The Maaloula Aramaic Speech Corpus (MASC): From Printed Material to a Lemmatized and Time-Aligned Corpus

Ghattas Eid, Esther Seyffarth, Ingo Plag
Heinrich-Heine-Universität Düsseldorf
ghattas.eid@hhu.de, esther.seyffarth@hhu.de, ingo.plag@uni-duesseldorf.de

Motivation and corpus content

Western Neo-Aramaic is an endangered Neo-Aramaic (Semitic) variety spoken in the Syrian villages of Maaloula, Jubbaadin, and Bakhaa. 'Maaloula Aramaic' refers to the dialect of Maaloula.

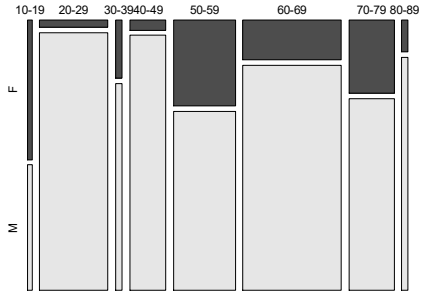
Neither an electronic text corpus nor a speech corpus with time-aligned transcriptions was available. Therefore, we designed the Maaloula Aramaic Speech Corpus (MASC, Eid et al., 2022), the first electronic speech corpus of this variety.

MASC is a multi-purpose corpus available to the scientific community at https://doi.org/10.5281/zenodo.6496714

Data included in the corpus:

1. the transcriptions of tape-recorded narratives that Werner Arnold collected in the 1980s (Arnold 1991a, 1991b):
   - 173 monologues (different text types)
   - 64,845 tokens, 12,220 types
   - tokens per narrative (M = 375, Min = 19, Max = 4,340)
   - 45 speakers (32 males, 13 females) aged 13 - 89 years
   - tokens per speaker (M = 1,441, Min = 42, Max = 10,688)

Distribution of tokens by age and gender



2. the audio files of these narratives, available at the Semitisches Tonarchiv website of Heidelberg University (Arnold, 2003): 173 mp3 files (10 hours)

- Permission to use the data was granted by Harrassowitz Verlag and Werner Arnold.

1. Transcriptions

These are the digitized transcriptions that contain no annotation (except for a few informative tags). The digitization process involved using an optical character recognition (OCR) program and collaborating with a native speaker consultant to produce an error-free text.

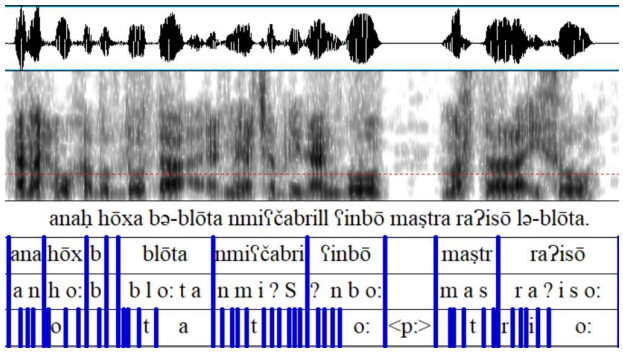
Extract from the MASC dataframe

Root	Lemma	LemmaFreq	Word_Form	Word_formFreq
ʔmr	amar yīmar	1925	amella	185
ʔmr	amar yīmar	1925	amelle	393
ʔmr	amar yīmar	1925	amellen	2
...	...	...	...	...

3. Time-aligned transcriptions

The original and denoised audio files are included in our corpus. They can be opened in Praat (Boersma & Weenink, 2021) together with their TextGrid files to conduct acoustic analyses. For the time alignment process, we used the WebMAUS tool (Schiel, 1999, 2015) by BAS Web Services (Kisler et al., 2017).

Screenshot from Praat displaying the TextGrid tiers



2. Lemmatized transcriptions

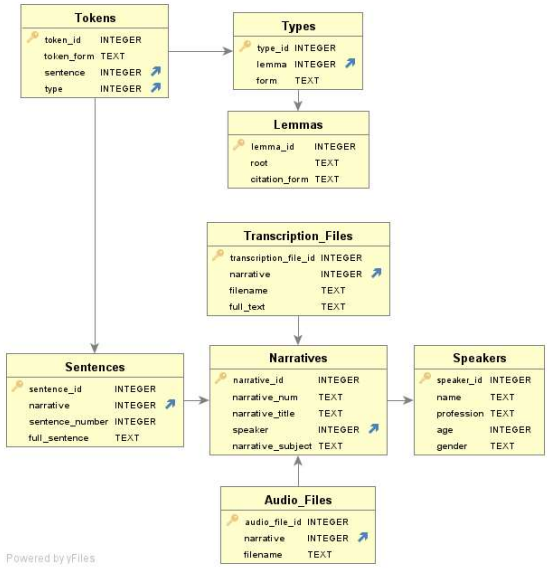
Each word is followed by the citation form of its lemma as it appears in Arnold's (2019) dictionary.

A lemmatized sentence:

```
anaḥ<anaḥ> hōxa<hōxa> bə<b->-blōta<blōta>
nmiʕčabrill<iʕčbar> yiʕčbar> ʕinbō<ʕenəpṭa>
maṣtra<maṣtra> raʔisō<raʔisa> lə<l>-blōta<blōta>.
```

These files are the result of a lemmatization process that was completed with the help of our language consultant.

4. SQLite database



REFERENCES Arnold, W. (1991a). *Das Neuwestaramäische: III. Volkskundliche Texte aus Ma'lūla*. Otto Harrassowitz. Arnold, W. (1991b). *Das Neuwestaramäische: IV. Orale Literatur aus Ma'lūla*. Otto Harrassowitz. Arnold, W. (2019). *Das Neuwestaramäische: VI. Wörterbuch*. Harrassowitz. Boersma, P., & Weenink, D. J. (2021). *Praat. Doing phonetics by computer* [Computer program] (Version 6.1.49) [Computer software]. Version 6.1.49, retrieved from http://www.praat.org/. Kisler, T., Reichel, U. D., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language, 45*(September 2017), 326–347. R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. In *Proceedings of the ICPhS 1999* (pp. 607–610). San Francisco, August. Schiel, F. (2015). A statistical model for predicting pronunciation. In *Proceedings of the ICPhS 2015*. Glasgow, UK, paper 195.

LANGUAGE RESOURCE REFERENCES Arnold, W. (2003). *Semitisches Tonarchiv (SemArch)*. Heidelberg University. Available online at http://semarch.ub.uni-heidelberg.de/. Eid, G., Seyffarth, E., Rihan, E., Arnold, W., & Plag, I. (2022). *The Maaloula Aramaic Speech Corpus (MASC)* (v1.0). Heinrich-Heine-Universität Düsseldorf. Zenodo. https://doi.org/10.5281/zenodo.6496714