Constituting an oral lexical Database in Laz (South Caucasian)

MOTS-CLES : Laze, Caucase, unités lexicales, linguistique de terrain
KEYWORDS: Laz, Caucasus, lexical units, field-work linguistics

The aim of this presentation is to share some challenges of a work in progress consisting of the creation *ex nihilo* of an oral lexical Database in Laz, an endangered language spoken in Georgia and Turkey. Theorical, empirical, methodological, and digital issues will be considered. This work represents the first attempt at an oral lexical database for a South Caucasian language.

The study focuses on the Laz variety of Georgia, spoken in only one village, Sarpi, on the Turkish border. This variety is critically endangered since only a dozen older adults retain some knowledge of the language. All of the data presented are first-hand data.

1 Collecting the data

We collected the data during fieldwork in Sarpi, on the basis of the questionnaire elaborated in Gérardin (2024) for Georgian dialects. This questionnaire consists in a slideshow showing pictures rather than a written word-list in order to avoid interferences with the vehicular language (here Georgian). We adapted it for Laz, which shows less interferences with Standard Georgian than do the Georgian dialects. Additionally, we selected lexical items both for their relevance to genetic comparison and for the study of diatopic variation.

The recording on which the present study is based includes around 60 items (nouns, verbs, adjective, some grammatical categories, etc.), elicited from a native speaker aged 88. The data were recorded using Zoom H5 and a Samsung HyperDIS 65x camcorder.

2 Data processing

For the present study, we used a recording of 1 hour and 19 minutes, containing 124 lexical units. The first step was to clean the file and isolate the lexemes. For each item on the list, we obtained at least one exploitable corresponding oral realization. We processed the mp4 file in DaVinci Resolve to perform data alignment and to extract the lexical items.

In parallel, we compiled the data in a spreadsheet to create a written database linked to the audio files. The spreadsheet contains the transcription of each item in Laz, Georgian, English, and French, as well as the start and end timecodes for each pronunciation, which we have exported automatically.

3 First results

So far, we have completed the processing of the entire recording mentioned above (124 items). This has allowed us to enrich the word list with their corresponding oral realizations. The following figures provide an overview of the results. Figure 1 shows the alignment of video, audio and textual data, highlighting the isolation of the lexical items, while Figure 2 illustrates an excerpt from the associated spreadsheet:

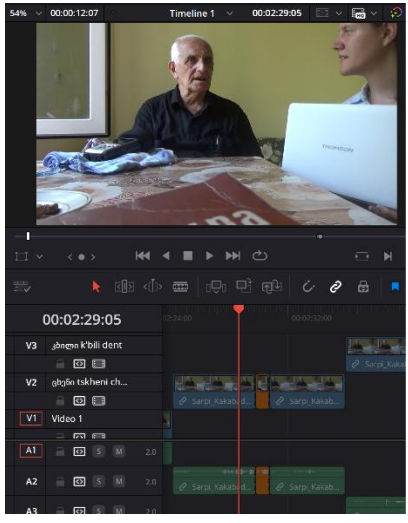

FIGURE 1: Extract from the data managing using DaVinci Resolve

Item	Laz	Georgian	English	French	Timecode_in	Timecode_out
					HH:MM:SS:DSDS	IDEM
2	tskheni	tskheni	horse	cheval	00:02:25:11	00:02:35:08
3	k1ibiri	k1bili	teeth	dent	**previous TC-OUT**	00:03:20:05
4	ley	niori	garlic	ail		00:04:13:21
5	topuri	tapli	honey	miel		00:04:54:13
6	gola	mta	mountain	montagne		00:05:38:24
7	toli	tvali	eye	œil		00:05:57:23
8	mtuti	datvi	bear	ours		00:07:26:03

FIGURE 2: Extract from the associated written spreadsheet

4 Future research

The completion of this initial work marks a decisive step in the automatic processing of spoken South Caucasian languages. The next stage will involve archiving the resulting database. Our goal is to make the most comprehensive word list possible accessible to the scientific community, in the form of a clickable list: clicking on each written word opens the corresponding audio file.

In the medium and long term, this database is intended to be expanded — first with the other recordings we collected during our fieldwork mission (approximately ten hours), and later with future recordings, both in Laz and in other South Caucasian languages. Researchers will be able use it to conduct various types of studies, such as distance algorithms, or diachronic and diatopic comparisons.

References

GÉRARDIN H. (2024). Building a Dialectological Lexical Database of Georgian Cognates for Digital Analysis. In *Digital Kartvelology*, 3, p41–52. DOI : 10.62235/dk.3.2024.8511

LACROIX René. (2008). *Description du dialecte laze d'Arhavi (caucasique du sud, Turquie) : grammaire et textes*. Thèse de doctorat, Université de Lyon.

PONIAVA Natia. (2023). *visc'avlot lazuri* [Let's study Laz], Art'anuji, Tbilisi. (in Georgian)