The Dutch LESLLA Corpus

Eric Sanders, Ineke van de Craats, Vanja de Lint
CLS/CLST Radboud University Nijmegen
[e.sanders, i.v.d.craat, v.delint]@let.ru.nl

Abstract

This paper describes the Dutch LESLLA data and its curation. LESLLA stands for Low-Educated Second Language and Literacy Acquisition. The data was collected for research in this field and would have been disappeared if it were not saved. Within the CLARIN project *Data Curation Service* the data was made into a spoken language resource and made available to other researchers.

Keywords: data curation, spoken language resource, low-educated second language and literacy acquistion

1. Introduction

A lot of research about second language acquisition (SLA) is carried out in the Netherlands and many researchers have collected spoken data. Still, only few SLA data collections are available to other researchers as spoken language resources consisting of sound files with transcriptions, metadata and documentation.

The *Jasmin Corpus* [1] is an extension of the Spoken Dutch Corpus [13] and contains speech of a miscellaneous group of second language (L2) learners (both children and adults, varying mother tongues and education levels) acquiring Dutch or Flemish.

The *European Science Foundation* (ESF) database [15] is a longitudinal and cross-linguistic database. It contains speech data from low-educated adult L2 learners in Western-Europe with six different source languages and five different target languages, collected over 27 monthly recordings. For Dutch as a target language, the source languages are Turkish and Moroccan Arabic. The speakers are relatively young adults learning Dutch with hardly any formal instruction.

Low-educated L2 learners with a low level of literacy form a special population in SLA research. Since researchers become more and more aware of the fact that SLA theory is mainly based on research with high-educated learners more research should be based on low-educates and low/non-literates.

The group of low-educated learners is generally called LESLLA-learners, since in 2005 the first symposium on Low-Educated Second Language and Literacy Acquisition (LESLLA) was held [16]. The data described here represents this population and was named the Dutch LESLLA corpus as an initiative for others to build similar corpora and stimulate research in this area.

The data collection started in 2003 and was completed in 2005. The data was collected 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 main research question in this project was to what extent the first language impeded the acquisition of the second language in the tutored context of a language course. Sub-questions were when and for which aspects this stagnation emerged and what features of L1 and L2 caused this stagnation.

Speech and transcriptions were on DVDs and unavailable for other researchers. In the framework of the Data Curation Service (DCS) the database has been curated and made publicly available. In this paper we report on the design of the corpus, its contents, the curation process and the scientific potential of the data.

2. Corpus Design

2.1 Speakers

The LESLLA corpus contains speech of 15 low-educated learners of Dutch as a second language. All of them are women; 8 are Turkish, 7 Moroccan. Turks and Moroccans are the two largest immigrant groups in the Netherlands. Participants were recruited through adult education centres, where they were taking Dutch classes. They were between 22 and 45 years old. Their time in the Netherlands differed from 0.5 to 26 years, while their schooling in Dutch differed from 0.5 to 2 years. The Turkish women all had received five or six year schooling in their fatherland, while for the Moroccan women this differed from none to seven years. Four of the Moroccan women had done a literary course in Roman script. For three Turkish and three Moroccan women stagnation in learning Dutch was expected by their teachers based on the results in the classroom and on proficiency tests administered during the last school year.

Tables 1 and 2 show an overview of the Moroccan and Turkish participants.

The mean level of spoken language proficiency was A1-A2 in terms of the Common European Framework of References (Council of Europe, 2001) [7].

Participant	Age	Years of schooling in Morocco	Years of L2 schooling in Netherlands	Years in Netherlands	Literacy course Roman script	Stagnation observed/ expected
Mina	23	0	2	4	yes	no
Zohra	41	5	0.5	8	no	no
Soad	34	4	0.5	12	no	no
Najat	25	4	1.5	4	yes	yes
Hayat	22	5	2	2	yes	yes
Nezha	38	0	1.5	3	yes	yes
Fatima	27	7	1.5	5	no	yes
Mean	**30**	**3.6**	**1.4**	**5.4**		

Table 1: Overview of the Moroccan participants.

Participant	Age	Years of schooling in Turkey	Years of L2 schooling in Netherlands	Years in Netherlands	Stagnation observed/ expected
Zilfi	30	5	1.5	11	no
Hülya	19	5	0.5	0.5	no
Emine	28	5	0.5	13	no
Hilal	19	5	1.5	2	yes
Ayfer	37	5	0.5	18	yes
Nazife	31	5	0.5	1	yes
Hatice	45	5	0.5	26	yes
Özlem	31	6	2	5	yes
Mean	**30**	**5**	**0.9**	**9.5**	

Table 2: Overview of the Turkish participants.

2.2 Tasks

The participants in the LESLLA study had to carry out five tasks which all involved spoken language but varied from strictly controlled to semi-spontaneous:

Sentence Imitation: production task, sentence imitation. The subjects heard a sentence that they had to repeat.
Elicitation was directed to: subject realization, possessive constructions, subject-verb agreement, verbs *zijn* (‘be’) and *hebben* (‘have’), subject-verb inversion, order in subclause, object-verb/verb-object order.

Discourse: production task, closed completion. The subjects heard (part of) a sentence that they had to finish. In some cases the answer was narrowed by a picture or one or two words that could be used to finish the sentence. Elicitation: same as Sentence Imitation.

Father and Daughter: production task with open elicitation. The subjects watched a silent cartoon movie [12] and were asked to recount the story and answer questions by the experimenter.

Snowman: production task with open elicitation (additional non-speech tasks were administered). The subjects read a picture story [11] which they then had to repeat orally followed by answering questions by the experimenter.

Quest: production task with open elicitation (additional non-speech tasks were administered). The subjects watched a silent cartoon movie and had to recount what they saw and answer questions by the experimenter. Elicitation was directed to expression of space.

2.3 Time Schedule

The recordings took place in three cycles of about 6 months each. Per cycle the same amount of tasks was performed by each participant. The recordings of one cycle were done in three separate visits (in order to avoid an overload for the participant). In the first visit the tasks Discourse, Sentence Imitation and Father and Daughter were recorded, in the second visit Snowman and in the third visit Quest. This amounts to 9 recording visits per participant over a period of 1,5 years. The first recording was in March 2003, the last in April 2005.

3. Material & Curation

3.1 Audio and Annotation

The original recordings were made on a Philips voice tracer and digitally transferred to a PC, on which it was converted to 16 kHz, 16-bit mono speech files. The data was transcribed orthographically in Praat [10] using separate tiers for the two speakers (experimenter and L2 learner). Special symbols were used to indicate non-speech and dialogue phenomena. The transcription protocol was for a large part following the CHAT conventions [2]. A session (one speaker doing one task) was split into segments of one or more utterances, cutting in natural pauses. All the data was stored on 135 DVDs in Praat collection files, which is a format in which speech and annotation are both stored in text format. These files can only be opened with Praat.

Task	*#files Moroccan*	*duration (s) Moroccan*	*#files Turkish*	*duration (s) Turkish*	*#files* **total**	*duration (s)* **total**
Discourse	630	6146	700	6634	**1330**	**12780**
Sentence Imitation	628	7275	720	8363	**1348**	**15638**
Father & Daughter	1492	12688	1934	13162	**3426**	**25850**
Snowman	2898	21007	3621	22226	**6519**	**43233**
Quest	809	5970	857	6019	**1666**	**11989**
total	**6457**	**53086**	**7832**	**56404**	**14289**	**109490**

Table 3: Number of files and durations of tasks per L1.

3.2 Curation

In October 2011 the Data Curation Service (DCS) started as a CLARIN-NL[5] project to rescue databases and corpora[14], often held by individual researchers, that might get lost otherwise. Within the DCS project, the LESLLA database was curated. That means that the data was converted to formats that are standard in the CLARIN framework and that a metadata profile is designed with ISOcat elements [8] and completed for all sessions. The data then becomes available through one of the CLARIN data centres, in this case the Max Planck Institute in Nijmegen [9].

In the curation process, the collection files were split up into Praat TextGrid annotation files and MS Wave binary audio files (mono, 16 bits/sample, 16 kHz). The TextGrids then were converted to ELAN [4] annotation files. All these formats are stored in the curated database in sessions with the following structure:

```
Task/L1/Speaker/Cycle
```

where
`Task` is *Discourse, SentenceImitation, FatherDaughter, Snowman* or *Quest*
`L1` is *Moroccan* or *Turkish*
`Speaker` is the name of one of the 15 participants
`Cycle` is *1, 2* or *3*

The name of the files is a coded version of the directory plus a serial number for the file. E.g. `d_m_f_2_023.wav` is the 23rd audio file of the second cycle of the discourse task done by the Moroccan Fatima. Table 3 shows the number of files in the corpus and the total duration per task and L1. A file contains typically one to three utterances. The average duration of the Discourse files is 9.5 seconds, that of Sentence Imitation 11.5 seconds. The files of the other tasks are about 7 seconds on average. In total there is 30 hours of material, almost evenly spread over the Moroccan and Turkish speakers. The number of files for the Turkish speakers is larger than that of the Moroccan speakers.
The metadata profile is made with the CMDI component registry [6]. The profile is on the level of a session and contains information about the project, the contents of the session, the speaker and experimenter, the annotation and audio files. The profile can be found at http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1375880372947

All speakers signed a statement that they agreed that the data can be used for research.

The corpus will be available at the CLARIN centre MPI. The metadata can be browsed from the IMDI-browser [3]. For free access to the whole data set, including audio files and transcriptions, individual permission has to be granted by the contact person.

4. Research

The corpus is suitable for all kinds of research like:

- Morphosyntactic research with the narratives whether the focus is on learners' L2 or the L1, e.g. the realization of the subject, the position of the verb, the use and the form of pronouns.
- Research on the non-realisation of difficult elements in syntax: e.g. with articles, complementizers, verbal inflection, pronouns.
- Research on non-native pronunciation (especially the imitation task can be used; the transcription lines up with the speech).
- The language development of low-educates over time.
- Speech technology for L2: e.g. acoustic model adaptation, test data.

Research done with this data so far focused on the morphosyntactic development of the verb in the sentence. The most remarkable findings were that both Turkish and Moroccan learners inserted auxiliaries in positions where a finite verb was expected in L2 Dutch. The two ethnic groups showed a preference for one specific 'dummy auxiliary' [17,18,19,20]. A few of the LESLLA tasks were done by Chinese immigrants and compared to the Moroccan and Turkish data with respect to their problem with inflection in Dutch [21].

The material can also be used for teaching purposes. The data can be used by language acquisition teachers at universities, vocational high schools, speech therapists, all those who want to show what L2 Dutch is at a very basic level.

5. Conclusion

In this paper a corpus is described that was lying idle on the shelf. Thanks to the DCS project it is now available for researchers worldwide. The LESLLA corpus is a complete collection of well annotated speech data from low-educated L2 learners of Dutch. It is suitable for all kinds of different research. We welcome researchers to use the data and to enrich the corpus. For example, a phonemic annotation or lexicon and a text-speech alignment on word or phoneme level would be a useful addition to the corpus.

The purpose of this paper is to describe the LESLLA corpus and its curation, but we would also like to stress the possibilities and importance of data curation in general. We encourage researchers to look for ways to have their (hidden) data curated for use by other researchers. It makes both reproduction of the experiments and new research possible. It will save researchers time and money when they do not have to build their own dataset and the original producers of the data have the satisfaction of their work still being used. The benefits of data curation are plenty.

6. References

[1] Cucchiarini, C., Driesen, J., Van hamme, H. and Sanders, E. (2008) Recording Speech of Children, Non-Natives and Elderly People for HLT Applications: the JASMIN-CGN Corpus. Proceedings of the Sixth International Conference on Language Resources and Evaluation

[2] http://childes.psy.cmu.edu/manuals/chat.pdf

[3] http://corpus1.mpi.nl/ds/imdi_browser

[4] http://tla.mpi.nl/tools/tla-tools/elan/

[5] http://www.clarin.nl/node/147

[6] http://www.clarin.eu/node/3219

[7] http://www.eui.eu/Documents/ServicesAdmin/LanguageCentre/CEF.pdf

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

[9] http://www.mpi.nl/

[10] http://www.praat.org

[11] http://www.thesnowman.co.uk/

[12] http://www.youtube.com/watch?v=MgdsfRDxIeQ

[13] Oostdijk, N., Goedertier, W. , Van Eynde, F., Boves, L., Martens, J.P., Moortgat M., Baayen., H. (2002) Experiences from the Spoken Dutch Corpus Project. Proceedings of the third International Conference on Language Resources and Evaluation: 340-347.

[14] Oostdijk, N., Van den Heuvel, H., Treurniet, M. (2013) The CLARIN-NL Data Curation Service: Bringing Data to the Foreground. The International Journal of Digital Curation, Volume 8, Issue 2

[15] Perdue, C. (ed.) (1993) Adult Language Acquisition. Vol 1: Field Methods. Cambridge University Press

[16] Van de Craats, I., Kurvers, J. & Young-Scholten, M. (eds.) (2006) Low-Educated Adult Second language and Literacy Acquisition. Proceedings of the inaugural symposium – Tilburg 05. Utrecht: LOT Occasional Series 6.

[17] Van de Craats, I. (2007) Obstacles on Highway L2. In N. Faux (ed.) Low-Educated Second Language and Literacy Acquisition. Proceedings of the Second Annual Forum. Richmond, Virginia: The Literacy Institute at Virginia Commonwealth University: 149-163

[18] Van de Craats, I. (2009) The role of IS in the acquisition of finiteness by adult Turkish learners of Dutch. Studies in Second Language Acquisition 31: 59-92.

[19] Van de Craats, I. (2011) A LESLLA corpus: L1 obstacles in the learning of L2 morphosyntax. In Ch. Schöneberger, I. van de Craats, and Jeanne Kurvers (eds.) Low-Educated Adult Second Language and Literacy Acquisition. 6th Symposium Cologne: 33-48.

[20] Van de Craats, I. and Van Hout, R. (2010) Dummy auxiliaries in the L2 acquisition of Moroccan learners of Dutch: Form and function. Second Language Research 26: 473-500.

[11] Oldenkamp, L. (2013) The trouble with inflection for adult learners of Dutch. PhD dissertation, Radboud University Nijmegen