Nederlab: Towards a Single Portal and Research Environment for Diachronic Dutch Text Corpora

Hennie Brugman[1], Martin Reynaert[34], Nicoline van der Sijs[12], René van Stipriaan[1] , Erik Tjong Kim Sang[1], Antal van den Bosch[2]

Meertens Institute Amsterdam[1], CLS[2] – CLST[3] / Radboud University Nijmegen, TiCC / Tilburg University[4],
hennie.brugman@meertens.knaw.nl, reynaert@uvt.nl, post@nicolinevdsijs.nl, stipriaan@planet.nl,
erik.tjong.kim.sang@meertens.knaw.nl, a.vandenbosch@let.ru.nl

Abstract

The Nederlab project aims to bring together all digitized texts relevant to the Dutch national heritage, the history of the Dutch language and culture (circa 800 – present) in one user friendly and tool enriched open access web interface. This paper describes Nederlab halfway through the project period and discusses the collections incorporated, back-office processes, system back-end as well as the Nederlab Research Portal end-user web application.

Keywords: Dutch, Diachronic Corpora, Online Portal, Nederlab

1. Introduction

Many digital Dutch text collections, from the eighth century to the present day, are available for digital humanities researchers. These texts reflect the dynamic development of Dutch language and culture. However, these collections typically are hosted by different institutions, are described with different metadata, and cannot be searched simultaneously. Advanced software tools necessary for the analysis of the texts in these collections are available, but are in most cases specific to particular goals and formats, rather than being generally applicable.

The Nederlab project is building a comprehensive research corpus that brings together collections from various sources in the Netherlands and Flanders. It also brings together relevant tools for processing, searching and analyzing these data, in one virtual research environment[1] that primarily supports historians, literary scholars, and linguists. The specific goal of Nederlab is to discover patterns of change over time and space.

Collections that are aggregated and integrated into Nederlab are processed using a carefully designed Nederlab collection pipeline (see Section 4.). This pipeline is applied to a number of basic resource types (several types of titles, persons/authors). It deals with metadata, text content, some vocabularies (e.g. genres, places) and linguistic annotations. It first aggregates data, then maps and converts it to a limited number of pivot formats, it identifies and links related resources. The pipeline involves an editorial team that is responsible for data selection, quality assessment and manual and semi-automatic data curation and interlinking. For texts that originate from Optical Character Recognition automatic spelling correction, OCR post-correction and text normalization is applied.

To all texts we apply sentence splitting, tokenization, lemmatization, part of speech tagging, and named entity recognition, resulting in massive amounts of linguistic annotations. This uniform annotation treatment enables us to obtain the necessary statistics for word tokens and types, lemmata, PoS tags and entities across all incorporated corpora. Federated Search (Stehouwer et al., 2012) across non-uniformly annotated corpora simply cannot provide this.

Metadata, text and linguistic annotations are indexed in one large, powerful search index. Finally, a broker orchestrates searching and other data services, and provides a uniform web-based access point to all Nederlab collections for end-user tools. The most important of these tools is Nederlab's own virtual research environment (see Section 6.).

Section 2. of this paper introduces the Nederlab project. Section 3. discusses the wide range of collections that play a role in the project. Section 4. presents our collection pipeline, it describes how we integrate collection data into the Nederlab Portal. Sections 5. and 6. discuss scientific usage scenarios and what part of those are already available in the Nederlab research environment. Section 7. presents a number of additional web services that we built, and that are valuable services in other contexts as well.

2. The Nederlab project

The Nederlab project started in 2013. It aims to bring together all digitized texts relevant to Dutch national heritage, in particular to the history of Dutch language and culture, in one user-friendly and tool-enriched open access web interface, allowing scholars to simultaneously search and analyze data from texts spanning the full recorded history of the Netherlands. An early vision of the project was described, in Dutch, as a 'Portal to a Research Paradise' by (van der Sijs, 2012).

Partners actively involved in the construction of Nederlab are Meertens Institute, Institute for Dutch Lexicology, Huygens ING and Radboud University. The project ties in with other major projects and initiatives: for collections Nederlab collaborates with academic libraries and institutions in the Netherlands and Flanders, for infrastructure with CLARIN (Odijk, 2010) and CLARIAH[2], for tools with eHumanities programmes such as NWO CATCH[3] and IM-

[1] Online at `http://www.nederlab.nl/`

[2] `http://www.clariah.nl/en/`

[3] `http://www.nwo.nl/en/research-and-results/programmes/Continuous+Access+To+Cultural+Heritage+%28CATCH%29`

PACT[4], and with language technologists from scientific institutions, again mainly in the Netherlands and Flanders.

The Nederlab project is currently about halfway. We have a solid collection pipeline in place for processing metadata, text content, and vocabularies, and have tested and applied this pipeline on three collections. The Nederlab index now contains 13.5 million searchable documents. We developed software for end users, the Research Portal, and have produced scripts and interactive tools for our back-office processes. Efficiently adding new collections to Nederlab is also a major point of attention. In this Nederlab acts as a 'user' to the CLARIAH project PICCL in which a corpus building work flow is further being developed (Reynaert et al., 2015).

The main focus for the rest of project period is on search and analysis facilities for linguistic annotations on a massive scale and on providing analytic tools in the context of scientific use cases.

3. Collections

Nederlab is designed to contain any historical text in the Dutch language from the Middle Ages up to now. Because of Intellectual Property Rights or IPR issues the main focus will be on texts from the Middle Ages until 1900, or in some cases 1940. Nederlab has to rely on texts that are already digitized, during the last twenty years, by institutions, companies and individuals. Nevertheless, this already results in massive amounts of data. Our original inventory of potential source collections contains 46 collections with an estimated 2.5 million titles containing tens of billions of words.

The accuracy of the texts but also of the bibliographical metadata differs quite a lot among collections. Some have very good bibliographical metadata, but quite poor and unstructured text files (often digitized by means of Optical Character Recognition). Other collections might contain very sophisticated XML-based text enrichment, in combination with minimal or lacking bibliographical metadata. To make all these data available in a homogeneous, or at least transparent, way, is quite a challenge.

Bibliographical and biographical metadata are key for the functioning of Nederlab, because those data provide the researcher with information about the precise dating and localization of texts, which is essential for a diachronic corpus as Nederlab is meant to be. Besides that, Nederlab offers information about text genres and about its authors, not only their names, but also a date and place of birth and death, gender, and other works written by the same person. Each bibliographical item is connected to a thesaurus of authors and in future to a thesaurus of titles, which will enable the user of Nederlab to find (or filter out) variant forms of the same texts, or to collect all texts by the same author, including the variant forms. In order to maintain the quality of its own thesauri Nederlab has to deal with the fluctuating and unstable quality of the metadata it receives from its partners and other data providers. Quality assessment, mapping, semi-automated harmonization of datastrings, but also manual data curation are essential to the editorial process that focuses on the possibilities, and also the comfort and speed that the end-user will experience in composing his own specific (virtual) datasets, in the preparatory phase of any corpus investigation.

Since March 2015 a beta version of the Nederlab Research Portal is online. It provides access to the first three of many collections. The Digital Library of Dutch Literature or DBNL collection[5] (van Stipriaan, 2009), contains high-quality transcribed texts and extensive, well-curated metadata. The second collection, Early Dutch Books Online[6], contains historical digital texts digitized by means of OCR. For this collection, Nederlab contains two alternative text versions per paragraph: the original OCR text and an automatically OCR post-corrected version for which (Reynaert, 2014) offers a description and an evaluation. The third collection was chosen partly to test scalability issues: the KB's newspaper collection[7] from 1618 up to 1900.

Central in the preparatory phase of Nederlab were the data of DBNL, which show a, rather rare, combination of elaborate and thesaurized metadata, next to high quality transcriptions of historical, mainly literary, texts. The DBNL contains TEI XML transcriptions of ca 12,000 titles, in total around 4 million pages of text, from the Middle Ages up to our days. The collection is considered as highly representative for the whole of Dutch literary fiction, although it shows some typical academic biases, for example more 'high' than 'low' culture. Besides literary texts it contains a vast amount of secondary literature on Dutch literature and language, as well as reference works and all kinds of journals (literary, scientific or popularizing).

4. The Nederlab collection pipeline

Within the CLARIN project considerable experience has been gained with harvesting and harmonizing metadata descriptions from various sources using the CMDI metadata framework (Broeder et al., 2010). Within the Nederlab project the CMDI approach in principle has been selected as the preferred method of metadata delivery as this provides optimal chances for automated mapping and ingest procedures. However, in practice the number of collection providers following the CMDI approach is still limited and although some data providers, such as the Dutch National Library (KB), have expressed interest in becoming CMDI data providers, metadata and data are currently delivered in various formats and using various delivery protocols. Metadata formats encountered from these data providers include DIDL[8], ALTO[9] and idiosyncratic TEI[10] formats.

In almost all cases Nederlab has to deal with customized import processes to ingest metadata and data into the portal. These are the steps in our per-collection workflow:

Acquisition and IPR arrangements Collection providers are requested to provide metadata and text

[4] http://www.digitisation.eu/

[5] http://www.dbnl.org/

[6] http://www.earlydutchbooksonline.nl/nl/edbo

[7] http://www.delpher.nl/nl/kranten/

[8] http://xml.coverpages.org/mpeg21-didl.html

[9] http://www.loc.gov/standards/alto/

[10] http://www.tei-c.org/

content of complete collections to be processed by Nederlab. There are potential problems with this that are discussed with them early in the acquisition process: Nederlab transforms collections to its own homogeneous representation suiting its own objectives, which may not always comply with those of the collection providers. This results in a newly annotated version of the collection which is republished via the Nederlab Research Environment. These collections often contain material that is subject to access restrictions because of Intellectual Property Rights. Therefore we aim for a simple, standard contract under the aegis of an authoritative institution, the KNAW or Royal Netherlands Academy of Arts and Sciences – to help ascertain future, free availability for research purposes, and explicit agreement on what access policies we implement to enforce this contract. For image content of the original materials, for instance, this often involves the Nederlab Portal linking back to and displaying the original version and attendant information at its home site.

Quality Assessment and collection description We systematically collect information about each collection. This information is used to support internal data processing and curation, and to inform end users about status and quality of the collection's data.

Metadata mapping We designed a fixed metadata schema for the four basic Nederlab resource types (Titles - in the sense of a work; Dependent Titles - only existing as part of another Title; Series - like newspapers or periodicals; and Persons - most importantly Authors) and represented and documented this schema using the CMDI framework. Metadata of incoming collections is either mapped to Nederlab metadata, or imported as 'collection specific' metadata, or ignored. This mapping is executed by the Nederlab editorial staff with a tailor-made metadata mapping tool.

Metadata conversion For each collection custom conversion and mapping scripts are written. The converted metadata is stored in a project-internal relational database. This database is then used for all curation tools and as source for the indexing process.

Text extraction and conversion Text content is extracted from the collection's textual resources, sometimes with a different granularity than the original (e.g. each newspaper article is extracted as a separate Nederlab title). Often the texts contain additional metadata or annotations of specific text segments (for example titles versus body text, or whether a segment is to be considered editorial matter). These metadata and annotations are also extracted. The information is then converted to the FoLiA XML format (van Gompel and Reynaert, 2013) and stored on a project internal FoLiA store. Subsequent text enrichments are added to the FoLiAs in the store. These FoLiA documents can also be retrieved for indexing purposes or to display them to end users.

Curation by editorial staff To facilitate the need for high quality data the ingest process is supervised and monitored by an editorial team. An integrated editorial environment supports this team in performing quality assessment checks, modifications and data alignment tasks. All metadata can be manually curated. The focus for this work is on author information, dates, genres and on the identification of identical resources (e.g. re-publications). Authors are 'thesaurized': in a semi-automated process authors from newly integrated collections are linked to already existing authors. The same process will be applied to titles. For this, a special tool is developed that guides the editor through a cascade of steps where different combinations of metadata values are systematically compared, for example, the combination of authors' surnames and years of birth. This 'thesaurizing' process is especially important for users that want to collect reliable statistics about (selected parts of) the Nederlab collection.

Text processing and enrichment Since part of the Nederlab corpus consists of rather low quality data that have been automatically digitized through Optical Character Recognition (OCR) techniques it was deemed necessary to raise the quality of these digitized texts. For this, a customized version of TICCL (Text-Induced Corpus Cleanup) (Reynaert, 2010) is used to reduce the amount of spelling variation introduced by the OCR process. Furthermore, the data is automatically enriched with lemmata and Part-of-Speech tags and Named Entities labels by means of Frog (Van den Bosch et al., 2007). Frog is developed for modern Dutch, and the results for historical variants of OCR-post-corrected Dutch vary from reasonable to mediocre; we are working on improving this. For this reason we started a pilot in which we *translate* historical text to a modern variant, using a translation lexicon learned from parallel seventeenth century text versus its modern equivalent. The translation step appeared to improve the accuracy of assigned Part-Of-Speech tags in seventeenth century texts by more than 10% (Tjong Kim Sang, 2015). In the future, we plan to apply parallel lexicons for the modernization of texts for other time periods as well. These lexicons will be derived from parallel texts, where available, and from the historical lexicons developed by the Institute for Dutch Lexicography in Leiden.

Indexing Incoming metadata and texts are periodically indexed. This index allows the user to efficiently search text and metadata, and to select a personal research corpus on the basis of the main corpus.

Currently, we are upgrading to the next generation of our indexing and search software that, in addition, is capable of searching for complex patterns of multi-layered linguistic annotations and generating statistics about those. We closely collaborate with the Institute for Dutch Lexicology (INL), who provide the corpus back-end BlackLab[11] and with the developers of WhiteLab[12], further being developed in the CLARIN-NL project OpenSoNaR-CGN, the sequel to (Reynaert et al., 2014).

5. User perspective and requirements

Academic end users of Nederlab will be mainly humanities scholars, in particular historians, literary scholars and linguists. The main focus of Nederlab is to enable research on patterns of change, over time and over space. More in

[11] `https://github.com/INL/BlackLab/wiki`

[12] `https://github.com/TiCCSoftware/WhiteLab`

particular, the following groups of use cases are, a.o., to be supported[13]:

- Detect the beginning of changes: When do new concepts occur, when are new word forms used for the first time and when are particular word combinations appearing?
- Study the spread of changes in language and culture: Researchers try to discover what phenomena and what types of phenomena have spread over time – from one text type to another, from one author to others, from one dialect to another, or to the standard language – and what phenomena have disappeared, and how.
- Find connections and networks: Which entities (persons, locations) occur in texts and what are their co-occurrence relations? How can texts be classified, for example by style characteristics or genres?
- Systematically compare (sets of) texts: Which text fragments within sets of documents are similar? What is the provenance of these fragments? In how far are the text characteristics of sets of documents similar, or different, for example comparing different dialects or authors?

To implement this wide range of use cases, at the scale that is foreseen, it is clear that Nederlab needs a powerful back-end that can retrieve a range of result types. Users have to be able to search for documents and filter those on the basis of associated title and author metadata. Next, they have to be able to find word forms, but also sequential patterns of word forms, and even sequential patterns over text and/or associated annotation layers (minimally, patterns such as can be specified by the Corpus Query Language - CQL (Christ, 1994). Finally, it has to be possible to retrieve frequency lists: these lists contain each occurring word type or annotation value in any subcollection of Nederlab, sorted by frequency of occurrence.

For these result types there are different ways to present to the researchers, depending on what they are interested in. The most obvious way is as sorted lists: lists of documents (with associated metadata), lists of text snippets showing highlighted hits with some context, or keyword-in-context (KWIC) concordances that show more complex hit patterns with some left and right context and all of the associated annotation values.

A second way is to organize results in groups and show each group, together with the number of results in this group. Results can be grouped on the basis of their values themselves, on the basis of a part of the result (e.g. adjectives preceding the lemma ”heart”), or on the basis of one or more metadata dimensions associated with the result (e.g. per genre and then distributed over time periods). Especially the last case lends itself very well for visualizations.

An even more aggregated way to present results is as a statistical overview, both about the results themselves and about the documents containing these results (e.g. total numbers, average numbers per document, spread, etc).

On top of this back-end Nederlab will have a number of analytical tools, for example tools to compare subcollections on the basis of their frequency lists.

Many researchers prefer to use Nederlab only for the first stages of their research. For those researchers export of results, in a number of ways and using a number of formats, will be available. An example is export of multi-dimensional arrays to be used in packages such as R.

The next section describes what we implemented so far in our research environment and our plans for the near future.

6. Virtual Research Environment

All users have access to the Nederlab search interface. They can select the way in which their search results are represented: as a pageable list of result snippets, as keyword-in-context concordance or, visually, as a time distribution graphic showing the numbers of matching documents over time. An article, in Dutch, describing the envisioned user facilities for Nederlab is (van der Sijs, 2015).

Users can gradually refine the set of matching search results by adding constraints on either metadata or text content, in any order. At this moment, Title metadata allows users to filter on title string, type of title, genre, source collection, availability of text, availability of corrected text and year of publication. It is possible to exclude more recent copies of Titles that are present in Nederlab in an older version. Person metadata can be used to filter Titles on characteristics of persons (authors, editors, illustrators, etc.) that are associated with these Titles. Person metadata includes person names (including varieties of these names collected over all source collections), life years and places of birth and death.

Nederlab uses its own genre vocabulary. Genres of source collections are mapped to Nederlab genres, enabling Nederlab wide selection on the basis of genre, across different source collections.

For text search the Nederlab Portal supports (boolean combinations of) search strings, expansion of search strings with historical lexical variations and the possibility to select the specific text layer or version in which to search (original text, corrected text and/or case-sensitive text).

To enjoy the full functionality, users have to log in with a user account of one of organizations participating in the CLARIN federation. We consider such a log-in to be proof of use for research purposes, meaning the user is to be considered an authorized user. Authorized users have three additional benefits over non-authorized users: they are allowed to inspect more text content, they are able to store their queries as virtual research collections in their personal workspaces, and they have access to a growing number of analytical tools to work on these virtual research collections. Currently, a limited initial set of analytical tools is available. These tools are mainly focused on exploring and visualizing metadata such as distributions over genre, locations, or gender and age information of authors. It also is possible to visualize combinations of these dimensions. These visualizations do not only provide visual means for representing metadata, but also new ways of filtering and searching as the visualizations can be made navigable.

[13] For more examples see the original subsidy application at `http://www.nederlab.nl/cms/wp-content/uploads/2015/10/Nederlab_NWO_Groot_English_aanvraagformulier.pdf`

In future versions we will expand the set of available analytical tools. This requires that we first replace the current search engine with a new one, that enables the functionality described in Section 4., and that gives results in term frequencies, next to document frequencies. We further plan to support uploading of the user's own collections to their personal workspace and enable searching and analysis of these collections in the context of the rest of Nederlab.

7. Additional web services

Lexicon service The Dutch Institute for Lexicology or INL is a partner in the project. INL contributions to Nederlab include a historical Dutch lexicon. This lexicon is accessible using a RESTful web service[14] and is actively used in the Portal to expand user queries with known diachronic word variants on demand.

User annotations and Alexandria Huygens ING, yet another project partner, builds Alexandria, a repository for text and annotations for Nederlab. It will be used to store and retrieve user generated annotations for all kinds of objects and object segments in Nederlab.

R based visualization service We chose to base Nederlab visualizations on an separate web service that is based on the R open source software environment for statistical computing and graphics[15]. This allows us, and potentially end users as well, to plug in custom R modules in the future.

8. Conclusion

Half way through the Nederlab project we have versions of most required components in place. We cover the whole trajectory from selecting and evaluating source collections all the way to generating statistical analyses over these collections in the context of all the other collections. Some of these components are still rudimentary, most of them need further development.

We have tested all of this by processing three very different collections. We make these collections available in our Research Portal in a homogeneous and useful way.

We have gained insight in the processes and technology we need and in scalability issues. At this stage, Nederlab is constructed as an extensible framework. It can be extended by adding a variety of scholarly tools, as well as more collections, both during and after the remaining project period.

9. Acknowledgments

The Nederlab project described in this paper has been made possible through project grants from the Netherlands Organisation for Scientific Research (NWO), KNAW, CLARIN-NL, and CLARIAH.

Broeder, D., Kemps-Snijders, M., Uytvanck, D. V., Windhouwer, M., Withers, P., Wittenburg, P., and Zinn, C. (2010). A Data Category Registry- and Component-based Metadata Framework. In Nicoletta Calzolari et al., editor, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC-2010)*, Valletta, Malta. ELRA.

Christ, O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. In *Proceedings of COMPLEX '94: 3rd Conference on Computational Lexicography and Text Research, Budapest)*.

Odijk, J. (2010). The CLARIN-NL project. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC-2010*, pages 48–53, Valletta, Malta.

Reynaert, M., van de Camp, M., and van Zaanen, M. (2014). OpenSoNaR: user-driven development of the SoNaR corpus interfaces. In *Proceedings of COLING 2014: System Demonstrations*, pages 124–128, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Reynaert, M., van Gompel, M., van der Sloot, K., and van den Bosch, A. (2015). PICCL: Philosophical Integrator of Computational and Corpus Libraries. In *Proceedings of CLARIN Annual Conference 2015 – Abstracts*, pages 75–79, Wrocław, Poland. CLARIN ERIC.

Reynaert, M. (2010). Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition*, 14:173–187.

Reynaert, M. (2014). Synergy of Nederlab and @PhilosTEI: diachronic and multilingual Text-Induced Corpus Clean-up. In *Proceedings of the Ninth International Language Resources and Evaluation Conference (LREC-2014)*, Reykjavik, Iceland. ELRA.

Stehouwer, H., Durco, M., Auer, E., and Broeder, D. (2012). Federated search: Towards a common search infrastructure. In Nicoletta Calzolari et al., editor, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey. ELRA.

Tjong Kim Sang, E. (2015). *Converting seventeenth century Dutch to modern Dutch.* Presented at the Workshop Morphosyntactic Enrichment of Historical Text, Utrecht, The Netherlands.

Van den Bosch, A., Busser, G., Canisius, S., and Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In P. Dirix et al., editor, *Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting*, pages 99–114, Leuven, Belgium.

van der Sijs, N. (2012). Toegangspoort tot digitaal onderzoeksparadijs. *Informatie Professional*, 10:20–23.

van der Sijs, N. (2015). TS Tools: Nederlab voor tijdschriftonderzoek. *TS. Tijdschrift voor tijdschriftstudies*, 37:53–62. OA tijdschrift.

van Gompel, M. and Reynaert, M. (2013). FoLiA: A practical XML Format for Linguistic Annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3.

van Stipriaan, R. (2009). At the Speed of Writing, and Beyond. A Short History of the Digital Library of Dutch Literature (DBNL). *Logos*, 20(1):74–78.

[14] `http://sk.taalbanknederlands.inl.nl/LexiconService/`

[15] `https://www.r-project.org`