Curation report

LESLLA corpus

CLARIN-NL Data Curation Service

August 2013

Maaske Treurniet, Henk van den Heuvel, Vanja de Lint
CLST, Radboud University Nijmegen

Executive summary

The LESLLA corpus dates from 2003-2005 and contains speech of 15 low educated learners of Dutch as a second language. The learners are all women. Eight learners have a Turkish background, seven learners have a Moroccan background. From these learners, data were collected over time in three cycles, with an interval of 5 months. In each cycle the participants took part in three types of tasks: (1) production tasks, (2) perception tasks, and (3) a perception task with a metalinguistic component. The production tasks included open elicitation (14,000 utterances), closed completion (4,000 utterances) and imitation (6,000 utterances). Apart from the audio files (WAV format), the data consist of orthographic transcriptions (PRAAT TextGrid files). For all 15 learners there are also metadata available.
Curation of the LESLLA Corpus involves making the production part of the corpus CLARIN-compliant, better searchable and accessible through different platforms (PRAAT, ELAN).[1] To this end the transcriptions that are available in the Praat TextGrid format were converted to ELAN EAF format. A suitable CMDI metadata schema was created to accommodate all available relevant metadata. Where necessary new concepts were created in ISOcat[1]. Ineke van de Craats is principal researcher with respect to this dataset and curation steps were taken in consultation with her. The metadata are made available in CMDI. As the designated archive, MPI will provide the persistent identifiers. A harvest and accessibility test will be conducted in collaboration with the MPI.

1. The LESLLA corpus

In the LESLLA corpus, three cycli of five tasks are available from 15 participants. This means each participant performed each task at each of the three different periods in time. Seven of the participants have a Moroccan Arabic language background and eight have a Turkish language background.

The five tasks are:
1. Sentence imitation
2. Discourse
3. Quest - narrative
4. Snowman - narrative
5. Father & Daughter - narrative

The 7 Moroccan participants are:
1. Fatima
2. Mina

[1] https://catalog.clarin.eu/isocat/interface/index.html

3. Najat A
4. Najat Z
5. Nezha
6. Soad
7. Zohra

The 8 Turkish participants are:
1. Ayfer
2. Emine
3. Hatice
4. Hilal
5. Hulya
6. Nazife
7. Ozlem
8. Zilfi

Explanation results in terms of CEF (Common European Framwork).
A1 and A2 are proficiency levels related to the *Common European Framework of Reference for Languages: Learning, teaching, assessment. (Council of Europe 2001) Cambridge: Cambridge University Press* (and the Dutch version: *Raamwerk NT2 , 2003*).

Explanation on literacy scores.
The literacy levels below A1 are defined in the Literacy Framework Dutch as a second language (Stockmann & Dalderop, 2005): literacy levels A, B and C.
W.Stockmann & K. Dalderop (2005) Portfolio Alfbetisering NT2. Arnhem: Citogroep.
Low-literate (functionally non-literate) is a person who has trouble understanding what he or she is reading; when writing, spelling is slow and not adequate.

2. Curation report

1) Restoring data

The data were copied from three hard disks to a server. However, some problems were encountered with diacritics in file names and corrupt files. To restore all original LESLLA data on a server, the back-up DVDs were used to retrieve the files that were lost during the copying process. A check, using checksum[2], was done to compare the files on the original hard-disks with the data on the server to make sure no files were lost during the process.

[2] http://msdn.microsoft.com/en-us/library/ms189788.aspx

The following (parts of) directories do not exist because the corresponding data are non-existent:
Directories
FatherDaughter/Turkish/Nazife/Cycle3
Quest/Moroccan/Fatima/Cycle1
Quest/Turkish/Hilal/Cycle1
Quest/Turkish/Ozlem/Cycle3
Parts of directories
Discourse/Turkish/Ayfer/Cycle1
Discourse/Turkish/Nazife/Cycle2
Discourse/Turkish/Nazife/Cycle3
Sentence imitation/Turkish/Ayfer/Cycle2

2) Setting up a metadata profile

To harmonize the metadata with other ESF corpora, metadata categories were renamed and restructured. Part of the metadata was consistently divided among the different metadata categories. Part of the metadata was re-phrased, because the original phrases were personal notes. Participant's surnames were removed from the metadata.
All metadata from different sources were converted from plain text files to a table and translated into English. Metadata categories were translated to English in the most appropriate ISOcat data category. For example, "schoolopleiding" was renamed as "Education" and "contacten" als "FriendsStructure" according to the information the respective field contained.
Recording dates were reconstructed with the help of the researcher's logbook. For creating CMDI files, the metadata profile that had been created for the Dutch Bilingual Database (another DCS curation project) has been used as a basis and has been adapted to fit the specifics of the LESLLA database.
The LESLLA metadata profile can be viewed through the following link: http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1375880372947. An example CMDI file is included in the appendix of this document.

3) Restructuring the database

In its original form the database was structured in either one of the following ways:

1. hard-disk of original file location - task - cyclus - nationality - participant - session - files.
2. hard-disk of original file location - task - nationality - participant - cyclus - session (- parts) (-part) - files

The structure of the database has now been adapted to a logical order from a user's point of

view. Data is divided in the order of:

experimental task - language background of participant - actor- cyclus - files.

Users of the curated version will of course typically access the database via the CMDI metadata files so that the actual directory structure is not relevant to them.

Files have been systematically renamed for a uniform, consistent result. We have followed the directory structure and used abbreviations. For example f_m_s_3_020 is used for the 20th utterance (020) of task Father & Daughter (f_), performed by the Moroccan (m_) participant Soad (s_) in the 3rd cycle (3_). An overview of abbreviations used follows.

Overview of abbreviations used

Experimental tasks	Abbreviations used
Father & Daughter	f
Quest	q
Sentence imitation	z
Snowman	s
Discourse	d
Language background of participant	Abbreviations used
Moroccan	m
Turkish	t
Actor	Abbreviations used
Fatima	f
Mina	m
Najat A	na
Najat Z	nz
Nezha	ne
Soad	s

Zohra	z
Ayfer	a
Emine	e
Hatice	ha
Hilal	hi
Hulya	hu
Nazife	n
Ozlem	o
Zilfi	z

4. Converting formats

All files in LESLLA were in a PRAAT .collection format. The .collection format is a PRAAT compatible package of a .wav file and a transcript in PRAAT (.textgrid format).
First, the total number of 14289 transcripts in .praat.collection format were converted into ELAN's .eaf format. In order to convert the files to .eaf format, first the .collection files had to be renamed to .textgrid files. A copy of the dataset was renamed to .textgrid. The conversion from .praat.collection.textgrid to .eaf was done by ELAN software. Several .textgrid files can be imported together through the function File > Import multiple files as... > Textgrid.
We then used a script to unpack the original .collection files into separate .wav and .textgrid files. Both .wav and .TextGrid files have been made available in the curated database together with the corresponding ELAN .eaf files that were obtained through conversion of the PRAAT .collection files. The original PRAAT .collection files remain available in the database, too, but are not included in the metadatafiles. Therefore, four files exist per utterance.
To illustrate, for the example utterance f_m_s_3_020 the database contains f_m_s_3_020.Collection, f_m_s_3_020.eaf, f_m_s_3_020.wav and f_m_s_3_020.TextGrid.

5. Accessibility

The LESLLA database will be publicly available and freely accessible to the public. This applies to transcriptions and mediafiles, as well as metadatafiles. For questions please contact clst@let.ru.nl, or more specifically i.v.d.craats@let.ru.nl and r.v.hout@let.ru.nl. For direct access, contact the MPI.

6. Persistent identifiers

MPI will provide the Persistent Identifiers for the LESLLA data collection.