and tools. Since a part of the DUELME data categories overlaps with data categories used in the Spoken Dutch Corpus (CGN), this was done in close collaboration with people working on the CGN data categories. An existing CMDI metadata description was adapted for the new database.

<sup>1</sup> <http://www.inl.nl/lexica/duelme>



---

A document has been produced in which limitations of and desiderata for the LMF standard, ISOCAT and other infrastructural aspects have been described. Furthermore, there is extensive documentation of the conversion software and of the web interface. The project has finished except that a metadata harvesting test still has to be performed. The DUELME database in LMF format will soon become available via the Dutch HLT Agency (“TST-Centrale”).

A number of presentations on the work being carried out in the *En Garde* project have been presented at various workshops<sup>2</sup>.

### References

- Grégoire, N. (2010), ‘DuELME: A Dutch Electronic Lexicon of Multiword Expressions’, *Language Resources and Evaluation* 44:1-2 pp. 23-39.
- Grégoire, N. (2010), “DuELME LMF project”. Presentation held at the Workshop “D-Spin meets RELISH: standardizing lexicon structures”, April 26, 2010, University of Frankfurt.
- Grégoire, N. (2010), “En Garde Project. The redesign of a Dutch Electronic Lexicon of Multiword Expressions.”, presentation held at the workshop “Lexicon Tools en Standards”, August 4, 2010, Max Planck Instituut, Nijmegen.
- Odijk, J. (2004), ‘A Proposed Standard for the Lexical Representation of Idioms’, in Geoffrey Williams and Sandra Vessier (eds.), *EURALEX 2004 Proceedings*, pp. 153-164. Lorient, Université de Bretagne Sud.
- Odijk, J. (2010), “The DUELME-LMF Project”, presentation held at the first CLARIN-NL project Meeting, Utrecht, Feb 19, 2010. See <http://www.clarin.nl/node/74>
- Odijk, J. (2010), “ISOCAT problems encountered at in DUELME-LMF”, presentation held at the ISOcat Workshop, Nijmegen, September 21, 2010
- Odijk, J. (2010), “Proposed solutions for problems encountered in DUELME-LMF”, presentation held at the ISOcat Workshop, Nijmegen, September 21, 2010

<sup>2</sup> See the references



# INTER-VIEWS

## *Curation of Interview Data*

---

*Project coordinator: dr. Henk van den Heuvel*

Interview data can be used in a number of ways, such as comparative research, re-study or follow-up study, re-analysis / secondary analysis, research design and methodological advancement, replication and validation of published work, and for teaching and learning.

Recent experiences with the re-use of interview data show that there is an enormous potential for this type of data in terms of multidisciplinary research, intersubjectivity and efficiency. Especially in the field of interview data related to the Second World War and other military conflicts several projects have successfully been carried out<sup>1</sup>.

In INTER-VIEWS the project partners (CLST, RU Nijmegen; DANS; Veterans Institute, Doorn; MPI, Nijmegen) have curated a corpus of 250 audio-interviews held with veterans of various conflicts, which can be searched through speech retrieval, and annotated and enriched by insights of the users through a so called *search and annotation tool*.

### ***The INTER-VIEWS database***

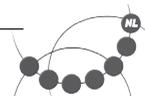
The INTER-VIEWS project has provided the curation of about 250 audio recorded life-history interviews with veterans, entailing:

- 120 World War II interviews which present a range of experiences of Dutch soldiers trained for and deployed in traditional warfare between 1935-1945.
- 100 interviews with veterans of the Dutch East Indies decolonization war between 1945 and 1950. This collection presents a large diversity of experiences at the local level in the domain of guerilla warfare.
- 30 interviews with veterans of the subsequent decolonization-war with Indonesia about the last colony New Guinea which ended in 1962. The interviews of this relatively unknown conflict present narratives about soldiers and local Papua-population left in uncertainty and isolation and the pressure of the international community to decolonize the area.

Each interview lasts between 1.5 and 2.5 hours.

The interviews are stored permanently by DANS and will be accessible for the research community. The audio is stored in wav-format. All interviews contain extensive metadata offered in a standardized (CMDI) format. This will facilitate search queries by researchers through the CLARIN webportal.

<sup>1</sup> <http://www.dans.knaw.nl/en/content/categorieen/projecten/getuigen-verhalen-telling-witnesses>  
<http://www.dans.knaw.nl/en/node/1233>



Relevant metadata include (but are not restricted to):

- Interview subject
- Information about the mission or conflict
- Time period of the veteran's mission or conflict
- Extensive Information about the interviewee
- Information about the interviewer
- Information about the audio, summary and transcripts

All interviews contain extensive textual summaries divided into time blocks of about 10 minutes, and topic keywords also arranged per 10 minutes. This information is stored in the metadata as well.

### The INTER-VIEWS annotation tool

In order to enlarge the search potential in the content of the interviews and to enrich these sources with the insights of the consultants a search and annotation tool has been developed. Access to this tool is arranged through a password protected website and this link: <http://www.dans.knaw.nl/content/categorieen/projecten/oral-history-annotatietool>

All interviews are treated by using automatic speech recognition so that relevant keywords are spotted in the audio. Some 10 interviews have been manually transcribed.

By entering keywords the user of the tool can find relevant sections in the interview, view the metadata and link directly to the audio. He or she can further add annotations to selected fragments of the interviews and upload files (pictures etc.) to illustrate the annotation. These annotations can (but need not) be made visible to other researchers

as well. In this way a wiki like working environment can be created containing a rich combination of raw interview data (audio, transcripts), metadata, and annotations. The project team is collecting funding to organize workshops for researchers of various disciplines in order to make them acquainted with this tool and to further tailor it to their needs.

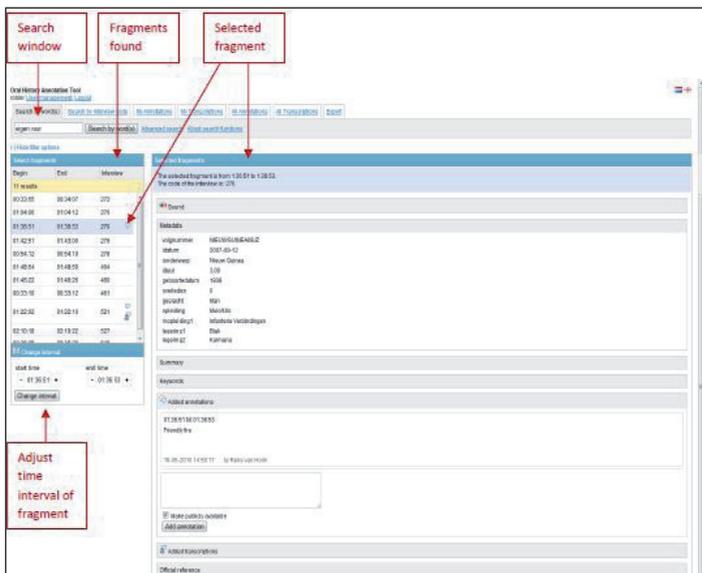


Figure 3. Search and annotation tool of interview data.



*Project coordinator: prof. dr. Sjef Barbiers*

The MIMORE tool enables researchers to investigate morphosyntactic variation in the Dutch dialects by searching three related databases with a common on-line search engine. The search results can be visualized on geographic maps and exported for statistical analysis. The three databases involved are DynaSAND, DiDDD and GTRP.

The data in DynaSAND, the dynamic syntactic atlas of the Dutch dialects<sup>1</sup>, were collected between 2000 and 2005 by oral interviews (fieldwork and telephone) in about 300 locations across The Netherlands, Belgium and a small part of north-west France. Dialect speakers were asked to judge and/or translate some 150 test sentences. DynaSAND makes available the full recordings and transcriptions of these interviews. Together, the DynSAND data cover the syntactic variation in the Dutch language area in the left periphery of the clause (the complementizer system and complementizer agreement), variation in subject pronoun form depending on syntactic position, subject pronoun doubling, cliticization on YES/NO, the reflexive system, fronting constructions (Wh-clauses, relative clauses, topicalization), word order and morphological variation in verb clusters, negation and quantification.

The data in DiDDD (Diversity in Dutch DP Design; <http://www.let.uu.nl/~Huib.Kranendonk/personal/>) were collected between 2005 and 2009 with oral and written interviews in about 200 locations in the Dutch language area, with a methodology highly parallel to DynaSAND. The data involve translations of and judgements on test sentences. For 30 interviews there are sound recordings which have been lined up with their transcriptions. The DiDDD data cover the morphosyntactic variation within nominal groups, in particular possessives, partitives, noun ellipsis, the demonstrative system, the numeral modification system, what-for constructions, quantitative *er*, adjectival inflection, negation and exclamatives.

The data in GTRP (Goeman, Tældeman, van Reenen Project)<sup>2</sup> were collected between 1979 and 2000 with oral interviews in about 600 locations in the Dutch language area. Informants were asked to translate words or short sentences. Parts of the transcriptions have been lined up with the sound recordings. The morphological data in GTRP include plural forms of nouns, diminutives, gender on nouns and adjectives, comparatives, superlatives, verbal inflection including participles, subject, object and possessive pronouns.

<sup>1</sup> <http://www.meertens.nl/sand>

<sup>2</sup> <http://www.meertens.nl/mand/database>



With the MIMORE search engine one can search these three databases simultaneously, with text strings, part of speech tags and syntactic variables. The researchers can combine categories and features into complex tags (see figure 4) or use predefined tags. All categories and features are linked to the ISOCAT-standards. Since all sentences have a location code, the morphosyntactic phenomena found in a set of sentences resulting from a search can be automatically plotted on a geographic map. It is possible to include more than one morphosyntactic phenomenon in one map, thus visualizing potential correlations between these phenomena. There is also a user-friendly function to export the data for external use, e.g. a statistical program.

The importance of combining the three databases can be illustrated with the following example. The GTRP database includes data on attributive possessive pronouns in singular and plural nominal groups and possessive pronouns in predicative position. The DiDDD database provides data on possessive pronouns combined with nominal possessors as in *Piet z'n auto* 'Piet his car' and *de bakker z'n auto* 'the baker his car', with noun ellipsis *de bakker z'n* 'the baker his' and with questions words (*wie zijn auto*). DynaSAND contains information on possessive pronouns in complex reflexives, e.g. *zijn eigen* 'his own'. By combining these data one can derive the complete possessive paradigm for each dialect that occurs in all three databases, including possessive inflection.

Geographic Conditions : no restrictions  
Resources & Limits : no restrictions  
Text : no restrictions  
Tags : N(1,pl)

Categories		Features					
<input type="checkbox"/>	Infmrk	aux	infin	fin	pres	past	pp
<input type="checkbox"/>	V	mod	imp	1	2	3	sg
<input type="checkbox"/>	D	pl	def	indef	m	f	neut
<input type="checkbox"/>	Adv	neg	comp	sup	dim	quant	pers
<input type="checkbox"/>	Conj	poss	refl	recipr	wh	rel	subord
<input type="checkbox"/>	P	coord	asp	caus	pass		
<input type="checkbox"/>	Part						

Output : table with results

Figure 4. MIMORE search engine.



*Project coordinator: dr. Onno Crasborn*

Tools for the transcription, annotation and exploitation of sign language resources are a relatively recent phenomenon. With the venue of digital video in desktop computers in the 1990s, it became possible for the first time to attach time-aligned annotations to video recordings. Multimedia annotation tools were developed in part on the basis of experiences with text and speech technologies in the decades before. ELAN is one of such tools. In the past ten years, it has become an indispensable tool for linguists working on signed languages and can be considered a de facto standard. Currently, ELAN is used for the annotation of most of the sign language corpora that are slowly starting to be created in different countries. The *Corpus NGT*, containing 72 hours of newly recorded dialogues of Sign Language of the Netherlands (abbreviated NGT, NederlandseGebarentaal), is one of such new corpora. It is used for both linguistic research and for the development of sign language technologies.

Lexicons of signed languages have seen a quite different development. They have a much longer history, starting with paper dictionaries for several languages, including NGT, in the 1970s and 1980s. The possibility of desktop video for short sequences in the 1990s enabled the production of dictionaries on CD-ROM and later on DVD. Currently, the Dutch Sign Centre produces DVDs side-by-side with online dictionaries for NGT. While being based on linguistic and lexicographic knowledge, they have emerged outside the broader context of language technology and the tools available there, in part because of the lack of integration of video.

Until now, there was no lexicon function of any kind built into ELAN. This makes it hard to annotate lexical items in continuous signing in a consistent manner, or to display properties of lexical items that would typically be coded in a lexicon. The SignLinC project aimed to make the standard CLARIN format LMF and the associated tool LEXUS of use to users of ELAN. In addition, a start was made with the linking to specific segments of annotation files from within LEXUS. This allows for the inclusion of examples of lexical forms in a specific context. The project created such a linking function in the form of URLs that link to the display of archived annotation files in ANNEX, the online browser for ELAN annotation documents. Further, two types of functions were added in ELAN. The possibility to link to an 'External Controlled Vocabulary' (and accompanying changes in the interface) allow for the use of very large lists of

---

words as controlled vocabularies. Secondly, a pane in the interface allows for the search and display of lexical entries that are stored in an online LEXUS lexicon. A subset of the lexicon of the Dutch Sign Centre has been converted to LEXUS to serve as a test set.



Variability in the choice of Dutch glosses for signs has been an area of major concern since the creation of the Corpus NGT in 2006-2008. The first link that is now created between ELAN and LEXUS is expected to dramatically improve the consistency of the lexical tagging of signs.



# TDS Curator

## A web-services architecture to curate the Typological Database System

---

*Project coordinator: dr. Alexis Dimitriadis*

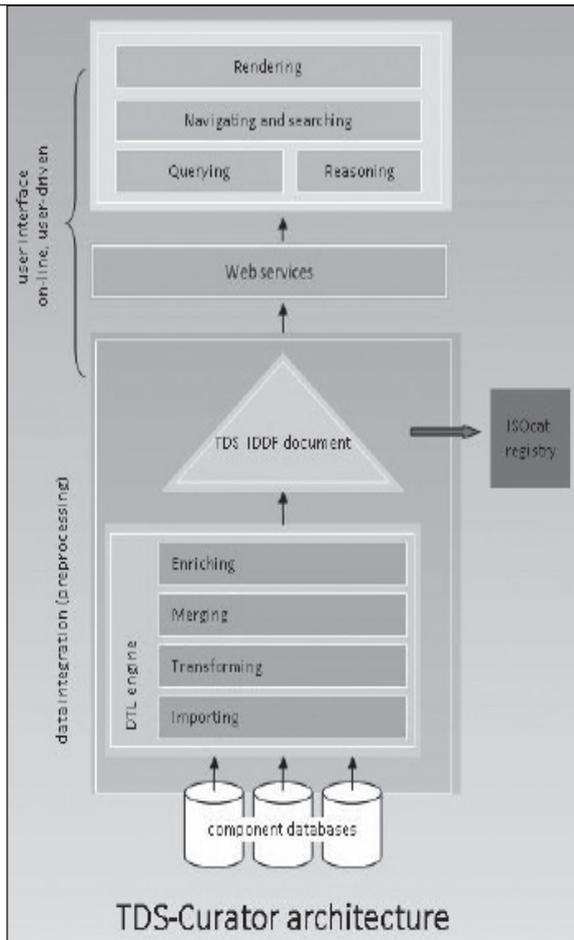
The Typological Database System (TDS) provides integrated access to multiple independently developed typological databases through a common web interface. Language typology, the study of the range of language variation and universals, is a data-intensive discipline that increasingly relies on electronic databases. Improved availability of the data collected in the TDS will enhance its potential to support linguistic research.

While the TDS is currently operational, TDS Curator will turn it into a sustainable service that conforms to CLARIN infrastructural requirements. A web services architecture that properly separates the user interface from the data access layer will make it possible for the latter to be queried by other components of the integrated environment envisaged by CLARIN. In this way, the data contained in the TDS will be transparently available to other tools and resources; the user interface itself would also be usable with other compatible data resources. The TDS Curator server will be hosted by DANS as an empirical test case for DANS serving as a Clarin A/B centre.

### **Data**

The TDS is currently comprised of approximately fifteen independently developed databases, originally created in a multitude of computer platforms and design structures. Together, they define well over a thousand data fields, which provide (partial) data for over a thousand languages. All data is freely accessible and contributed to the TDS by the creators of the component databases. The integration scripts are developed by TDS project members in consultation with the creators of the original databases, who provide the information needed to fully document the structure and linguistic meaning of the data. At the end of the project, all original data files will comply with the requirements of the international Data Seal of Approval (see: <http://www.data-sealofapproval.org/>).

The TDS data integration process does not manually modify the source databases; an integration script is created that imports the data, restructures and standardizes it in strictly delimited ways, and enriches it with documentation and other auxiliary information. If a component database is updated in the future, the new version can be imported and re-integrated with minimal manual intervention. The transformed data and documentation is stored in a self-describing, XML-based format developed by the project: The Integrated Data and Documentation Format (IDDF). The generated document is then queried using the TDS Curator interface.



## Interface

The TDS data collection can be queried like an ordinary database, with the caveat that the diverse sources of information mean that the data is not statistically balanced. Because of the great number of available data fields, users first search for data fields (attributes) relevant to their goals; these are then used to construct and execute a query over the data. The system can be used to help answer questions such as “what languages have the basic word order Verb-Object-Subject”, “what kind of phonological stress systems are common”, “are languages with subject-verb agreement more likely to allow null subjects than languages without it”, etc. The system is not an oracle: In all cases, only partial information is returned, as collected and deposited in the system by the creators of the component databases. But this information can be valuable to other researchers, either as a complete answer to a specific question or as the starting point for further research.



# TICCLops

## *Text-Induced Corpus Clean-up online processing system*

---

*Project coordinator: Martin Reynaert*

The Humanities are currently undergoing a profound paradigm shift in that paper, for many centuries the major means of disseminating and safeguarding knowledge, is gradually being replaced by digital means. This transition is not painless. Anyone who has seen the result of the digitalisation of previously paper-locked printed text by means of Optical Character Reading (OCR) knows that this is a process which may introduce a lot of undesirable noise. What may be perfectly readable on paper for humans, often proves not to be perfectly readable to the machine and the result is less than perfect electronic text.

At Tilburg University we have for years worked on better spelling error correction techniques for large amounts of text. We have gradually developed a system called Text-Induced Corpus Clean-up or TICCL to post-correct the electronic text obtained by OCR. It works fully automatically, because some digitised text collections are simply huge. Think about the eight million page selection of historical newspapers digitised by the Koninklijke Bibliotheek (KB), the National Library in The Hague, or the millions of books being digitised in libraries around the world by Google. As partner in this CLARIN-NL project, the KB has provided us with our demonstrator corpus: a book printed in 1789, recently digitised. Our second CLARIN-NL project partner, the Institute for Dutch Lexicology (INL), has produced the 'perfect' text for this book, i.e. the text as should have been recognised by the OCR.

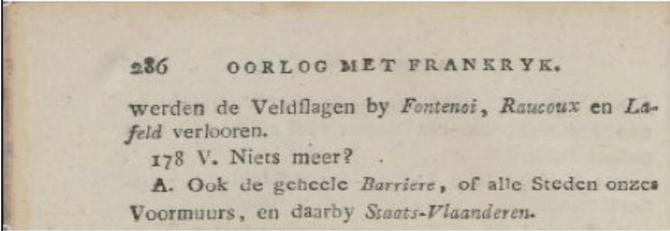
In many disciplines, in great numbers of research institutes, smaller more specialised digitalisation programs are underway or being planned.

In order to offer these researchers the chance to improve the results of their own digitalisation efforts, we have made TICCL into a web application and web service. At CLARIN Centre INL in Leiden, the TICCL online processing system or TICCLops is now available.

TICCLops addresses any kind of lexical variation and does not currently in its output distinguish between kinds of variation due to cause of origin. It effectively links morphologically related word forms, word forms due to historical spelling variation or variation caused by error, whether typographical, typesetting or OCR errors.

Although TICCL performs very well, it is nevertheless still under development. One should not currently expect any unsupervised system to correct words that have been very badly misrecognised. One may expect it to accurately correct words where the misrecognition is limited to two, perhaps three, characters. These cases typically account for about 90% of the variation actually present in an OCR-ed corpus.

1. Print version: image format



2. OCR version: xml format

```
<?xml version="1.0" encoding="UTF-8"?>
<text>
<p>£86 OORLOG MET FRANKRYK.</p>
<p>werden de Veldflagen by Fontenoi, Raucoux en Lafeld verlooren.</p>
<p>178 V. Niets meer? .</p>
<p>A. Ook de geneele Barriere, of alle Steden onzes Voormuurs, en
daarby Staats-Vlaanderen.</p>
```

3. TICCLops version: enriched xml format

```
<?xml version="1.0" encoding="UTF-8"?>
<text>
<p> £86 OORLOG MET FRANKRYK. </p>
<p> werden de <sc variants="veldslagen">Veldflagen</sc> by <sc
variants="fontenot;fontenoy">Fontenoi,</sc> Raucoux en Lafeld <sc
variants="verlooren">verlooren.</sc> </p>
<p> 178 V. Niets meer? . </p>
<p> A. Ook de <sc variants="geheeel">geneele</sc> Barriere, of alle
Steden onzes <sc variants="voormuur">Voormuurs,</sc> en daarby
Staats-Vlaanderen. </p>
```

Figure 6. Excerpt from the TICCLops demonstrator based on a 1789 book by Martinet digitized by the Koninklijke Bibliotheek - The Hague.

One of the major assets of TICCL is that it always actively uses the vocabulary of the corpus being processed. Validated lexicons are hard to get. In fact, we do not currently have a validated lexicon for the era in which the demonstrator book was written. The system can very well be run without any validated lexicon, in which case it uses the vocabulary of the corpus to be processed as its lexicon. Currently geared towards Dutch, it should perform well on other languages, too.



# TQE

## Transcription Quality Evaluation

Project coordinator: dr. Helmer Strik

The Transcription Quality Evaluation (TQE) tool is an instrument that automatically evaluates the quality of phonetic transcriptions. The application makes it possible to upload pairs of files consisting of an audio file and a transcription file and process them as follows: the audio signal and the phonetic transcription are aligned, segment boundaries are derived for each phone, and for each segment-phone combination it is determined how well they fit together, i.e. for each phone a TQE measure (a confidence measure) is determined, a number ranging from 0-100%, indicating how good the fit is, i.e. the quality of the phone transcription (see Figure 7). The higher the number, the better the fit. The output of the TQE tool consists of a TQE measure and the segment boundaries for each phone in the corpus (see Figure 7).

The transcription of the utterance shown in Figure 7 is 'd@wOlfmAn'. It can be observed that except for the @, all the TQE scores are higher than 90%. Subsequently, we deliberately introduced artificial errors in the transcription by swapping the A and O symbols (see Figure 8). Again the audio file and the new transcription files were processed by the TQE tool. The results can be seen in Figure 8. It appears that the TQE scores for the segments for which the phone symbols were swapped are lower now.

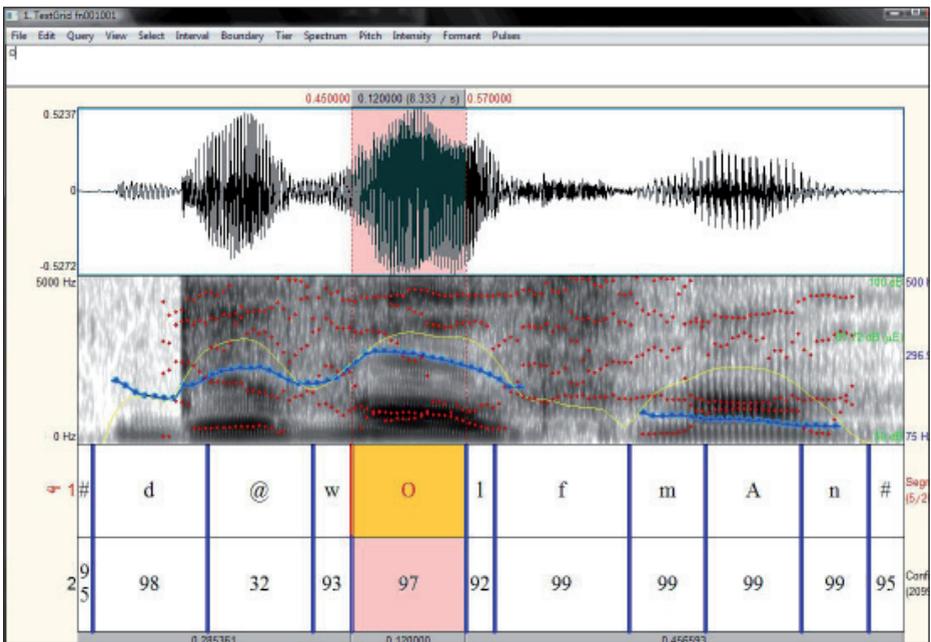


Figure 7. Screenshot of the results of TQE for a correct transcription.



In addition, the scores for the segments in the neighborhood of these erroneous segments are also lower. The reason is that swapping the phone symbols also influences the alignment calculated with the TQE tool, i.e. the boundaries of the segments change, and thus also the TQE scores.

The TQE tool thus makes it possible to find (sequences of) segments for which the match of the phone symbols with the audio signal is not optimal, in other words, the TQE tool can be used to check the quality of phonetic transcriptions. This can be useful for validating (manual) phonetic transcriptions, but also to compare and select ('competing') transcriptions, e.g. to study pronunciation variation. The TQE tool can thus be usefully applied in all research - in various (sub-)fields of humanities and language and speech technology (L&ST) - in which audio and PTs are involved.

Coordinator of the TQE project is Helmer Strik; and the other partners in the TQE project are Daan Broeder of the Max Planck Institute for Psycholinguistics (MPI), and Remco van Veenendaal of the Institute for Dutch Lexicology (Instituut voor Nederlandse Lexicologie, INL). Other people that have been involved in the project so far are Joost van Doremalen, Catia Cucchiari, Robin Oostrum, Robin Rutten, en Ferdy Hubers of the RU, Peter Withers and Tobias van Valkenhoef of the MPI, and Laura van Eerten of the INL.

For more information see <http://lands.let.ru.nl/~strikr/research/TQE/>

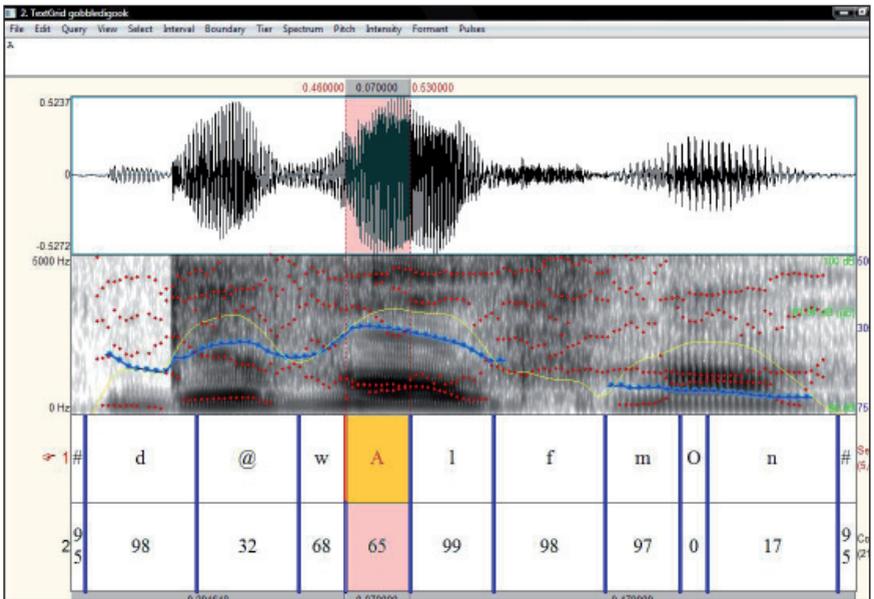


Figure 8. Screenshot of the results of TQE for a transcription with artificial errors.



*Project coordinator: drs. Hindrik Sijens*

The project's goal was the technical embedding of the *Wurdboek fan de Fryske Taal* (Dictionary of the Frisian Language, henceforth *WFT*) into in the Integrated Language Database of Dutch (GTB) at the Institute of Dutch Lexicology (INL) in Leiden.

In 1984 the Fryske Akademy in Leeuwarden (The Netherlands) published the first volume of the *WFT*. This dictionary is a scholarly, descriptive dictionary for the Modern West Frisian language. It has been published in 25 volumes of 400 pages each and contains about 115,000 lemmas. The dictionary entries provide the user with information about the spelling of the headword, its part of speech, its pronunciation and information about inflection and etymology. The semantic section shows information about the meanings of the headwords by means of definitions in Dutch or translations into Dutch. The idioms' section contains collocations and proverbs. The article concludes with compounds and derivations belonging to the headword.

Only 500 copies have been printed of each volume, and it can be assumed that the target audience is much bigger than the subscribers to the dictionary. The central question was how to reach a broader audience and how to provide them with as much linguistic information about Frisian as possible.

Making the dictionary available online was the most obvious option to reach this goal. The database structure enables thematic search operations. And matching the *WFT* with Dutch historical dictionaries included into the GTB enables comparative studies with Dutch material.

For this purpose data accessibility had to be enhanced by explicit tagging of information categories which can be exploited by a retrieval application. For this reason two projects were defined: The data curation of the *WFT* database and integration of the data into the dictionary web application of the GTB. The original dictionary data were stored in a BRS-database with almost no metadata added. BRS is a fully-inverted indexing system to store, locate, and retrieve unstructured data. In the curation project the original data were transferred into XML with TEI-annotation (TEI = *Text Encoding Initiative*, a standard for the representation of texts in digital form). Some preliminary work had to be done to achieve this: additional work on the *WFT*-data, correcting mistakes and inconsistencies, a set of metadata had to be defined, a dictio-

nary entry parser to convert the data to XML had to be developed. Before the TEI-annotated WFT could be integrated into the GTB, a list of word classes and an index of sources and references had to be compiled. To implement the WFT data into the GTB application, additional adaptations had to be made.



The back end (search layer) for retrieval in scholarly dictionaries is fully documented as a web service. This contributes to the establishment of a CLARIN standard for encoding of and retrieval in scholarly dictionaries as a linguistic resource.

The dictionary application was launched by her majesty Queen Beatrix on July 6<sup>th</sup> 2010. The project was supported by CLARIN-NL and carried out by the Fryske Akademy in Leeuwarden (<http://www.fryske-akademy.nl>) and the Institute of Dutch Lexicology in Leiden (<http://www.inl.nl>).

The URL of the dictionary application is: <http://gtb.inl.nl/>

## TSJELK

Woordsoort: znw. m./v.  
Modern-Nederlandse lemmavorm: **kelk**  
Uitspraak: **tsjɛlk**  
Datering: 1858—  
Flexie: Plur. **tsjelken**  
Dim. **tsjɛlkje**;

Varianten: **tsjilk**, Uitspraak: **tsjilk**, Datering: 1911—, Flexie: Plur. **tsjilken**; Dim. **tsjilkje**;  
**tsjelke**, Accent: 'tsjelke, Datering: 1911—, Flexie: Plur. **tsjelken**.; Dim. **tsjɛlkje**.

Etymologie: *Nederlands kelk, Latijn calicem (calix).*

Kelk.

- 1. van onder puntig of smal, naar boven wijd uitlopend glas of zo'n beker op een hoge voet.
  - ↪ **Figuurlijk**: in verbindingen meestal ontleend aan de vergelijking van het lijden van Jezus met een beker gevuld met een bittere drank.
    - Myn heit, as 't wæze mei, lit dizze tsjɛlk my foárbygéan. J.H. HALB, *ewangeelje*, 119, [1858].
    - By al de freugde fen de lêste wiken, is dit it dripkje alsem yn 'e tsjelke. J.P. WIERSMA, *oars*, 73, [1927].
- ↪ 2. buitenste krans van bloembekleedsel. (*plant*).

Samenstellingen: arons-, blom-, drinkens-, dúbel-, ljens-, mied-, mis-, nachtmielstsjɛlk.

Figure 9. Page with the entry *tsjelke* ('goblet', Dutch *kelk*)



*Project coordinator: prof. dr. Wijnand Mijnhardt*

A consortium of Dutch universities and cultural heritage and research institutions is building a web-based collaboratory (an online space for asynchronous collaboration) around a multilingual corpus of 20.000 letters of scholars who lived in the 17th-century Dutch Republic to answer the research question: how did knowledge circulate in the 17<sup>th</sup> century? Hereto, it will be necessary to analyze this large amount of correspondence systematically. Based on this (extendable) corpus, we will implement a content processing workflow that consists of iterative cycles of conceptual analysis, enrichment with several layers of annotation, and visualization.

With advice from CLARIN-EU and financed by CLARIN-NL, in the first stage of the project a demonstrator has been developed which implements techniques of topic modeling. Tools we have used to create this demonstrator include tokenizing, language recognition, stemming and LDA (Latent Dirichlet Allocation; Mallet implementation). Results are presented in a faceted search interface, allowing the researcher to create selections of letters, filtered by topic, correspondent, location and language. Besides the list of selected letters, the results are shown in a topic cloud, plotted on a map and on a time line. The personal network of a correspondent can also be visualized.

The second stage consists of evaluating existing more complex tools en techniques that can tackle one or more aspects of the targeted grammatical, content-related, and network complexity analysis, annotation, and visualization. In this phase a set of tools that can be readily utilized in CKCC shall be identified, as well as tools that need to be adapted or extended to the needs of CKCC; by the end of this phase, resources, requirements and risks shall become clear.

In the third stage the collaboratory will be developed further according to the description in the CKCC project goals, centering around the technique of concept extraction.



---

These three stages constitute the Work Package Analysis Tools, the core of the CKCC project, for which the support of CLARIN-NL was requested. Other Work Packages provide data and software tools needed to create a complete system:

- the editing collaboratory that will contain the letters (WP1);
- the archiving environment for data and software (WP2);
- the digital corpus of letters (WP6). The corpus consists of the correspondences of Hugo de Groot, Constantijn Huygens, René Descartes, Christiaan Huygens, Anthoni van Leeuwenhoek, Jan Swammerdam and Caspar Barlaeus. Main languages are Neolatin, French and Dutch, and a significant amount of letters is multilingual.

The CKCC project is a cooperation of the Descartes Centre of the University of Utrecht, the Huygens Institute – KNAW, DANS (Data Archiving and Networked Services) – KNAW, VKS (Virtual Knowledge Studio for the Humanities and Social Sciences) – KNAW, the National Library and the University of Amsterdam.

End date of the project is November 2012.

More information:

- <http://ckcc.huygens.knaw.nl/> (in English)
- <http://www.huygensinstituut.knaw.nl/ckcc-“geleerdenbrieven”> (in Dutch)



Project coordinator: Marc Kemps-Snijders

The transnational Dutch (CLARIN-NL) and Flemish (CLARIN Flanders) project TTNWW aims to integrate existing natural language processing, speech technology and computational linguistic applications which have been developed over the past years within projects such as CGN and STEVIN into web services workflows. This cooperation has already been in existence for over 10 years and is continued in the TTNWW project on the common language, Dutch. The project is situated in the larger European framework of CLARIN and aims to make both resources and technology available to HSS researchers with little or no technical background. Emphasis lies on storage and availability of data and services on a large scale, compatibility of data formats and interaction between tools. This should enable researchers to address research questions better or more easily and provide opportunities for formulating new types of research questions, i.e. research questions that previously could not or not adequately be addressed by the CLARIN community.

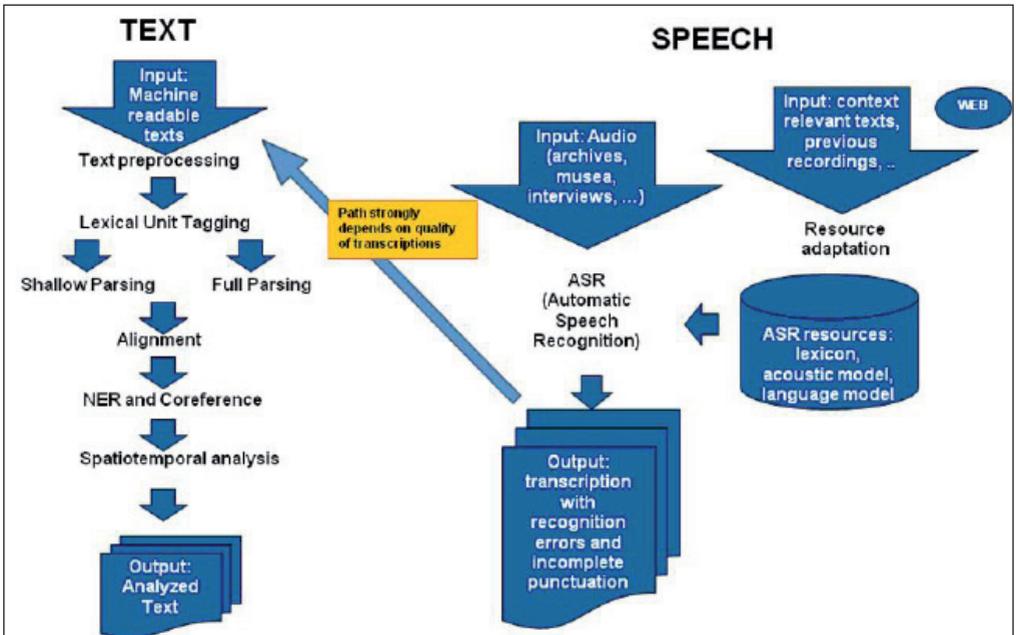


Figure 10. Overview of workflows created in TTNWW



---

The project comprises of two modalities: text and speech. For texts, workflows are created combining tools such as TICCLOPS, TADPOLE, ALPINO, MBSLR and STEx but also tools for text alignment, part of speech recognition and coreference to answer research questions on historical texts, forum discussions, translated novels and archeological data. For speech data, speech recognizers are used to make spoken data more easily accessible by creating searchable time aligned transcriptions. All results will be made available online on servers of recognized CLARIN centers.

Participating organizations to TTNWW include Tilburg University, Katholieke Universiteit Leuven, University of Antwerp, University College Ghent, Utrecht University, Meertens Institute, Institute for Dutch Lexicology, Huygens Institute, Radboud University, Twente University, Documentatie- en Onderzoekscentrum voor Religie, Cultuur en Samenleving, Katholiek Documentatie Centrum and Alletta.



---

## Call 2 Projects starting up in 2011

---

### **ArthurianFiction**

(*Arthurian Fiction in Medieval Europe*) is a mixed demonstrator and data curation project that aims to carry out data curation for two databases with data for literary research in the area of European Arthurian fiction and to create a demonstrator that will allow users access to the data for searching and editing.

### **C-DSD**

(*Curating the Dutch Song Database*) is a data curation project that aims to curate the Dutch Song Database (DSD), a database in the field of Literary Studies containing (meta-) data on 140.000 songs and their 15.000 sources from the Middle Ages to the present day.

### **COAVA**

(*Cognition, Acquisition and Variation Tool*) is a mixed demonstrator and data curation project proposal in which a tool is developed for easily exploring the linguistic characteristics of objects from two databases of language variation (historical dialectology) on the one hand and three first language acquisition databases on the other hand.

### **INPOLDER**

(*Integrated Parser and Lemmatizer Dutch in Retrospect*) is a mixed demonstrator and data curation project that aims to provide syntactically analysed corpus material for historical Dutch. It uses the Adelheid tagger for PoS-tagging and a variant of the Penn-Helsinki parser for historical texts for enriching the texts with parse trees.

### **IPROSLA**

(*Integrating and publishing resources on sign language acquisition*) is a resource curation project that aims to integrate two different data sets on sign language acquisition by documenting the two data sets with CMDI, and archiving them at the MPI language archive. The data consist of a set of longitudinal data of deaf children from deaf and hearing parents collected at the UvA, and a new collection of longitudinal data collected at the RU from hearing children of deaf parents.



---

## **NEHOL**

(*Negerhollands Database*) is a data curation project that aims to make available to the CLARIN community the data from the Dutch-lexifier Creole language Negerhollands, the now extinct Creole language of the Virgins Islands.

## **VU-DNC**

(*VU Diachronic Newspaper Corpus*) is a data curation project that aims to make a unique diachronic corpus of Dutch newspaper articles from five major Dutch newspapers from 1950/1951 and 2002 (2 MW) available to humanities researchers, to extend the discourse annotation with additional lexico-grammatical features and to create a gold standard benchmark for OCR-post-correction tools, all of this in a fully CLARIN-compliant manner.

## **WAHSP**

(*Web-application for historical sentiment mining in public media*) is a demonstrator project that aims to provide advanced forms of text mining (more specifically, sentiment mining) in large historical datasets of newspapers and journals in the form of a CLARIN compliant web-application, addressing research questions of historians and policy researchers.

## **WIP**

(*War in Parliament*) is a mixed demonstrator and data curation project that aims to make the Dutch Hansard database compliant with the CLARIN infrastructure and to provide an advanced search engine for it in order to make it possible to carry out historical and social science research, in particular with respect to the Second World War.







## Organisation and Contact

The CLARIN-NL executive board executes the daily policy as established by the board. The executive board is advised by a national and an international advisory panel.

### ***Executive board:***

Jan Odijk (Utrecht University)

Hans Bennis (Meertens Institute – KNAW)

Arjan van Hessen (University of Twente)

Daan Broeder (Max Planck Institute, Nijmegen)

### ***Chair of the board:***

Geert Booij (Leiden University)

### ***Contact:***

Address:

CLARIN-NL Office

Trans 10

3512 JK Utrecht

Telephone: +31 30 253 6279

Fax: +31 30 253 6000

Website: [www.clarin.nl](http://www.clarin.nl)

E-mail: [clarinnl@uu.nl](mailto:clarinnl@uu.nl)