CLARIN-NL: Major Results

Jan Odijk

UiL-OTS Utrecht University
Trans 10, 3512 JK Utrecht the Netherlands
j.odijk@uu.nl

Abstract

In this paper I will provide a high level overview of the major results of CLARIN-NL so far. I will show that CLARIN-NL is starting to provide the data, facilities and services in the CLARIN infrastructure to carry out humanities research supported by large amounts of data and tools. These services have easy interfaces and easy search options (no technical background needed). Still some training is required, to understand both the possibilities and the limitations of the data and the tools. Actual use of the facilities leads to suggestions for improvements and to suggestions for new functionality. All researchers are therefore invited to start using the elements in the CLARIN infrastructure offered by CLARIN-NL.Though I will show that a lot has been achieved in the CLARIN-NL project, I will also provide a long list of functionality and interoperability cases that have not been dealt with in CLARIN-NL and must remain for future work.

Keywords: CLARIN-NL, infrastructure, results

1. Introduction

In this paper I will provide a high level overview of the major results of the CLARIN-NL project so far, but I will also discuss a wish list of functionality that has not been provided by CLARIN-NL. Though the CLARIN-NL project is still running, and several services and data are still being produced, the major subprojects of CLARIN-NL have been defined and most of them have already finished.

The CLARIN-NL project provides the Dutch national contribution to the Europe-wide CLARIN infrastructure. CLARIN-NL runs from 2009 through 2014, has a budget of 9.01 m euro, is financed by the Dutch Science Foundation (NWO) as part of the National Roadmap for Large Scale Infrastructures, and is coordinated by Utrecht University. Over 33 partners participate, including universities, royal academy institutes, independent institutes, libraries, and data centres.

CLARIN is a research infrastructure intended for humanities and social science (SSH) researchers that work with language resources. CLARIN-NL has mainly focused on humanities researchers. The CLARIN infrastructure was prepared by the CLARIN preparatory project (CLARIN-PP, 2008-2011), coordinated by Utrecht University. As of February 2012, the CLARIN infrastructure is coordinated by CLARIN ERIC, hosted by the Netherlands. An ERIC is a legal entity at the European level specifically set up for research infrastructures.

The CLARIN infrastructure is virtual (i.e. runs as software on the internet) and distributed. The infrastructure is implemented by so-called *CLARIN Centres*. The CLARIN centres in the Netherlands are INL, Meertens, MPI, Huygens ING, and DANS. In addition, so-called *CLARIN Data Providers* include the National Library, the *Netherlands Institute for Sound and Vision* (NISV), and the library of Utrecht University. These centres make available data and services, as well as metadata for these data and services. They ensure visibility, accessibility, and long term preservation of these data and services.

In a nutshell, the CLARIN infrastructure aims to provide a research infrastructure in which an SSH researcher who works with language resources

- Can find all data relevant for the research
- Can find all tools and services relevant for the research
- Can apply the tools and services to the data without any technical background or ad-hoc adaptations
- Can store data and tools resulting from the research in CLARIN

and it should be possible to do this from a single portal.

I will dedicate a separate section to each of these aspects: search for data (section 2.), search for tools and services (section 3.), apply tools and services to data (section 4.), incorporate data and services in CLARIN (section 5.), and a portal to get access to CLARIN (section 6.). I invite researchers to start using the CLARIN infrastructure (section 7.), sketch some of the desired future extensions of the functionality in CLARIN (section 8.), and will end with the conclusions of this paper (section 9.).

2. Search for Data

Searching for data is done via searching in metadata. CLARIN requires CMDI metadata (Broeder et al., 2010) for all data in the infrastructure. Both data and metadata are stored at servers of CLARIN centres and made accessible to CLARIN users. The CLARIN preparatory project made the Virtual Language Observatory (VLO) with faceted browsing and geographical navigation functionality (Van Uytvanck et al., 2012). The VLO offers free text search and faceted browsing for a limited and fixed number of facets, which have been derived from metadata elements. A hyperlink to the metadata in the VLO for data and services mentioned below will be provided in this paper where available. In CLARIN-NL, an alternative Metadata Search engine was created. It offers free text search but also search for values of arbitrary metadata elements in the metadata, and incremental refinement of the search results (Kemps-Snijders et al., 2012). Though the Metadata Search engine was created in CLARIN-NL, it will of course search in all metadata in the CLARIN infrastructure, whatever their origin.

2.1. Data Curated

CLARIN-NL has hardly created any new data. It has focused on adapting ('curating') existing data in such a way that they become CLARIN-compatible, i.e represented in CLARIN-supported formats, with CMDI-metadata, with explicit semantics via CLARIN supported data category registries, e.g. ISOCAT, (Kemps-Snijders et al., 2010), and stored on a CLARIN-centre, which ensures visibility (via the metadata), accessibility via persistent identifiers, and long term preservation.

A whole range of data has been curated in CLARIN-NL. These data have been curated by interdisciplinary teams of humanities researchers, ICT researchers, and infrastructure specialists from research organisations and CLARIN Centres. They include lexical data, literary data, linguistically annotated corpora and databases, historical and contemporary text corpora, data for religion studies, and art history texts. This wide range shows that the CLARIN infrastructure truely is for all humanities researchers working with digital language data, including lexicologists, literary scholars, linguists, historians, political scientists, religion scholars, and art historians.

Lexical data

- Dutch Dialect Dictionaries for Brabant and Limburg (curated by the COAVA project and the Data Curation Service, see below);
- Cornetto data ((Vossen et al., 2013), a combination of Dutch Wordnet (Vossen, 1998) and the Dutch lexicon *Referentiebestand Nederlands* (Martin and Maks, 2005) in LMF and RDF format.
- DuELME database of Multiword Expressions in LMF format

Literary data

- Metadata for Arthurian Fiction are accessible via the VLO
- Metadata for the *Liederenbank* (Dutch Song Database) are accessible via the VLO
- Emblem metadata are available in a first version and are currently being finalized by the EMIT-X project.
- COBWWWEB WomenWriters database connected to other national collections in women's literature (expected in 2014)
- eBNM+: curated e-BNM collection of textual, codicological and historical information about thousands of Middle Dutch manuscripts kept world wide (expected in 2014)

Linguistically annotated corpora

- DISCAN text corpus enriched with discourse annotation
- IPROSLA project Sign Language data (license needed for access to the data)
- The DiDDD, Dynasand, and GTRP microcomparative databases for Dutch dialects (MIMORE project)
- Negerhollands data by the NEHOL project
- Historical news corpus by the VU-DNC project
- Curated Database of the Longitudinal Utrecht Collection of English Accents (D-LUCEA, expected in 2014)
- EXILSEA project enhancements of the Corpus NGT, the worlds first open access sign language corpus, by updating the existing IMDI metadata to CLARIN-standard CMDI descriptions using bilingual ISOcat categories (expected in 2014)
- FESLI curated specific language impairment data (expected in 2014)
- LAISEANG language documentation data for a wide range of languages from Insular South East Asia and West New Guinea (expected in 2014)
- Five existing, digital data sets of language pathology data collected in the Netherlands, primarily on Dutch, are being curated by the VALID project (expected in 2014)
- WIVU Hebrew Text Database is being curated by the SHEBANQ project (expected in 2014)

Historical and contemporary text corpora

- Loe de Jong's texts on the Second World War curated (by the *Verrijkt Koninkrijk* project) (obtainable via DANS)
- Curated maritime history legacy datasets curated with a tool chain and methodology developed by the DSS project (expected in 2014)

Data for religion studies

- Pilgrimage data curated in the PILNAR project
- The WIVU Hebrew Text Database is being curated by the SHEBANQ project (expected in 2014)

Art History data

- Rembrandt Documents (RemDoc) database linked with related resources created by the *Rijksbureau voor Kunsthistorische Documentatie* (RKD)[1], and with a university library catalogue (by the RemBench project, expected in 2014)

[1] State Office for Art History Documentation.

In addition, the Netherlands Institute for Sound and Vision (NISV) has made its Academia Collection metadata available in CMDI format, and the actual data are accessible with a special license. Similarly, Utrecht University Library has made its its digital data available via CMDI metadata, and the National Library is working on this (expected early 2014).

CLARIN-NL has also set up a Data Curation Service (Oostdijk and van den Heuvel, 2012), which has curated several resources, including IPNV Interviews with veterans, the dictionary of 'Gelderse' Dialects (Rivierengebied and Veluwe), LESLLA (Lower Education Second Language Learner Acquisition data), the Dutch Bilingualism Database / TCULT, and 6 Brabant dialect dictionaries. It has also harmonised organisation names for CLAVAS (see section 5.), and is working on Roots of Ethnolects data and Traces of Contact data.

3. Search for Tools and Services

One will find very few tools and services that have been contributed by CLARIN-NL in the VLO or the Metadata Search Engine. Many tools and services have been made available by CLARIN-NL, as we will describe in section 4.. However, metadata for tools and services have not been created systematically so far. A separate project has been set up to develop a CMDI profile and components for describing tools and services. This profile has been tested against 5 services in the CLARIN infrastructure (Westerhout and Odijk, 2013), and is currently being refined and applied to all tools and services offered by CLARIN-NL . It is expected that metadata for these tools and services will be visible via the VLO by the end of 2014.

4. Apply Services to Data

I have categorized tools and services globally in three categories:

- services for searching in and through data
- services for analyzing data and visualising the analysis results
- services for enriching data

I will discuss each category in a separate subsection.

4.1. Search

Many services for searching in and through specific data sets have been created and made available in the CLARIN infrastructure. They include services for searching in lexical data, in linguistically annotated text corpora, in literary data, in historical and contemporary text corpora, and in data for religion studies.

Lexical Data These include the COAVA application Dialect Lexicon Browser, the search interface to the Cornetto lexico-semantic database, the DuELME multiword expression database search interface, the GTB (Integrated Language Bank) including the WFT-GTB Frisian dictionary in the GTB, as well as a search interface for searching in a Greek-Dutch dictionary (letter π only).

Linguistically annotated data These include the COAVA application CHILDES Browser, a search interface to the Corpus Gysseling provided by INL, the FESLI Search application for search in language selective impairment acquisition data, a simple interface to search for grammatical relations between words in Dutch sentences from LASSY Small (65k sentences) and the wiki part (8.5 million sentences) of LASSY Large (van Noord et al., 2013). It also includes GrETel, which is a result of CLARIN Flanders in the context of the CLARIN-NL/CLARIN Flanders Cooperation, and which enables searching for grammatical constructions in the LASSY-Small and Spoken Dutch Corpus (CGN, (Oostdijk, 2000)) treebanks by providing an example of the construction and marking which aspects of the example are crucial in defining the construction. The Mimore search engine enables search in the micro-comparative databases for Dutch dialects *DiDDD*, *DynaSand* and *GTRP* described in section 2.. The OpenSoNaR tool allows researchers to explore the SoNaR-500 (Oostdijk et al., 2013) reference corpus (preliminary version available, full version expected in 2014). The TDS-Curator project created a search interface to the Typological Database System (TDS). Finally, work is ongoing for the SHEBANQ web application that enables researchers to perform linguistic queries on the curated WIVU web resource and preserve significant results as annotations to this resource (expected in 2014).

Literary data Metadata for Arthurian Fiction can be searched through the Arthurian Fiction web application. Metadata for the *Liederenbank* (Dutch Song Database) can be searched via its web application. Namescape offers a search interface for searching named entities in literary works. The COBWWWEB scholar application for research on the WomenWriters Database and its connected databases, as well as the eBNM+ web application for consultation, using facetted search, and collaborative editing are expected in 2014.

Historical and contemporary text corpora BILAND offers a multilingual application for search and discourse analysis in historical text corpora. The CKCC (*Geleerdenbrieven*) project, which was partially funded by CLARIN-NL, yielded ePistolarium, which enables researchers to browse and analyze around 20,000 letters that were written by and sent to 17th century scholars who lived in the Dutch Republic. A web application for search in Loe de Jong's work on the Second World War was developed by the Verrijkt Koninkrijk project. The WAHSP project created a search engine and interface for sentiment mining in the digital newspaper collection of the Dutch Royal Library. The WIP project produced a search engine for searching in the proceedings of the Dutch Hansard (*Handelingen der Staten-Generaal 1930-1995*). Quamerdes is working on an application for quantitative content analysis of television and printed media, and a CLARIN-NL financed subpart of the Nederlab project will provide data and tools for the longitudinal study of Dutch language and culture (both expected in 2014).

Data for Religion Studies PILNAR provided a web application for search in Pilgrimage data. The SHEBANQ web application that is under development has been described above.

4.2. Analysis and Visualisaton

Several tools enable one to analyse data and to visualise the results of the analysis, often in multiple ways. In some services this is combined with search functionality.

Gabmap is a web application that analyses and visualises dialect variations. MIGMAP enables one to analyze and visualise migration patterns by creating maps that show the dispersion of people in the Netherlands during the 20th century at the level of municipalities. MIMORE enables one to analyse and visualise micro-comparative data from three databases for Dutch dialects. WIP, WAHSP, BILAND, and Quamerdes offer rich opportunities for analysis and visualisation of the data. The CKCC ePistolarium enables visualizations of geographical, time-based, social network and co-citation inquiries. The Polimedia project yielded an application for cross-media analysis. Namescape offers next to its search interface a barcode browser, and a visualiser for analysing and visualising named entities in literary works.

4.3. Enrichment

A wide range of tools for enriching and annotating data have been created or adapted to be CLARIN-compatible.

Several enhancements have been made to ELAN and ANNEX, tools for the annotation and display of time-based resources such as audio and video.

The AAM-LR CLAM web service can be called from ELAN and distinguishes segments in an audio-visual signal as speech or non-speech, and will provide a rough phonetic transliteration of the speech. Enhancements have also been made for multimodal collocations. Enhancements for annotating sign language data were made to the LEXUS and ELAN tools. The ELAN and ANNEX applications are also being enhanced with a referencing and note exchanging system (expected in 2014), and with enhancements making use of the multilingual features of ISOCAT (expected in 2014)

The Oral History Annotation Tool has been used for the annotation of a collection of 250 interviews from the *Interview Project Nederlandse Veteranen* (Dutch Veterans). Huygens ING, one of the Dutch CLARIN Centres, is extending eLaborate and making it CLARIN-compatible (expected in 2014). The Adelheid web service, tokenizer, lexicon and editor enable tokenization, lemmatisation and part-of-speech tagging for historical Dutch. INPOLDER is an application for parsing Historical Dutch and also includes a work flow in which it is combined with the Adelheid Tagger. TICLLops is an application for orthographic normalisation, which has also been incorporated in the TTNWW work flow system and is further elaborated in the @PhilosTEI project, which will provide an open source, web-based, user-friendly workflow from textual digital images to TEI (expected in 2014). TTNWW is a work flow system created in a cooperation project between CLARIN-NL and CLARIN Flanders, which includes web services for spelling normalisation (TICCLops), Part of Speech-tagging (Frog), Parsing (Alpino parser), Named Entity Recognition, Semantic Role Assignment, Assignment of co-referential relations and transcription of speech files. In TTNWW a generic solution for turning existing software into web services was created with the Computational Linguistic Application Mediator (CLAM, and a new candidate for a standard format for annotated text corpora (FoLiA) emerged, which currently is widely in use in the Netherlands (van Gompel et al., 2011; van Gompel and Reynaert, 2013). NameScape also provided a Named Entity Tagger. Transcription Quality Evaluation (TQE) is tool for assessing the quality of phonetic transcriptions for spoken resources. eBNM+ is working on a web application for consultation, using facetted search, and collaborative editing, and DSS creates a tool chain and methodology for converting legacy datasets in the area of maritime history. The latter two are expected to be available in 2014.

5. Incorporating Data and Services in CLARIN

Research projects may yield new data and new services. These must be incorporated in CLARIN, and therefore be CLARIN-compatible. Various facilities to support this are offered in the CLARIN infrastructure.

5.1. Metadata

All data and services in the CLARIN infrastructure must have CMDI metadata. CLARIN offers metadata profiles and components that can be re-used as well as tools to create new metadata profiles and components via the the metadata registry. Initial versions have been created in CLARIN-PP but they have been further elaborated in CLARIN-NL. There are also editors for creating and modifying metadata (e.g. ARBIL), and metadata profiles and components have been created for software, as described in section 3..

5.2. Formal and semantic interoperability

CLARIN supports formal interoperablity by requiring that data and tools conform to CLARIN-supported standards and best practices (Kemps-Snijders et al., 2009). For semantic interoperability the data category registry ISOCAT (Kemps-Snijders et al., 2010) plays a crucial role. It offers a web interface, web services, and documentation such as manuals, help, and tutorials. It has turned out that ISOCAT alone was not enough to ensure semantic interoperability. For this reason new registries such as RELCAT and SCHEMACAT have been created, and via CLAVAS a single entry point to multiple concept and data category registries has been made available.

5.3. Ingesting data

Several centres use special software to ingest new data and software. This automates the ingestion process largely and ensures the consistent application of this process. The Language Archive (TLA) uses the LAMUS software for this purpose, and DANS offers EASY. Adaptations to this software have been made and are being made to ensure that they effortlessly operate with CLARIN-compatible data and make these data and their metadata available and accessible in a CLARIN-compatible manner.

6. Portal

A portal for the Netherlands part of the CLARIN infrastructure is under construction and the exact relation of this portal (and other national portals) to the CLARIN ERIC portal is being defined. It is expected that a first version of the portal will be up by the middle of 2014. As a temporary stand-in, this page provides a brief overview of what CLARIN-NL has produced, ordered more or less as in this paper.

7. Invitation

CLARIN-NL is starting to provide the data, facilities and services in the CLARIN infrastructure to carry out humanities research supported by large amounts of data and tools. These services have easy interfaces and easy search options (no technical background needed). Still some training is required, to understand both the possibilities and the limitations of the data and the tools. To this end, educational modules are being developed for selected functionality by CLARIN-NL (expected in the second half of 2014).
Actual use of the facilities leads to suggestions for improvements and to suggestions for new functionality. I therefore invite researchers to start using (elements from) the CLARIN infrastructure offered by CLARIN-NL. If one encounters problems or has questions, one can turn to the CLARIN-NL Helpdesk. Provide feedback so that we can further improve CLARIN and so that you can improve your research!

8. Future Work

Though a lot has been achieved, as is clear from the preceding sections, there is still a lot to do. We mention a few items:

- Not all data (even some crucial data) are visible via the VLO or via Metadata Search
- Very few tools and web services are currently visible via the VLO
- Many tools are still prototypes or first versions
- There are good search facilities for some individual resources but not for all
- The search facilities so far are aimed at a single resource, or a small group of closely related resources.
- Federated content search[2], which enables one to search with one query in multiple, quite diverse, resources, is still being worked on but has turned out to be quite difficult.

Other functionality on the wish list includes

- Increased interoperability by providing converters and/or wrappers to enable the use of all CLARIN-supported standard formats. There are often multiple supported standard formats for the same data type, but most tools allow only one of these standard formats as input.
- Enrichment of data must be followed by analysis or search. For example, TTNWW enables automatic enrichment of text corpora. But that is just a first step. No researcher is interested in that in itself. It should be possible to offer the enriched results to a search engine or to sophisticated analysis tools. This is currently not possible in such a way that the targeted users of CLARIN, humanities researchers, can do this by themselves (unless they have specific technical knowledge).
- Flexible work flows for search / analysis services / visualisation services. Search queries applied to large data often yields large results. These cannot be analyzed by hand. Therefore, each search tool should yield output formats suitable for existing analysis and visualisation software (e.g. appropriate formats for input to Excel, Calc, R, SPSS, etc.). And it should be made possible and easy to reapply a search application to its own output for incremental refinement of the search results.
- For federated search it may be wise to start with simpler cases of federated search, and then improve incrementally. At this moment, federated content search is restricted to string (keyword) search. Full-fledged federated content search is not possible yet. But much simpler cases of search through multiple resources are not possible either.[3] For example
  - Search with one query in multiple Dutch lexical resources: CGN-lexicon, CELEX, GTB, Cornetto, DuELME-LMF,
  - Search with one query in multiple Dutch PoS-tagged text corpora: CGN, D-COI, SONAR-500, VU-DNC, Childes corpora,
  - Search with one query in multiple Dutch treebanks: CGN treebank, LASSY-Small, LASSY-Large

  These might be steps in an incremental approach to get to full-fledged federated content search.
- Chaining Search. An engine such as GrETEL enables a researcher to search for grammatical constructions on the basis of morpho-syntactic and syntactic information. But often selecting the relevant set of data for a grammatical construction also requires phonological, morphological or semantic conditions. For example, if one is interested in studying bare noun phrases (i.e. noun phrases without a determiner), then usually one is specifically interested in noun phrases with a singular *count* noun as head. But the distinction between *count* and *mass* is semantic in nature, not included in the treebanks that GrETEL operates on, and hence one cannot select for such noun phrases with GrETEL alone. We need a way to select for semantic distinctions such as *count/mass* as well. A second example is *binominal* noun phrases (i.e a specific type

[2] The main federated search system has been created by CLARIN-D.

[3] MPIs TROVA offers some of the functionality described here, though not as federated search.

of noun phrase with two nouns in it (Broekhuis and Keizer, 2013)). Selecting for such noun phrases purely on the basis of morpho-syntactic and syntactic information is not possible: semantic information (e.g. that the first noun must denote a quantity or a group) is required as well. If we could combine GrETEL with the Cornetto lexico-semantic database, we could make such selections, but at this moment this is not possible yet.

Another example concerns morphology or morphological potential. In certain constructions in Dutch the adjective exceptionally does not have an inflectional suffix, e.g. in *het bijvoeglijk(*e) naamwoord*[4] (Odijk, 1992). Selecting these using GRETEL is easy, but the results will always include a lot of examples with adjectives in which the distinction between the inflected and the non-inflected form is neutralized (e.g. *eigen* 'own'). Such examples cannot be used for an analysis of the absence of inflectional suffixes, so they should be filtered out. The morphological information is not encoded in the treebanks that GrETEL operates on, but it is available in or can be computed from lexicons with morphological information such as CELEX or the CGN-lexicon.

Finally, adjectives in Dutch only optionally get an inflectional ending (a schwa) when it ends in 2 syllables containing a schwa. We thus would like to filter out such examples by making use of a lexicon that contains such phonological information (e.g. CELEX) and regular expressions over such phonological representations. But again, currently this is not possible.

- Parameterized queries (batch queries). It is very easy to get a large number of results from certain queries. And often a lot of operations have to be applied to each of these results. Doing this by hand is too tedious, so automating this is required. For example, in a study on the modification potential of the Dutch words *zeer*, *heel* and *erg* (which are synonyms and mean 'very'), I wanted to do treebank searches with all synonyms/hyponyms of these words as provided by Cornetto. But there are more than 60 of them! Technically, it is pretty easy to automate this (if you can program and have access to the right APIs), but currently this has not been done and therefore this functionality is lacking for the users targeted by CLARIN.
- Requirements on tools for replicability One of my student tried to replicate similarity measure calculations on Wordnet of (Patwardhan and Pedersen, 2006) and (Pedersen, 2010). He did this in an excellent team: Piek Vossen and his research group. He did it with the help of one the original authors: Ted Pedersen. They used the exact same software and data as in the original paper. Nevertheless, they failed to reproduce the original results! And the reason for this is that 'properties which are not addressed in the literature may influence the output of similarity measures' (Fokkens et al., 2013). Many experiments and Pedersens unpublished intermediate results were required to determine the original settings of all parameters (e.g. treatment of ties in Spearman ρ), and which aspects of the data had been used and how. As one step towards a solution for this, the following is needed:

  - All tools must allow input of metadata associated with data
  - All tools must provide provenance data
  - All tools must provide a list with settings of all parameters as part of the provenance data, and this list must also be usable as a configuration file for the tool
  - All tools must generate new metadata for its results based on the input metadata, the generated provenance data, and possibly some manual input of a user

  However, at this moment, very few if any tools meet these requirements.

9. Conclusions

CLARIN-NL is starting to provide the data, facilities and services in the CLARIN infrastructure to carry out humanities research supported by large amounts of data and tools. These services have easy interfaces and easy search options (no technical background needed). Still some training is required, to understand both the possibilities and the limitations of the data and the tools.
Actual use of the facilities leads to suggestions for improvements and to suggestions for new functionality. All researchers are therefore invited to start using the elements in the CLARIN infrastructure offered by CLARIN-NL.
Though a lot has been achieved in the CLARIN-NL project, I have provided a long list of functionality and interoperability cases that have not (yet) been dealt with in CLARIN-NL. These should be taken up in future work.

Acknowledgments

This work was funded by the NWO CLARIN-NL project.

10. References

Broeder, D., Kemps-Snijders, M., Uytvanck, D. Van, Windhouwer, M., Withers, P., Wittenburg, P., and Zinn, C. (2010). A data category registry- and component-based metadata framework. In Calzolari, N., Maegaard, B., Mariani, J., Odijk, J., Choukri, K., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 43–47, Valetta, Malta. European Language Resources Association (ELRA). URL.

Broekhuis, Hans and Keizer, Evelien. (2013). *Syntax of Dutch: Nouns and Noun Phrases, Volume 1*. Comprehensive Grammar Resources. Amsterdam University Press, Amsterdam, the Netherlands. PDF.

[4] lit. 'the adjectival noun', meaning 'the adjective'

Fokkens, A., van Erp, M., Postma, M., Pedersen, T., Vossen, P., and Freire, N. (2013). Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1691–1701, Sofia. ACL. PDF.

Kemps-Snijders, Marc, Bel, Núria, Wittenburg, Peter, Broeder, Daan, Van Uytvanck, Dieter, Romary, Laurent, Hinrichs, Erhard, and Budin, Gerhard. (2009). Standards for LRT. CLARIN report, Nijmegen, January. URL.

Kemps-Snijders, M., Windhouwer, M.A., and Wright, S.E. (2010). Principles of ISOcat, a data category registry. Presentation at the RELISH workshop Rendering endangered languages lexicons interoperable through standards harmonization Workshop on Lexicon Tools and Lexicon Standards, Nijmegen, The Netherlands, August 4-5, 2010. PPTX.

Kemps-Snijders, M., de Bruin, M.J., Kunst, J.P., van der Peet, C.M., Zeeman, R.H.M., and Zhang, J. (2012). Applying CMDI in real life: the Meertens case. In *Proceedings of the Workshop 'Describing Language Resources with Metadata'*, Istanbul, May 22. LREC 2012. URL.

Martin, Willy and Maks, Isa. (2005). Referentiebestand Nederlands: Documentatie. Report, Free University Amsterdam, Amsterdam, April. PDF.

Odijk, Jan. (1992). Uninflected adjectives in Dutch. In Bok-Bennema, Reineke and van Hout, Roeland, editors, *Linguistics in the Netherlands 1992*, number 9 in AVT Publications, pages 197–208. John Benjamins, Amsterdam, the Netherlands.

Oostdijk, Nelleke and van den Heuvel, Henk. (2012). Introducing the CLARIN-NL data curation service. In *Proceedings of the LREC 2012 Workshop Challenges in the Management of Large Corpora*, pages 29–34. ELRA. URL.

Oostdijk, Nelleke, Reynaert, Martin, Hoste, Véronique, and Schuurman, Ineke. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In Spyns, Peter and Odijk, Jan, editors, *Essential Speech and Language Technology for Dutch. Results by the STEVIN-programme*, volume XVII of *Theory and Applications of Natural Language Processing*, chapter 13. Springer, Berlin, Germany. URL.

Oostdijk, N.H.J. (2000). The design of the Spoken Dutch Corpus. In Peters, P., Collins, P., and Smith, A., editors, *New frontiers of corpus research. Papers from the twenty first international conference on English language research on computerized corpora*, volume 36 of *Language and Computers*, pages 105–112, Amsterdam, the Netherlands. Rodopi.

Patwardhan, Siddharth and Pedersen, Ted. (2006). Using Wordnet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, Trento, Italy.

Pedersen, Ted. (2010). Information content measures of semantic similarity perform better without sense-tagged text. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2010)*. [PDF.

van Gompel, Maarten and Reynaert, Martin. (2013). FoLiA: A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81, 12/2013. PDF.

van Gompel, M., Reynaert, M., and van den Bosch, A. (2011). CLAM: Computational linguistics application mediator. Presented at CLIN21, Gent.

van Noord, Gertjan, Bouma, Gosse, van Eynde, Frank, de Kok, Daniel, van der Linde, Jelmer, Schuurman, Ineke, Tjong Kim Sang, Erik, and Vandeghinste, Vincent. (2013). Large scale syntactic annotation of written Dutch: Lassy. In Spyns, Peter and Odijk, Jan, editors, *Essential Speech and Language Technology for Dutch. Results by the STEVIN-programme*, volume XVII of *Theory and Applications of Natural Language Processing*, chapter 9. Springer, Berlin, Germany. URL.

Van Uytvanck, Dieter, Stehouwer, Herman, and Lampen, Lari. (2012). Semantic metadata mapping in practice: the Virtual Language Observatory. In Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, Doğan, Mehmet Uğur, Maegaard, Bente, Mariani, Joseph, Odijk, Jan, and Piperidis, Stelios, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA). URL.

Vossen, P., Maks, I., Segers, R., van der Vliet, H., Moens, M-F., Hofmann, K., Tjong Kim Sang, E., and de Rijke, M. (2013). Cornetto: a combinatorial lexical semantic database for Dutch. In Spyns, Peter and Odijk, Jan, editors, *Essential Speech and Language Technology for Dutch. Results by the STEVIN-programme*, volume XVII of *Theory and Applications of Natural Language Processing*, chapter 10. Springer, Berlin, Germany. URL.

Vossen, Piek, editor. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.

Westerhout, Eline and Odijk, Jan. (2013). Metadata for tools: creating a CMDI profile for tools. presentation held at CLIN 2013, Enschede, Jan 18, 2013. PDF, January.