Curation status report

TOC & RoE

CLARIN-NL Data Curation Service

February 2014

Vanja de Lint, Eric Sanders, Henk van den Heuvel
CLST, Radboud University Nijmegen

1. The Material

1.1 The TOC corpus

The project “Traces of contact: Language contact studies and historical linguistics” aimed to establish criteria by which results from language contact studies can be used to strengthen the field of historical linguistics. It did so by applying the scenario model for language contact studies to a number of concrete settings, which differ widely in their level of aggregation and time depth: the languages of the Amazonian fringe in South America, the complex multilingual setting of the Republic of Suriname, the multilingual interaction of immigrant groups in the Netherlands, and two groups of multilingual individuals. New methods from structural phylogenetics were employed, and the same linguistic variables (TMA and evidentiality marking, argument realization) have been studied in the various projects. These projects made use of a shared questionnaire, so that comparable data was gathered. By applying the scenario model at various levels of aggregation, a more principled link between language contact studies and historical linguistics has been established.

The TOC project consisted of 4 subprojects: Multilingual Processing, Multilingual Netherlands, Suriname and the Carribean, and South America. An inventory of data for all subprojects was done at the start of the curation (see Appendix file “ToCdatabasesOverview-tm 2013-11-05”).

The global objective of the Suriname Group within the Traces of Contact project was to investigate to what extent and in which way the grammatical systems of the various languages spoken in the highly multilingual Republic of Suriname have converged, and in which way the changes these languages have undergone result from language contact or are internally motivated. The Suriname Group of the Traces of contact project has investigated language contact in a contact scenario of a medium time depth. In terms of methodology, the Suriname component therefore falls squarely within the range of contact linguistics as an academic discipline and is less concerned with historical and comparative linguistics. Like the other TOC Research Groups, the Suriname Group has focused on tense, mood and aspect as well as argument realization. Research concentrated on altogether seven languages of Suriname: Sranan, Surinamese Dutch, Suriname Hindustani (Sarnámi), Ndyuka, Surinamese Javanese, Hakka. For comparative purposes and as part of the component "Heritage Languages in the Netherlands", Sranan, Sarnámi and Hakka data are also gathered in the Netherlands. The Suriname Group has two fully associated researchers: Kofi Yakpo (Postdoctoral Researcher) and Robert Borges (PhD researcher).

1.2 The RoE corpus

The project 'The Roots of Ethnolects' was founded in 2005 by Pieter Muysken from the Radboud University Nijmegen and Frans Hinskens from the Meertens Institute & the VU University Amsterdam. It is mainly funded by the Netherlands Organisation for Scientific Research (NWO). This project aimed to explore the roots of ethnolects, which result from the interaction between second language acquisition, multilingual language use, and ingroup/outgroup dynamics in urban settings. For this purpose, the interactions of members of three ethnic groups were studied in two cities: Dutch spoken by adolescents with a Dutch, Turkish or a Moroccan background in either Nijmegen or Amsterdam. The aim was to find out which features are characteristic for these ethnolects and how to explain them. Therefore the researchers looked at the influence of the mother tongue and processes of second language acquisition, but also at the influence of dialects like the Amsterdam dialect and the dialect of Nijmegen. Furthermore, they were interested to see if (some of) these features spread to peers with a non-immigrant Dutch background. The DCS started with the curation of Linda van Meel’s data within the Multilingual Netherlands Group.

2. Curation report

1) Restoring data

For the Suriname subproject, 135 transcription files and 252 audio files were transfered from Kofi Yakpo. Original data are stored on Applejack:
vol/bigdata/users/treurniet/TracesOfContact/Suriname/

Bob Borges’ data are stored on Applejack:
vol/bigdata/datasets/DCS/TOC/Borges_SurinameData

For the Roots of Ethnolect subproject, there were 277 audio files, totalling about 88,8 GB. The smallest file is 1,34 MB, the largest one 804 MB. On average, most files are between 500 and 600 MB. Additionally there are transcription files, protocol files and metadata files.
All ROE data are stored on Applejack (/vol/tensusers/esanders/DCS/ROE/).

The following scripts for processing, reorganizing and renaming files were created and can be found on Applejack: home/vdelint/DCS/TOC or home/vdelint/DCS/ROE.

TOC:
ConvertTranscriptions.py
CopyFiles.py

ROE:

CopyFiles.py
MakeCMDIs.py

2) Setting up a metadata profile

Based on our experience with similar databases (DBD, LESLLA) we set up a metadata profile in the CLARIN component registry named “ROE” for Linda’s data. The profile can be viewed through the following link:
http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1381926654446
The list of metadata categories was originally established by Maaske Treurniet, in consultation with Linda van Meel and other researchers of the TOC project. It has been updated to accommodate the ROE data specifically. The metadata were transfered to us in the form of an sql database. Linda van Meel used a web-based interface to enter all metadata into that database. The categories were labeled in Dutch and have been translated into English. Somehwere in the curation process, an extra metadata category was added by Maarten van de Peet (from the Meertens Institute that hosts the web interface that Linda was using) about the recording settings (“stereo: speaker 1 left, speaker 2 right”, “stereo: speaker 1 right, speaker 2 left”, or “mono”); the values entered into this category have been placed under “Resources>MediaFile>RecordingConditions”.

A similar metadata profile will be suitable for Kofi and Bob’s data (and most likely for data from other TOC subprojects).

3) Restructuring the databases

Kofi’s data were originally structured in the following way:

TOC>Suriname>recorded_in_country>target_language

recorded_in_country was either one of the following three options:
recorded_in_mauritius
recorded_in_netherlands
recorded_in_suriname

target_language was either one of the following options:
bhojpuri
kreyol
javanese
sarnami

sranan
ndyuka
saramaccan
dutch
hakka
hindoestaans?

divided over the recording countries in the following way:
recorded_in_mauritius:
- mauritian_bhojpuri
- mauritian_kreyol

recorded_in_netherlands:
- java_javanese
- sarnami
- sranan

recorded_in_suriname:
- ndyuka
- saramaccan
- sarnami
- sranan
- suriname_dutch
- suriname_hakka
- suriname_javanese

NB: HIN=hindoestaans? = Sarnami?

Filenames were codes based on target language, year of recording, the first three letters of the speaker(s)'s first name and the task

f.e. SAR11aar_fe stands for **SAR**NAMI 20**11** **aar**i? **f**rog story **e**licitation

Kofi's data are to be restructured in the following way: language - country of recording - speaker(s). Each folder is then to contain all corresponding recordings (divided by the different tasks), both audio and transcription files.
There is no perfect correspondence between audio and transcription files. Many transcriptions are missing, there are mismatches in name, and some files contain references to unknown tasks (such as "metadata_nickerie" and "sickwoman"). This still has to be sorted out, preferably with help of Kofi.

For Roots of Ethnolect, the data are structured as follows:

```
ROE > eafs > eaf files
      metadata > 3 csv files (‘gesprek’, ‘locaties’, ‘spreker’)
      protocollen > xls files
      wavs > wav files
```

The wav, xls and eaf files are linked in name:[code speaker1]-[first name speaker1]_[code speaker2]-[first name speaker2].[extension]. For example: n12m02-amine_n12d02-ricardo.eaf. Where the code for the speaker is built up in the following way:
[location: Amsterdam (a) or Nijmegen (n)][age group: 12 or 20][language background: Dutch (d), Moroccan (m), Turkish (t), or control (c)][subject number within pool: ‘00-06’, or ‘0x’ if just serving as interlocutor to other subject in this session].

Inconsistencies in the filenames such as in capitalization, or in spaces, hyphens and underscores, were removed, but some issues remain solved. There are cases where it is unclear which eaf files belong to which wav files. For example, n12m02-amine_n12d02-ricardo.eaf and n12m02-amine_n12d02-ricardo_02.eaf are linked to n12m02-amine_n12d02-ricardo_02.wav, but alongside exist n12m02-amine_n12d02-ricardo_01.wav, n12m02-amine_n12d02-ricardo_04.wav and n12m02-amine_n12d02-ricardo_05.wav to which no .eaf file seems to be linked. The script for the creation of cmdi files will put all eafs, wavs and protocol xls belonging to one session (dialogue) into one folder, leaving the unclarities about the correspondence between those files as they are.
The script has yet to be finished.

.

4) Converting formats

Kofi’s data:
Transcription files were in .doc, .docx, .rtf or .txt. We chose .txt as the target format and converted all .doc/.docx/.rtf transcription files into .txt format.

Audiofiles existed in either .wav, .mp3 or .WMA format, sometimes in multiple formats. All mp3 files had a .wav alternate with the same name; Kofi had explained that the mp3’s had only been created for the purpose of annotation. We chose .wav as the target format for the audiofiles, because it transfers better quality and was already available for the vast majority of files.

Bob’s data still have to be fully inventarised, in order to see if converting of formats has to be done.

ROE data:
The audio files and transcription files were using CLARIN standard formats (wav and eaf). The protocol files are in .xls,and may have to be converted. The metadata are to be extracted from

the .csv files and copied into the appropriate cmdi files through the script. The script was in the making at the time of closure of this report.

5) Anonymization and accessibility

Kofi's data:
Transcription files were anonymized by filtering out the names of the speakers: we searched for the names, as they appeared in the metadata file, and replaced them with a three-letter code between brackets (f.e. [aar]), where the three-letter code was taken from the filenaming code for the speaker in question, as used originally by the researcher.
Although Kofi signed the IPR forms and authorised DCS to curate the data, he would ideally be contacted to help us answer the remaining questions, which are:

- about the project

All the audiofiles are here stored within a directory named "suriname" (is that the official name of your subproject?), but are then divided over 3 directories named "recorded_in_mauritius", "recorded_in_netherlands" and "recorded_in_suriname". How are all these collections related exactly? What do the data recorded in Mauritius have to do with the languages spoken in Suriname? Or how is Javanese, recorded in the Netherlands, relevant? Do all of these data belong to the same Traces of Contact project? Can you please explain for the purposes of curation?

- transcriptions

There are only about half as many transcriptions (135; of which at least 24 are doubles in a different format) as there are audiofiles (252). Is that correct?
There are no English translations. Is that correct?
Some of the transcriptions contain comments by the transcriber, directed to you for example. Do you want to keep those there?
In december 2013 I was told by Linda van Meel that Luis Miguel Rojas Berscia was working on transcriptions of your haka-data. Is that supposed to be included in this database? If so, can you provide us with all the necessary information?

Bob's data: contact has to be sought for further curation (a.o. extraction of metadata from audiofiles).

ROE:
It was agreed upon by Frans Hiskens, Pieter Muysken, Linda van Meel and the DCS that the ROE database ought not to be made publically available until after the defense of Linda van Meel. This in order to protect the authenticity of her research. After that, all data can be made publicly available; first names are considered not to be problematic due to the time that has gone by

since the recordings.

6) Persistent identifiers

Persistent identifiers will have to be assigned by the envisaged data centers. These are Meertens for RoE and the MPI for ToC.