

# **Project proposal: Corpus of Spoken Modern Western Armenian**

## **1. Short description of the project**

Data collection for a balanced multi-purpose corpus of spoken Modern Western Armenian (MWA), containing 27 different 4,000-word conversation fragments and 30 read scripted samples of sentences with phonetically varied sounds (300 words each).

## **2. Motivation**

A well-constructed corpus of a sufficient size is the main precondition for methodologically sound linguistic research to be carried out. Nowadays, language corpora are used as the basis for reference grammars, they provide natural non-contrived examples for text books and serve as an enhancement of foreign language teaching in “data-driven learning”. In theoretical linguistics, they provide data to study language variation, specific linguistic constructions and language acquisition. Introspective judgements can be used to formulate hypotheses but only a corpus offers means for their objective verification.

Thanks to the current technical developments, automated corpus studies are highly reliable and little time-consuming. While corpora of written texts are often easier to construct (and undoubtedly useful), spoken corpora of natural conversations reflect more truthfully the state of the language: only a small segment of the speakers of a language creates literary texts and news reports but all of them are engaged in spontaneous dialogues. (It has been noted that a strictly ‘proportional’ corpus would have to contain roughly 90% conversations.) Spoken language corpora are also needed for acoustic analyses (e.g., study of sound systems and of prosody) and for pragmatic research (e.g., conversation analysis).

There is currently no existing spoken corpus for MWA that could be used in the sense described above. A creation of the corpus would serve as a motivation and a precedent for local Armenian communities, as well as scholars in the West.

## **3. Design**

### **3.1 Selected variant**

The sociolinguistic situation of Modern Western Armenian is complex. Out of the few communities around the globe, the Syrian Modern Western Armenian has been selected as the representative variant for the corpus. Unlike in Europe and in America, the development of MWA in the Middle East has been continuous (no generations of speakers who learned the language later in life, e.g., from textbooks). All the three major dialects of MWA in Syria (Damascus, Aleppo, Northern Syrian dialect) would be represented in the corpus.

### **3.2 Recordings**

The corpus will consist of digitized recordings in the ‘wav’ format, later to be transcribed in UNICODE and ASCII (with transliteration key), in accordance with the Text Encoding Initiative guidelines. Information regarding the gender, age, level of education, dialect variation and social context of the conversations will be provided in the documentation. Names of the speakers will be edited out, to preserve their anonymity. Prior to the start of the recording session, the speakers will be given in writing a description of the project. Written permissions will be obtained from the speakers regarding the use of the recordings. Information about relevant biographical data will be requested from the speakers (biographical form). In order to enhance naturalness of the final sample, each conversation will be recorded as lengthy as possible, so that both the beginning and the end of the recording, as well as possible disturbances can be filtered out (approximately 70 minutes for 4,000-word samples). For the same reason, recordings will be made in the actual environment

(controlled for low-frequency noise). More data will be collected than what will be used in the end (the estimate is that only 4-5 out of 10 recordings will be usable). The sampling technique will be non-probabilistic, a combination of “judgement” and “convenience” sampling (as used, e.g., in the International Corpus of English). In practice, this means that every effort will be made to collect speech from a balanced group of constituencies but pragmatic decisions will have to be made on the spot.

### **3.3 Structure of the corpus**

Spontaneous conversations exhibit a high rate of internal variation. Therefore, in order to create a representative corpus, collected samples will vary across the following variables: gender, age, level of education, dialect variation, social context and relationships (disparate/equal, intimate/non-intimate).

The corpus will be structured as indicated below, with each dialogue type recorded in all three dialects (in total, 27 recording sessions). Selection of speakers for scripted monologues (10 per dialect) will be balanced with respect to the above noted variables.

#### **A. Dialogues**

##### **a. private**

- i. female-female
- ii. male-male
- iii. male-female-female
- iv. male-male-female
- v. female – child
- vi. male – child

##### **b. public**

- i. academic
- ii. business transactions
- iii. legal setting

#### **B. Scripted Monologues**

**a. Damascus dialect** (10 speakers)

**b. Aleppo dialect** (10 speakers)

**c. North Syrian dialect** (10 speakers)

### **3.4 Computerization**

Digitization of the recordings will be done using a soundboard and CoolEdit (freeware).

### **3.5 Future work**

According to existing studies, 2,000 words take ca 15-20 hours to transcribe. Assuming that the transcribers would be given an appropriate financial compensation, manual transcription of the data is beyond the possible financial scope of the given project (ca 1.080 hours of work in total).

### **3.6 Distribution**

The recordings will be freely available for research and future work. The authors would be happily willing to make copies for research institutions as well as individuals.

#### **4. Recording Equipment**

After a careful consideration of available field recording equipment, taking into account their portability, durability, convenience with respect to subsequent digitization, and most importantly, quality of the recordings, DAT recorders have been given preference over MiniDisc and CD recorders (their compression system distorts some amplitude and frequency components, necessary for acoustic analysis).

- microphone: Audio-Technica AT803b – lavalier condenser
- pre-amp/mixer: Shure FP23
- recorder: USBPre
- media: Maxell DAT tapes (125 min), CD-R

#### **5. Project participants and their expertise**

**Melanie Keledjian** – responsible for communication with the speakers (selection of suitable candidates, organization of the recording sessions)

Native speaker of MWA, Armenian Language teacher at Inalco (Institut National des Langues Civilisation Orientales, Paris) she has taught the Western Armenian Class of the University of Michigan Summer Program in 2001 and 2002.

**Marie Šafářová** – responsible for the technical part of the project (equipment, recording of the sessions, digitization of the recordings).

Currently a PhD student in linguistics at the University of Amsterdam. In 2000-2001 she participated in the development of a Western Armenian text-to-speech synthesis system. In 2001 she received a University of Michigan scholarship to study at the Armenian Summer Language Institute in Yerevan. She has been doing research on MWA morphosyntax.