Data curations by the Dutch Data Curation Service

Henk van den Heuvel, Nelleke Oostdijk, Eric Sanders, Vanja de Lint
CLS / Centre for Language and Speech Technology (CLST)
Erasmusplein 1, Radboud University Nijmegen, The Netherlands
E-mail: {h.vandenheuvel|n.oostdijk|e.sanders|v.delint}@let.ru.nl

Keywords: language resources; sustainable infrastructure; data curation; research data management

1. Introduction

In line with developments we see at the European level, in CLARIN-NL (Odijk 2010) substantial efforts are made to contribute towards the development of an infrastructure that will support the sharing and re-use of resources, and that will open up new avenues of research as it allows for combining various resources in new and unforeseen ways. Apart from work on the implementation of the technical part of the infrastructure, there are several resource curation and/or demonstration projects which should bring this infrastructure to life and promote its actual use. The Data Curation Service (DCS) hosted at the Centre for Language and Speech Technology in Nijmegen is a centre of expertise set up to assist researchers, especially those without the time, money, or know-how, in preparing their data for delivery to one of the CLARIN centres that operate as hubs in the CLARIN infrastructure (Oostdijk & van den Heuvel, 2012). Data curation involves digitizing data (where necessary), converting the data so as to conform to CLARIN accepted standards or preferred formats, providing metadata and adding documentation. The DCS acts as intermediary between the researcher and the eventual data centre.

In this contribution we focus on the data curations that were carried out by the DCS so far, and show the diversity of language resources taken care of.

2. Data curation

In the two years that the CLARIN-NL DCS has now been operational, its focus has been on the curation of data collections residing with and used by individual researchers or research groups in the Netherlands. Candidates for curation are identified and for each it is assessed as to (1) whether it is desirable to have the resource curated and (2) whether successful curation is feasible. A more elaborate description of how these criteria can be operationalized is given in Oostdijk et al. (2013).

Most of the data collections targeted by the DCS are collections that were compiled in projects that were already finished and of which many did not receive any follow up, so that in effect the data were at risk of being lost. Curation of such collections can be challenging, especially when they were created in a context where little or no thought was given to the idea of sharing or re-use. Often IPR has not been settled or if it has, the arrangements did not anticipate the distribution or wider use of the data. Typically data formats are diverse, metadata and documentation incomplete. Since settling IPR for already existing collections was deemed problematic, the DCS has refrained from taking on the curation of resources for which any IPR issues remained to be settled.

The curation of resources involves various actions which can be summarized as follows:

- Data collection: obtaining and agreeing upon the complete and final set of data;
- IPR check: so as to make sure the data can be published in CLARIN context
- Conversion of data formats into standard formats of CLARIN;
- Anonymization of the data; this is typically done in transcriptions, metadata and file names;
- Finding an appropriate CMDI metadata profile (http://www.clarin.eu/CMDI) and modifying it where necessary;
- Filling the metadata profile with the metadata belonging to the database;
- Writing documentation and the curation report; the curation report addresses the above issues and describes the steps taken to curate the resource along these lines;
- Packaging and delivery at a CLARIN data centre. The data centre takes care of adding persistent identifiers and storage of the curated database.

3. Curated language resources

So far a variety of language resources have been curated by the DCS. We will report on the curation of the databases by grouping them into several categories following the language resources typology presented in Gavrilidou et al. (2012):

1. Lexical resources: Dialect databases
2. Multimodal and multilingual corpora: Language acquisition databases
3. Oral/spoken corpora: IPNV interviews

3.1 Lexical resources: Dialect databases

The dialect resources were delivered in various formats including exports of MySQL, MS Access, and FileMaker

Pro. None of these formats is an accepted CLARIN format, the LMF format, however, is. LMF stands for Lexical Markup Framework and is an XML standard which is typically suited to capture hierarchical lexicon structures (Francopoulo, 2013). We departed from a first LMF model used in the COAVA project[1] and made an extended version of this. Our LMF model is based on three head features associated with Lexical Entry.

- Form
- Sense
- Location

Two further features are Definition and Context (both positioned under Sense). Each feature is linked to an ISOcat[2] data category (see Windhouwer & Wright, 2013) as shown in Table 1. Only Form Keyword is mandatory.

LMF feature	**Corresponding ISOcat element**
Form Keyword=	278 keyword
Form Representation AggregatedKeyword=	278 keyword
Form Representation Lexvariant=	5585 lexical variant
Form Representation Morphologicalvariant=	5758 morphological variant (new, defined by DCS)
Form Representation GrammaticalInformation=	2303 grammatical unit
Form Representation Dialectform=	1851 geographical variant
Form Representation Phoneticform=	1837 phonetic form
Form Representation standardizedform=	1851 geographical variant
Sense lemma-id=	288 lemma identifier
Sense Lemma=	286 lemma
Sense Meaning=	464 sense
Definition Definition=	168 definition
Definition sourcelist= sourcebook=	5759 source list (new, defined by DCS) 471 source
Definition sourcelistnumber= sourcebookpage=	5760 souce list number (new, defined by DCS) 4126 pages
Context Timecoverage=	3664 Time coverage
Context Example=	3778 example
Context Comment=	4342 Comment
Location Place=	3759 source
Location Area=	3814 region
Location Subarea=	3814 region
Location informant-id=	3597 speaker id
Location kloeke=	3651 Kloeke geo-reference

Table1: LMF features in the LMF model for dialect databases and corresponding ISOcat elements

We could capture all dialect databases in this framework. All databases were converted to Excel which was considered the intermediary format. Excel files can be converted and imported by tools that are typically used by dialectologists. Care was taken that all data was encoded using UTF-8. The databases were exported as tab separated text files and converted to LMF by means of a Perl script. Phonetic transcriptions (as found to occur in the WBD (Dictionary of the Brabant Dialects[3]) and the WLD (Dictionary of the Limburgian Dialects[4]) were preserved in SIL IPA.

Metadata for each lexical database was entered in the WND profile[5], a CMDI profile created for the COAVA project (Cornips et al. 2011).

In this way the following dictionaries were curated:

- WLD and WBD part III (Dutch dialect dictionaries from Brabant and Limburg)
- Woordenboek Gelderse Dialecten, Rivierengebied
- Woordenboek Gelderse Dialecten, Veluwe

[1] http://www.meertens.knaw.nl/coavasite/

[2] http://www.isocat.org/

[3] http://dialect.ruhosting.nl/wbd/index.htm

[4] http://dialect.ruhosting.nl/wld/index.htm

[5] See http://catalog.clarin.eu/ds/ComponentRegistry/#

- Melis-van Delst (2011) Bikse Praot. Prinsenbeeks Dialectwoordenboek. (Dialect dictionary of the town Prinsenbeek in Brabant)
- Swanenberg, A.P.C. (2011). Brabants-Nederlands Nederlands-Brabants: Handwoordenboek. (Dictionary Brabantic-Dutch, Dutch-Brabantic)
- Panken, P.N. (1850) Kempensch taaleigen. (Dialect dictionary of the town Bergeijk in Brabant)
- Hendriks, W. (2005) Nittersels Wóórdenbuukske. Dialect van de Acht Zaligheden. (Dialect dictionary of the town Netersel in Brabant)
- Laat, G. de (2011) Zoo prôte wèij in Nuejne mi mekaâr. (Dialect dictionary of the town Nuenen in Brabant)
- Bergh, N. van den, et al. (2007) Um nie te vergeete. Schaijks dialectboekje. (Dialect dictionary of the town Schaijk in Brabant)

All curated dialect databases were transferred to the Meertens Institute where they were assigned persistent identifiers and stored.

3.2 Multimodal and multilingual corpora: Language acquisition databases

LESLLA

The LESLLA corpus was collected between 2003-2005 in the framework of the research project *Stagnation in L2 acquisition: under the spell of the L1?* sponsored by NWO (the Dutch Organisation for Scientific Research). The corpus contains valuable data for studying low-educated second language and literacy acquisition, but had been lying idly on the shelf ever since the project came to an end.

The main research question in the project was to what extent the first language impeded the acquisition of the second language in the tutored context of a language course. The participants in the original study had to carry out five tasks which all involved spoken language but varied from strictly controlled to semi-spontaneous. The recordings took place in three cycles of about 6 months each. In each cycle the same tasks were repeated by each participant. The recordings of one cycle were done in three separate sessions (in order to avoid an overload for the participant). Thus there were 9 recording sessions per participant over a period of 1.5 years.

The data was stored on 135 DVDs in PRAAT[6] collection format, which is a text-based format with both the speech signal and the annotation. The files were split into MS riff wave files and PRAAT TextGrids. The TextGrids were converted to ELAN[7] transcription files by using ELAN's export function. The database was restructured into sessions with the structure Task/L1/Speaker/Cycle. All files were renamed in the same structure, using a fixed format in such a way that each file could be uniquely identified by its name. As only first names were used in the database there was no need for anonymisation.

The metadata profile for LESLLA was adapted from the DBD (see below, next section). The metadata was stored in an MS Excel file and CMDI files were created using a python conversion script.

The data has been made available through one of the CLARIN data centres: the Max Planck Institute in Nijmegen. For the time being it is available in IMDI[8] format[9] and at the end of 2014 it will be available in CMDI format.

DBD/TCULT

The Dutch Bilingual Database (DBD) is a rather substantial collection of data (over 1,500 sessions) from a number of projects and research programmes that were directed at investigating multilingualism. It comprises data originating from Dutch, Sranan, Sarnami, Papiamentu, Arabic Berber and Turkish speakers. At the basis of the collection lies the research project TCULT (1998-2002) in which intercultural language contacts in the Dutch city of Utrecht were studied. Many more bilingual datasets collected over the period 1985 – 2005 were later added to the database.

The DBD corpus was stored at the Max Planck Institute with the metadata in IMDI format. During the curation process, missing CHAT[10] files (i.e. files that belonged to the database but had not before been included), were added. Because all data was already in CLARIN approved format, there was no need for any data conversion.

A new DBD metadata profile was set up in CMDI, based on the existing IMDI profile. A shell script was created to convert the IMDI files to CMDI files. Where necessary information was made consistent and missing information (e.g. about file sizes) was added. New ISOcat elements were introduced that were submitted to the ISO committee for formal approval.

The database will be available at the MPI in CMDI at the end of 2014.

3.3 Oral/spoken corpora: IPNV interviews

The IPNV Corpus is a corpus originally compiled by the Veteraneninstituut (VI). It comprises a collection of more than 1,100 (recorded) interviews with veterans who were involved in wars and other military actions that the Dutch military forces took part in. The average duration of an interview is 2.5 hours. Most interviews are with veterans of World War II, the decolonization wars with Indonesia and New Guinea, the UN action in Korea, the UN observe mission in Lebanon, UN missions in Cambodia and former Yugoslavia, and the NATO missions in Iraq and Afghanistan. Some 100 interviews are with veterans who

[6] http://www.Praat.org
[7] https://tla.mpi.nl/tools/tla-tools/elan/
[8] http://www.mpi.nl/IMDI/
[9] https://corpus1.mpi.nl/ds/asv/?openpath=node:1893295
[10] http://childes.psy.cmu.edu/

were involved in small-scale observation, monitoring and humanitarian missions.
In the INTER-VIEWs project[11] 246 of the interviews were curated: the audio recordings (in riff wav format) of the interviews were transferred to DANS[12] and the metadata were made available in CMDI/ISOcat format adopting the profile *OralHistoryInterview* in CLARIN's component registry. The data and metadata can be accessed through the DANS EASY system.

For the remaining interviews all recordings are in WAV format as well. They have also been transferred to DANS by the Veteraneninstituut. With these data, some metadata (at least covering Dublin Core categories) is available. The Veteraneninstituut has provided additional metadata (in an MS Access database) such that the metadata are comparable (and thus compatible) with the metadata for the 246 interviews that were curated in the INTER-VIEWs project.

Around 950 interviews were curated (including an update of the 246 previously curated interviews). All corresponding CMDI metadata files were delivered at DANS. DANS has been authorised to publish various aspects of the metadata in accordance with their agreement (Convenant) with the Veteraneninstituut.

4. Conclusion

So far the DCS has focused on existing data collections which means that most of its efforts have been directed at trying to make the resources conform to CLARIN preferred formats, allowing for their integration in the larger CLARIN infrastructure and the application of various services offered within this infrastructure. Thus one could say that the DCS has been working on curating a backlog of resources that were created in the past. In the near future, however, we expect the task of the DCS to change in the light of the evolving vision of an infrastructure for language resources. In a recent position paper presented at LREC2014 (Oostdijk & Van den Heuvel, 2014), we have described our view on the future role for researchers and other stakeholders in the data curation process in the evolving language resources infrastructure .

5. Acknowledgement

The research for this paper was funded by CLARIN-NL (http://www.clarin.nl) under grant numbers CLARIN-NL-10-025 and CLARIN-NL-11-005.

6. References

Calzolari, N.; Quochi, V. and Soria, C. (2014) *The Strategic Language Resource Agenda*. http://www.flarenet.eu/sites/default/files/FLaReNet_Strategic_Language_Resource_Agenda.pdf. Retrieval date: 20 March 2014.

Cornips, L.; Kemps Snijders, M.; Snijders M.; Swanenberg, J. and de Vriend, F. (2011). Bridging the gap between first language acquisition and historical linguistics with the help of digital humanities. *Proceedings Supporting Digital Humanities.* Copenhagen, 17-18 November 2011. Retrieved from: http://www.meertens.knaw.nl/coavasite/wp-content/uploads/2011/11/Paper-SDH.pdf

Gavrilidou, M.; Labropoulou, P.; Desipri, E.; Piperidis, S.; Papageorgiou, H.; Monachini, M.; Frontini F.; Declerck, T.; Francopoulo, G.; Arranz, V. and Mapelli, V. (2012) The META-SHARE Meta Schema for the description of language resources. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC-2012,* Istanbul, Turkey.

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

Oostdijk, N. and Van den Heuvel, H. (2012). Introducing the CLARIN-NL Data Curation Service. In *Proceedings of the Workshop Challenges in the management of large corpora*. *LREC2012,* Istanbul, 22 May 2012. http://www.lrec-conf.org/proceedings/lrec2012/index.html. Retrieval date: 20 March 2014.

Oostdijk, N.; Van den Heuvel, H and Treurniet, M. (2013). The CLARIN-NL Data Curation Service: Bringing Data to the Foreground. *The International Journal of Digital Curation,* Vol. 8, Issue 2, 134-145.

Oostdijk, N. and Van den Heuvel, H.(2014). The Evolving Infrastructure for Language Resources and the Role for Data Scientists. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC-2014, Reykjavik.*

Sanders, E.; Van de Craats, I. and De Lint, V. (2014). The Dutch LESLLA Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC-2014, Reykjavik.*

Van den Heuvel, H.; Sanders, E.; Rutten, R. and Scagliola, S. (2012). An Oral History Annotation Tool for INTER-VIEWs. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC-2012*, Istanbul, Turkey.

Windhouwer, M. and Wright, S.E. (2013). LMF and the Data Category Registration: Principles and application. In: G. Francopoulo (ed.): *LMF Lexical Markup Framework.* Chapter 3. Wiley-ISTE. ISBN: 978-1848214309

[11] Project funded by CLARIN-NL under grant number CLARIN09-015.

[12] DANS (Data Archiving and Networked Services) is one of the Dutch CLARIN centres. See also http://www.dans.knaw.nl