Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6513–6520
Marseille, 20-25 June 2022

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

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

Abstract

This paper presents the first electronic speech corpus of Maaloula Aramaic, an endangered Western Neo-Aramaic variety spoken in Syria. This 64,845-word corpus is available in four formats: (1) transcriptions, (2) lemmatized transcriptions, (3) audio files and time-aligned phonetic transcriptions, and (4) an SQLite database. The transcription files are a digitized and corrected version of authentic transcriptions of tape-recorded narratives coming from a fieldwork trip conducted in the 1980s and published in the early 1990s (Arnold, 1991a, 1991b). They contain no annotation, except for some informative tagging (e.g. to mark loanwords and misspoken words). In the lemmatized version of the files, each word form is followed by its lemma in angled brackets. The time-aligned TextGrid annotations consist of four tiers: the sentence level (Tier 1), the word level (Tiers 2 and 3), and the segment level (Tier 4). These TextGrid files are downloadable together with their audio files (for the original source of the audio data see Arnold, 2003). The SQLite database enables users to access the data on the level of tokens, types, lemmas, sentences, narratives, or speakers. The corpus is now available to the scientific community at https://doi.org/10.5281/zenodo.6496714.

Keywords: language documentation corpus, lemmatization, time alignment

1. Introduction

Aramaic is a Semitic language that has been spoken in the Middle East for more than three millennia. It has survived, however, not as a single language but as a number of varieties collectively given the hypernym 'Neo-Aramaic'. These Neo-Aramaic varieties fall into four groups: Western Neo-Aramaic, Central Neo-Aramaic, North-Eastern Neo-Aramaic (NENA), and Neo-Mandaic (Heinrichs, 1990, pp. x–xv; Khan & Noorlander, 2021, p. xvii).

Western Neo-Aramaic is spoken in the three Syrian villages of Maaloula, Jubbaadin, and Bakhaa. In addition to the inhabitants of these villages, the native speakers who moved from these villages to bigger cities, such as Damascus and Beirut, still speak the language (Arnold, 2011, p. 685). Western Neo-Aramaic is now considered "definitely endangered" by the UNESCO Atlas of the World's Languages in Danger (Moseley, 2010). Similarly, Eberhard et al. (2021) report that the EGIDS level for Western Neo-Aramaic is 7 (Shifting).[1] Level 7 is exactly between 6b (Threatened) and 8a (Moribund). Since each of these three villages has its own dialect, and since this paper focuses only on the dialect spoken in Maaloula, we will use the term 'Maaloula Aramaic' to refer specifically to the Western Neo-Aramaic dialect of Maaloula.

Currently, neither a text corpus in electronic format nor a speech corpus with audio files and time-aligned transcriptions is available. This does not imply, however, that there is no well-documented written or audio material on Maaloula Aramaic. Transcriptions of authentic narratives coming from fieldwork trips have been published sporadically for more than a century (e.g. Bergsträsser, 1915, 1933; Reich, 1937; Spitaler, 1957; Arnold, 1991a, 1991b). An online archive of audio files, albeit without accompanying transcriptions, has existed for around 20 years (see section 2).

The importance of such transcriptions and audio archives to language documentation and preservation is undeniable, but the extent to which they can facilitate empirical linguistic research in their current format is rather limited. For example, a phonetician interested in the acoustic properties of the Maaloula Aramaic sounds will need to listen to the audio files and simultaneously go through the textbook pages to match the transcriptions with the pronounced segments. This is because these transcriptions are mainly available in paper format. By the same token, a morphologist studying a certain inflectional process will have to collect the examples manually from these textbooks.

The electronic corpus presented in this paper meets these and other empirical research requirements by benefitting from and complementing the existing resources. These resources are the result of many hours of work involving finding the native speakers, recording their speech in situ, and painstakingly transcribing the recordings. Therefore, turning part of them into a speech corpus is a more efficient process than having to repeat all these steps from the beginning.

However, compiling a corpus that would cover a wide array of potential research needs should go beyond the digitization of available transcriptions. Therefore, we decided to design a multi-purpose corpus and make it available to the scientific community in four different formats: (1) transcriptions (e.g. for lexical and sociolinguistic analysis), (2) lemmatized transcriptions (e.g. for morphological and lexicographical analysis), (3) audio files and time-aligned phonetic transcriptions (e.g. for phonetic and phonological analysis), and (4) an SQLite database through which the data can be accessed on the level of tokens, types, lemmas, sentences, narratives, or speakers, thus enabling all sorts of inquiries on any of these levels. Such formats are now considered state-of-the-art, as evidenced by the growing number of speech corpora which include time-aligned phonetic transcriptions, such as the TIMIT corpus (Garofolo et al., 1993), the Switchboard corpus (Godfrey et al., 1992; Godfrey & Holliman, 1993), and the Buckeye Corpus (Pitt et al., 2007).

In this paper, we will introduce the Maaloula Aramaic Speech Corpus (MASC, Eid et al., 2022). We will first

[1] EGIDS stands for the Expanded Graded Intergenerational Disruption Scale and is used to evaluate how endangered a language is.

describe the data we included in the corpus (Section 2). In Section 3, we will explain how the transcriptions were computerized and annotated. In Section 4, we will describe the corpus composition, structure, and use. Finally, in Section 5, we will discuss the current and potential applications of the corpus.

For the descriptive statistical procedures presented in this paper, we used R (R Core Team, 2021). The concordances were generated with the corpus analysis toolkit AntConc (Anthony, 2020). For displaying the spectrograms, waveforms, and TextGrid annotations, we used Praat (Boersma & Weenink, 2021).

Throughout the paper, we use the Roman numbers III and IV to refer respectively to Arnold's volumes (1991a) and (1991b). The Arabic numbers refer either to page numbers (e.g. III.28 refers to Arnold 1991a, p. 28) or to text file numbers if followed by *.txt* (e.g. III.28.txt refers to the 28th narrative in Arnold 1991a).

2. The Data to be Included in the Corpus

The data chosen for inclusion in the Maaloula Aramaic Speech Corpus consist of the transcriptions of tape-recorded narratives that Werner Arnold collected during his field research in Maaloula between 1985 and 1987. These transcriptions alongside the translation into German appear in two publications (Arnold 1991a, 1991b). These two particular sources were chosen for two main reasons.

First, the audio files of these narratives are available at the *Semitisches Tonarchiv* 'Semitic Sound Archive' website of Heidelberg University (see Arnold, 2003). They are fully accessible to the scientific community, as the Semitisches Tonarchiv "was established by support of the Deutsche Forschungsgemeinschaft and it can therefore be used by all scientists for research purposes" (Arnold, private communication). Each audio file is further supplemented by valuable metadata (e.g. name, gender, age, and occupation of the speaker; the year and place of recording; and reference to the textbook that contains the transcription).

Second, these texts are varied with regard to their content and the sociolinguistic variables pertaining to their narrators. In terms of content, these texts consist of 173 monologues that belong to different text types, such as fairy tales, fables, and legends; local and religious traditions, customs, and beliefs; personal experiences and autobiographies; daily, occupational, and agricultural activities; jokes and anecdotes; songs and poems (see Arnold, 1991a, pp. vii–x, 1991b, pp. vii–ix for a comprehensive classification of the individual narratives).

In terms of their sociolinguistic properties, these monologues are also varied as they were narrated by 45 native speakers (32 males, 13 females) between the ages of 13 and 89. There are no substantial differences between the age of female speakers (mean = 50.8 years) and male speakers (mean = 52.6 years) (see Arnold, 1991a, pp. 381–382, 1991b, pp. 345–346 for the name, age, and occupation of each speaker).

Now we turn to how we computerized and annotated these transcriptions.

3. Data Computerization and Annotation

This step involved carrying out the following tasks:

- scanning and digitizing the transcriptions
- correcting the errors manually and adding informative tags
- lemmatizing the transcriptions
- denoising the audio recordings
- automatically aligning the transcriptions with the corresponding recordings

In what follows, each task will be introduced and explained individually.

3.1 Scanning and Digitizing the Transcriptions

The two volumes (Arnold 1991a, 1991b) were scanned, and the transcriptions were computerized with the help of the optical character recognition (OCR) software ABBYY FineReader 10.[2] However, since Maaloula Aramaic is not one of the languages that the OCR software can recognize, the computerized text was far from perfect, as example (1) shows:

(1) **OCR output:** *anah höxa b^{d}-blöta nmiScabrill Sinbö mastra ra?isô P-blöta*
Desired text: *anaḥ hōxa bə-blōta nmiʕčabrill ʕinbō maṣtra raʔisō lə-blōta*
'We, here in the village, consider grapes to be a main source for the village.' III.28

While some errors were predictable and somehow automatically correctable (e.g. *S*, *c*, and *ö ~ ô* could be replaced with *ʕ*, *č*, and *ō* respectively), other errors were impossible to correct automatically. For example, the contrast between similarly written characters (e.g. *š* and *ṣ*, *ḳ* and *k*, *ḥ* and *h*) was neutralized completely by the OCR software, which displayed all these characters without the diacritic marks (e.g. *ana**h*** rather than *ana**ḥ*** 'we' in (1)). As a result, manual correction was inevitable.

3.2 Correcting Errors and Adding Informative Tags

In order to produce an error-free text, we hired a native speaker consultant who compared the scanned texts with both the original transcriptions and audio files. During this phase, two types of errors, spelling inconsistencies, and mismatches were identified and corrected. The first type consists of spelling errors and inconsistencies in the original transcription, such as the words in (2). The errors, here, were not made by the original narrators. They are the result of the transcription process itself. Therefore, we corrected them without adding any textual marking.

(2)

Misspelled	Corrected		
sōləfṯa	*soləfṯa*	'story'	IV.140
bēʕṯa	*beʕṯa*	'egg'	III.326
ḳuttōra	*ḳuṭṭōra*	'quarrel; fight'	IV.8
m-ʕa	*maʕ*	'from; about'	IV.8
ḳʕōle ~ kʕōle	*ḳʕōle*	'he sat'	III.304 ~ IV.8

The second type consists of errors made by the narrators themselves. In these cases, we tried to remain as faithful as possible to the audio files even if this meant that some of our new passages would be different from the original

[2] We were granted permission to use the published transcriptions by the Harrassowitz Publishing House.

transcriptions. For this type, we added explicit textual marking. Whenever a narrator made an error, we would transcribe their words the way they were said, but we would mark the error by inserting *sic* in square brackets immediately after it and give our consultant's suggested correction in parentheses without changing the narrators' actual words, as shown in (3). In this example, the narrator inadvertently made a subject-verb agreement error. He used the verb *ṯōle* which is inflected for the third person masculine singular although it is followed by the feminine subject *eḥḏa*.

(3) *ṯōle* [sic] *(= ṯalla) eḥḏa*
ṯō-l-e *(=ṯal-l-a)* *eḥḏ-a*
come.PRET-OM-3M.SG come.PRET-OM-3F.SG one-F
'Someone (F) came.' III.132

In the original transcriptions, only one form appears (usually the corrected one).
The second type also includes false starts, self-corrections, and extraneous remarks. Whenever a narrator reformulated their words after a false start or some hesitation, both forms would be kept, but the false start would be followed by points of ellipsis, as example (4) shows. This practice was already adopted in the original transcriptions, but we extended it to cover all similar cases.

(4) *battax... battaḥ nibəx baḥar, lōb ṭaššrīčnaḥ*
batt-ax *batt-aḥ* *ni-bəx* *baḥar* *lōb*
will-2M.SG will-1PL 1-cry.SBJV a lot if
ṭaššr-īč-n-aḥ
leave.PRET-2M.SG-LM-1PL
'You (M.SG) will… We will cry a lot if you (M.SG) leave us.' IV.116

If a word is interrupted, it is marked with two consecutive hyphens (--) (e.g. *amrō-- amrōle* 'she said to him' IV.14). We chose a different symbol for interrupted words to distinguish them from false starts, self-corrections, and extraneous remarks. This is because the interrupted words are always ungrammatical as they are cut off before reaching their end (e.g. **amrō*). They are not part of the lexicon of the language. However, the words followed by points of ellipsis are meaningful and grammatical on their own (e.g. *battax* 'you (M.SG) will' in (4)) but they are either redundant or in disagreement with the following syntactic units.
We kept the punctuation marks and numbering of the individual sentences as they appear in the original text. We also kept the original loanword annotation which marks the non-aramaicized, infrequently occurring Arabic loanwords (Arnold, 1991a, p. 24). We only changed the symbols used in this annotation from the original superscript *A* letters, as in (5a), to the tags <ar> and </ar>, as in (5b).

(5) a. Original text: ^{A}*fa*^{A} *bess yiṯḳan aylul*
b. Corpus text: <ar> *fa* </ar> *bess yiṯḳan aylul*
'When September comes.' III.28

3.3 Lemmatizing the Transcriptions

Lemmatization is a type of corpus tagging whereby the inflected word forms are linked to their lemmas. Lemmatization is a handy feature for many research tasks, and is particularly useful for highly inflectional languages (McEnery et al., 2006, pp. 36–36). Being a Semitic language with complex root-and-pattern morphology, Maaloula Aramaic is such a language. This is illustrated in (6).

(6)

Lemma			Word form		
ḏōḏa	'uncle'	III.220	*ḏaḏōye*	'his uncles'	III.256
ḏōrča	'house'	IV.138	*ḏaryōṯa*	'houses'	IV.68

We decided to lemmatize the transcriptions to maximize the benefit of this corpus.
Since there were no electronic resources available for Aramaic that would have allowed automatic lemmatization, we did this manually, implementing the following procedure.
As a first step, and with the help of our native speaker consultant, we created a word list, which consisted of all of the 12,220 unique word forms, and supplied each word form with its lemma and root as they appear in Arnold's (2019) Aramaic-German dictionary. We excluded 614 forms because they were interrupted or misspoken words, individual letters, Arabic loanwords, or proper nouns. Although we kept these word forms in the list, we provided them with tags rather than lemmas, such as [interrupted], [sic], [NA], [loanword], and [proper noun].[3] The resulting lemma list (exemplified in Table 1) consists of 3,781 different lemmas derived from 1,932 roots.

Root	Lemma	Word form
zbn	*zappen yzappen*	*mzappnin*
zbn	*zappen yzappen*	*nimzappella*
zbn	*zappen yzappen*	*nimzappen*
zbn	*zappen yzappen*	*nimzappnilla*
zbn	*zappen yzappen*	*nimzappnille*
zbn	*zappen yzappen*	*nzappīlle*
zbn	*zappen yzappen*	*nzappillēle*
zbn	*zappen yzappen*	*nzappnell*

Table 1: Extract from the lemma list

Based on the hand-crafted list of form-lemma mappings, the transcription files were enhanced to indicate the lemma for each word form. Lemmas were added in angled brackets immediately after the word form, making this version of the corpus easy to use with AntConc (see Section 4.2.2 for the advantages of this format).

3.4 Denoising the Audio Recordings

Since the original audio files were tape-recorded several decades ago, some amount of noise was present in the data. We used the REAPER Digital Audio Workstation software (https://reaper.fm) with the ReaFIR plugin to create a noise profile for the audio files and to generate a denoised version of each file.

3.5 Automatically Aligning the Transcriptions with the Recordings

One of the goals of this work was the creation of Praat TextGrid files in which the audio files are aligned with their transcriptions. Since Maaloula Aramaic is a relatively small and underdocumented language, no pre-trained language-specific alignment tool is available for it. We

[3] We noticed later that we could exclude more Arabic loanwords and proper nouns, but we did not proceed because classifying a word as aramaicized or not did not prove straightforward.

used the WebMAUS tool (Schiel, 1999, 2015) provided by BAS Web Services (Kisler et al., 2017, available at https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface) to align the denoised audio files with the transcription files. WebMAUS provides a language-agnostic model which can align speech signals with phonetic transcriptions represented in SAMPA format. We created a mapping of the characters appearing in our corrected transcription files to their corresponding SAMPA characters and used a SAMPA-encoded version of our text files as input to WebMAUS, together with the denoised audio files. Denoising the audio files prior to processing led to significantly better results with regard to alignment quality. For instance, noisy periods in the original audio files were often analyzed as long fricatives by WebMAUS, while the denoised files allowed WebMAUS to more reliably recognize pauses. The TextGrid files were then extended by a sentence tier, in addition to the word- and phoneme-level tiers provided by the WebMAUS output.

4. Corpus Composition and Structure

In this section, we describe the composition of the corpus. We present statistics on the word tokens that make up the corpus (i.e. the number of word tokens per file, per speaker, per gender, and per age group). We also describe the different formats in which the corpus is available, where to find the corpus, and how to use it.

4.1 Corpus Composition

Following Arnold's original organization of texts and audio files, we divided the transcriptions into 173 text files, which contain 64,845 tokens in total, and saved them in UTF-8 format.[4] The speech data vary considerably in the number of tokens per file (*M* = 374.8, *Mdn* = 227, *Min* = 19, *Max* = 4,340, *SD* = 470) and in the number of tokens per speaker (*M* = 1,441, *Mdn* = 754, *Min* = 42, *Max* = 10,688, *SD* = 2,232.9). As can be seen from Figure 1, four speakers (represented by the leftmost bars) provided many more tokens than any of the other speakers. They produced 31,988 tokens, making up 49.3% of the entire corpus, whereas all the other 41 speakers produced a total of 32,857 tokens (50.7%).

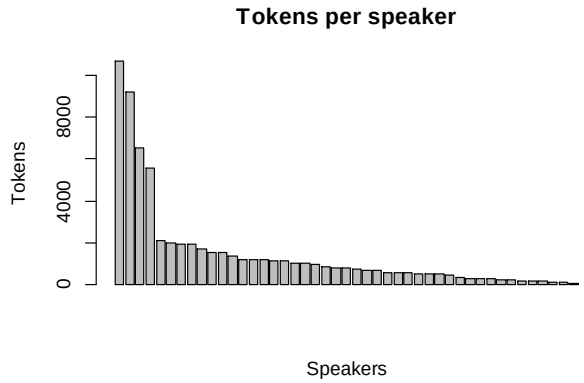


Figure 1: Distribution of tokens by speaker

Around 63% of the produced speech data come from older speakers (aged 50-79) (see Figure 2). This trend is more prominent for female speakers where 86% of the tokens come from these age groups. Although the same trend is noticeable for male speakers, the 10,688 tokens produced by only one 26-year-old speaker (represented by the leftmost bar in Figure 1 above) have partly masked this trend by giving more weight to age group 20-29.

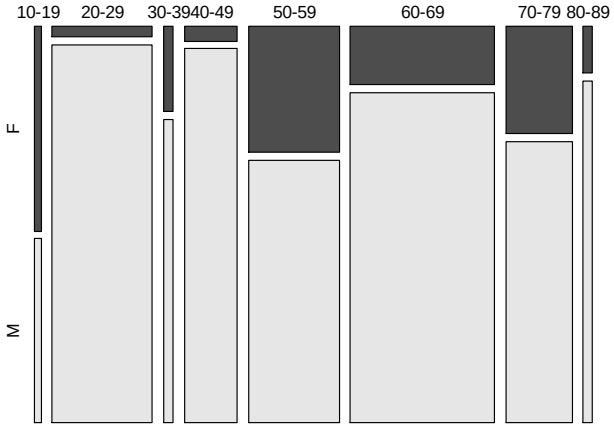


Figure 2: Distribution of tokens by age and gender

Figure 2 also clearly shows that the corpus contains more words spoken by male speakers (53,922 tokens, 83.2%) than by female speakers (10,923 tokens, 16.8%). This distribution is expected given that the male speakers outnumber the female speakers (see Section 2) and the four main speakers are all men.

4.2 Corpus Structure and Use

The corpus is available to the research community at https://doi.org/10.5281/zenodo.6496714. As already mentioned above, the corpus data come in four different formats: (1) transcriptions, (2) lemmatized transcriptions, (3) audio files and time-aligned phonetic transcriptions, and (4) an SQLite database.

4.2.1 The Transcriptions

These text files are the digitized transcriptions that contain no annotation at all (except for the informative tags presented in Section 3.2, e.g. [sic], <ar>, </ar>). These plain transcription files (as well as the lemmatized transcription files presented in Section 4.2.2) can be used with any regular programming language, such as Python. Researchers not familiar with programming can access and analyze these files via a corpus analysis toolkit. We chose to set up the files in a format compatible with the corpus tool AntConc (Anthony, 2020) because it is user-friendly, free, and available to Windows, Macintosh OS X, and Linux users.

Using the corpus analysis toolkit, researchers can investigate the unannotated corpus by carrying out basic tasks, such as generating frequency lists, examining concordances, and analyzing collocations and keywords. For example, Table 2 shows part of the key word in context (KWIC) display for *xōla* 'food' within a window of two words to the left and right.

[4] Corpus users will notice, however, that the corpus consists of 65,722 tokens, which additionally include the informative tags, *sic* and *ar*, and corrected words in parentheses.

KWIC			File
xett mišwin	xōla	alūla aw	III.55.txt
čiḳʕēx billa	xōla	?” (15) amelle	IV.23.txt
“ē, šwēn	xōla	atar, baḥ	III.23.txt
htīṯa aw	xōla	aw ščū	III.52.txt
ʕammaxell lanna	xōla	.” (28) aḳa bib-	IV.07.txt
ʕammaxlōl lanna	xōla	. (29) aḳa ḥakīna	IV.07.txt
ḏikkil iṭlab	xōla	, aḳam hann	IV.33.txt

Table 2: KWIC concordance of *xōla* ‘food’

Using wild cards, such as the asterisk, researchers can conduct basic morphological analyses. For example, to generate a list of the words that contain the root *ṭʕn*, the search string **ṭ*ʕ*n** can be used. Table 3 shows only seven out of the 197 concordance hits that this search finds in the corpus.

KWIC			File
čimbaṭṭel	čiṭʕun	ḥ-ḥaṣṣax?” (6)	IV.22.txt
amrillax lā	čiṭʕun	ḥ-ḥaṣṣax?” (11)	IV.22.txt
w čzellax	čiṭʕun	ḥ-ḥaṣṣax?” (19)	IV.22.txt
batta	čšaṭiʕenne	šaṭranž. (19)	IV.15.txt
lā baḳḳrič	čṭuʕnenne	w čišwenne	IV.34.txt
w hanna	ʕamṭōʕen	ḥ-ḥaṣṣe, bann	IV.22.txt
mazal čū	ʕamṭōʕna	ḳuṭʕā w	IV.55.txt

Table 3: KWIC concordance of words containing the search string **ṭ*ʕ*n**

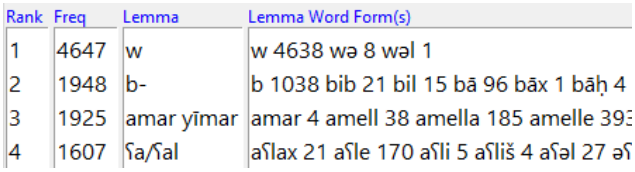
However, raw data like these may contain many irrelevant words. For example, although the fourth word, *čšaṭiʕenne* ‘play (SBJV 3F SG) [e.g. chess] with him’, contains the search string **ṭ*ʕ*n**, it should be weeded out manually because its root is *šṭʕ* rather than *ṭʕn* (see Arnold 2019, p. 761).

The lemma list we provide as part of our corpus is a more elegant and timesaving solution to the problem of having to find and remove the irrelevant results manually. This solution enables the corpus users to investigate the lemmas as well as all their inflectional variants by uploading a lemma list to the corpus tool. For the lemma list (presented in Section 3.3) to be processed by AntConc, its layout was modified slightly. Example (7) shows the modified layout of the lemma list whereby the lemma is separated from its word form(s) by an arrow (->).

(7) The AntConc-friendly lemma list layout

```
ḥaẓẓūra -> ḥaẓẓūr, ḥaẓẓūra, ḥaẓẓurō
ḥbōka -> ḥbōka
hbulya -> hbulya
ḥḏawṯa -> əḥḏawōṯa, ḥḏawōṯa
ḥḏučča -> əḥḏuččaḥ, ḥḏučča, ḥḏučče,
ḥḏuččōṯa, ḥḏuččun
ḥḏūṯa -> əḥḏūṯa, ḥḏūṯ, ḥḏūṯa, ḥḏuṭō
```

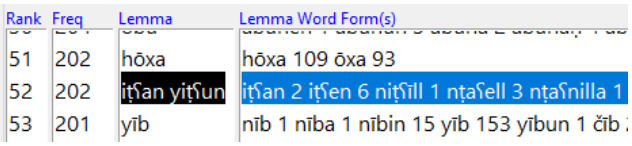
For the corpus users to load the lemma list to AntConc, they need to upload the Maaloula Aramaic Speech Corpus first, and then choose the Word List category in the Tool Preferences tab and click on the Lemma List Load button. When a word list is created, the lemma (rather than the word form) and its frequency are given first, followed by the individual word forms and their frequencies, as in Figure 3.



Rank	Freq	Lemma	Lemma Word Form(s)
1	4647	w	w 4638 wə 8 wəl 1
2	1948	b-	b 1038 bib 21 bil 15 bā 96 bāx 1 bāḥ 4
3	1925	amar yīmar	amar 4 amell 38 amella 185 amelle 39
4	1607	ʕa/ʕal	aʕlax 21 aʕle 170 aʕli 5 aʕliš 4 aʕəl 27 əʕ

Figure 3: Screenshot from AntConc: A lemma frequency list

Using the same search string (i.e. **ṭ*ʕ*n**) in the Word List pane and the numbers in the Search Only box, we can examine the lemmas that contain the root *ṭʕn*. The search yields only six results this time, three of which contain the root *ṭʕn* and three are irrelevant. Figure 4 illustrates one of these six lemmas (highlighted).



Rank	Freq	Lemma	Lemma Word Form(s)
51	202	hōxa	hōxa 109 ōxa 93
52	202	iṭʕan yiṭʕun	iṭʕan 2 iṭʕen 6 niṭʕill 1 nṭaʕell 3 nṭaʕnilla 1
53	201	yīb	nīb 1 nība 1 nībin 15 yīb 153 yībun 1 čīb

Figure 4: Screenshot from AntConc: A lemma containing the search string **ṭ*ʕ*n** and its word forms

It can be seen that all the inflectional forms of this lemma which the corpus contains are listed together with their frequencies to the right of the lemma.

For the corpus users who want to conduct further analyses and, therefore, need the output to be organized in a dataframe with each variable receiving a column, we provide a spreadsheet for this purpose. The spreadsheet is called “MASC_dataframe.csv” and is downloadable with the corpus. It contains all the 12,220 unique word forms, their frequencies, their lemmas, the frequencies of their lemmas, and their roots. Table 4 shows the first few rows of the spreadsheet.

Root	Lemma	LemmaFreq	Word_form	Word_formFreq
w	w	4647	w	4638
w	w	4647	wə	8
w	w	4647	wəl	1
b	b-	1948	b	1038
b	b-	1948	bā	96
b	b-	1948	bāḥ	4
b	b-	1948	bāx	1

Table 4: Extract from the MASC dataframe

4.2.2 The Lemmatized Transcriptions

In these files, each word is followed by the citation form of its lemma in angled brackets, as in (8). These files are the result of the lemmatization process introduced in Section 3.3.

(8) Two lemmatized sentences from file III.01.txt

```
(2) anaḥ<anaḥ> hōxa<hōxa> bə<b->-blōta<blōta>
nmiʕčabrill<iʕčbar yiʕčbar> ʕinbō<ʕenəpṭa>
maṣtra<maṣtra> raʔisō<raʔisa> lə<l>-
blōta<blōta>. (3) <ar<[annotation]>> fa<fa>
</ar<[annotation]>> bess<bess/bessi>
yiṭḳan<iṭḳen yiṭḳan> aylul<aylun/aylul>
yiščawyan<iščwi yiščwi> ʕinbō<ʕenəpṭa>
ʕa<ʕa/ʕal> maẓbuṭ<maẓbuṭ>, tōr<tōr>
batte<batt-> yizlullun<zalle yzelle>
ʕa<ʕa/ʕal> šṭōḥa<šṭōḥa>.
```

Researchers can use this lemmatized corpus in different ways, using a corpus analysis toolkit. For example, they

can search for the lemma itself, as in Table 5. In this example, the search for the lemma *iṯḳen yiṯḳan* 'to become' (a lemma chosen from example (8) above) yields 476 hits, seven of which are shown in the table.

KWIC	File
iṯḳen<iṯḳen yiṯḳan >. amelle<amar yīmar>	III.32.txt
yiṯḳan<iṯḳen yiṯḳan >." (18) amellon<amar	IV.20.txt
ṯōḳna<iṯḳen yiṯḳan >." (13) amrōle<amar	IV.56.txt
tiḳniṯ<iṯḳen yiṯḳan > ana<ana> nnōḥeč<inḥeč	III.53.txt
tiḳniṯ<iṯḳen yiṯḳan > ana<ana> yaṯma<yaṯma>	III.99.txt
ṯōḳen<iṯḳen yiṯḳan >, ana<ana> mn<m-/mn->	IV.15.txt
tiḳninnaḥ<iṯḳen yiṯḳan > ana<ana> w<w>	IV.58.txt

Table 5: KWIC concordance of the lemma *iṯḳen yiṯḳan* 'to become'

If a researcher is not sure what the exact lemma is, they can look it up by searching for any of its word forms.

AntConc provides the option of hiding these tags completely or partially (from the Tags category in the Global Settings tab). If the option Hide Tags is chosen, the tags will be hidden completely, and the files will appear in their plain form (as in Section 4.2.1). However, if the option Hide Tags (Search in Conc/Plot/File View) is chosen and the lemma is typed explicitly in the search window with the surrounding brackets and a preceding asterisk (e.g. *<iṯḳen yiṯḳan>), then the lemma itself will not be revealed, but the relevant word forms will be marked.

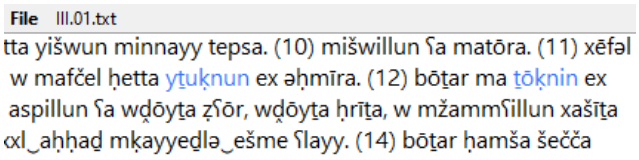


Figure 5: Screenshot from the File View window in AntConc (hidden lemma tags)

Figure 5 is a screenshot from the File View window in AntConc. All tags, including the searched lemma *iṯḳen yiṯḳan* 'to become', are hidden, but the relevant word forms *yṯuḳnun* 'become (SBJV 3M PL)' and *ṯōḳnin* 'become (PRS 3M PL)' are marked in blue.

4.2.3 The Audio Files and Time-aligned Phonetic Transcriptions

The audio files are included in our corpus in the form of 176 mp3 files (10 hours of audio material).[5] Both the original and denoised audio files are available and can be opened in Praat (Boersma & Weenink, 2021) together with their corresponding TextGrid files to conduct different types of acoustic analyses, such as measuring segment duration, vowel formants, and pitch.

The TextGrid annotations consist of four tiers, as shown in Figure 6. The first tier represents the sentence level. The second and third tiers represent the word level in the normal script (Tier 2) and SAMPA (Tier 3). The fourth tier represents the segment level, which is also transcribed in SAMPA.

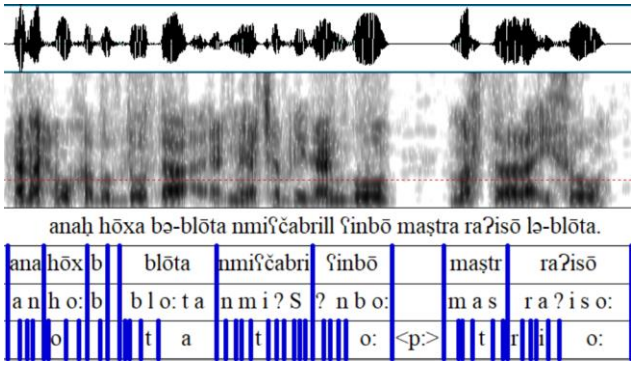


Figure 6: Screenshot from Praat displaying the four tiers as well as the corresponding spectrogram and waveform

4.2.4 The SQLite Database

The SQLite database consists of 8 interconnected tables in which the tokens, types, lemmas, sentences, narratives, speakers, audio files and transcription files that appear in the corpus are associated with each other. The structure of the database is visualized in Figure 7.

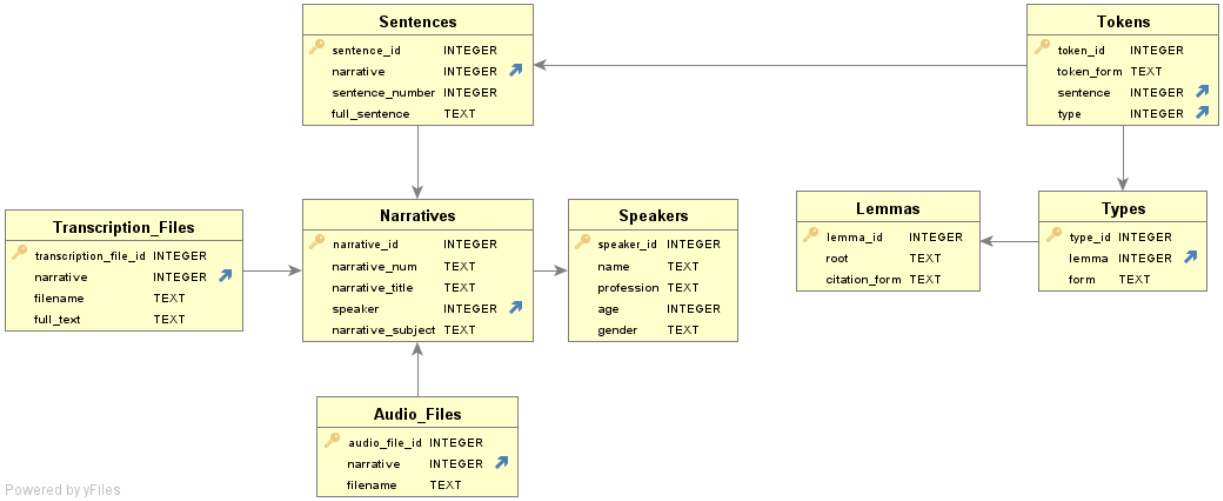


Figure 7: The structure of the database

[5] During the time alignment process, we had to divide a 44-minute audio file into four pieces. This explains why we have 176 (rather than 173) mp3 files.

This provides a way to conduct statistical analyses that optionally take metadata into account. For instance, the database can be queried to answer questions such as:

- Which words are most often used by female speakers, and which words are most often used by male speakers?
- Which words are specific to one subject area, and which words appear in the context of a variety of topics?
- Which words are exclusively used by speakers belonging to a particular profession?
- Do younger speakers produce longer sentences than older speakers, or shorter sentences?

An example query selecting all sentences uttered by female speakers under 40 is presented in Figure 8.

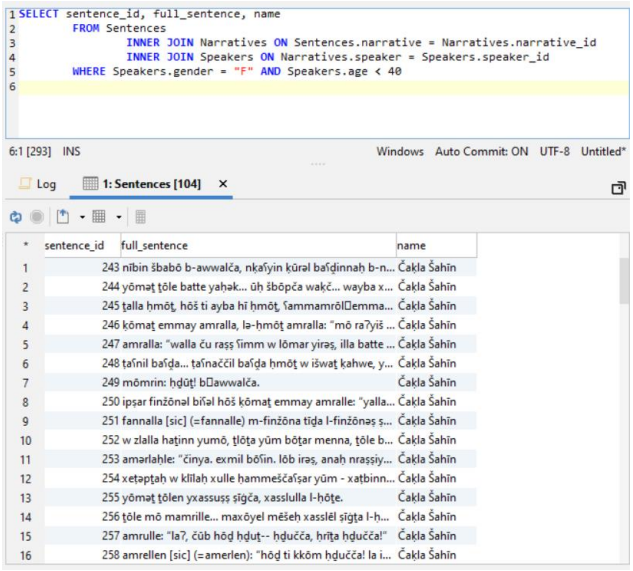


Figure 8: Example query on the MASC database

5. Discussion: Applications

As previously noted, one of the main goals of creating the Maaloula Aramaic Speech Corpus is to facilitate empirical linguistic research. This goal has been put to the test in one study, conducted by two of the co-authors of this article, which investigated vowel epenthesis in Maaloula Aramaic from a syllable-based perspective (Eid & Plag, 2021). As can be seen in Figure 9, the corpus was an essential component of the research process adopted in this study. It was used to generate the words that exemplify the descriptive generalization found in previous accounts on Maaloula Aramaic as well as the words that represent counterexamples not captured by the generalization. The numerous examples and counterexamples provided by the corpus helped us reformulate and formalize the generalization.

In a different study employing acoustic analysis, the TextGrid files will be used to measure vowel formants and to compare the durations of singletons with those of geminates (Eid, in prep.).

Further studies based on the Maaloula Aramaic Speech Corpus are possible in the future. For example, since the corpus provides authentic speech production data, it may be useful for studies of speech production that want to test the effect of word frequency or morphological processes (e.g. affixation) on phonetic implementation in a language that has never been explored from this perspective.

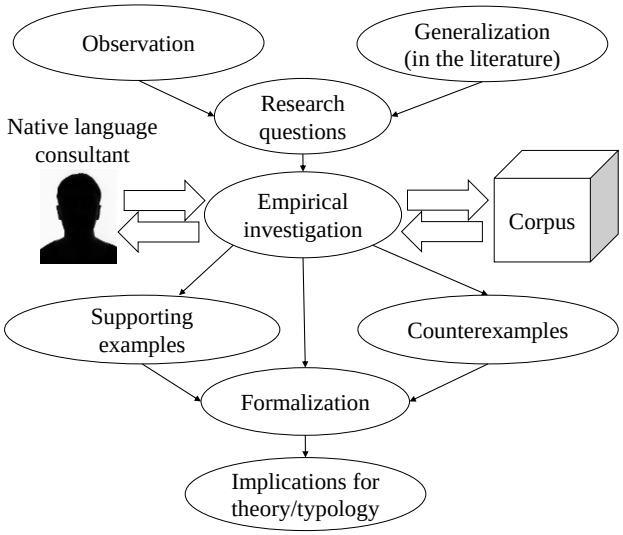


Figure 9: The research process adopted in Eid & Plag (2021)

6. Acknowledgements

We would like to thank our colleagues at Heinrich Heine University Düsseldorf and the reviewers for their helpful feedback. We are also grateful to our native language consultant, Emad Rihan, for matching the scanned texts with the original transcriptions and audio files, correcting the errors, and helping us create the lemma list. We would like to thank Harrassowitz Verlag for the permission to include the published transcriptions in the Maaloula Aramaic Speech Corpus, and Werner Arnold for allowing us to use the audio data. We thank Roman Seyffarth for support in denoising the audio files.

7. Glossing Abbreviations

1	first person
2	second person
3	third person
F	feminine
LM	linking morpheme
M	masculine
OM	object marking
PL	plural
PRET	preterit
SBJV	subjunctive
SG	singular

8. Bibliographical References

Anthony, L. (2020). *AntConc [Computer program]* (Version 3.5.9) [Computer software]. Version 3.5.9, retrieved from https://www.laurenceanthony.net/software

Arnold, W. (1991a). *Das Neuwestaramäische: III. Volkskundliche Texte aus Maʿlūla.* Otto Harrassowitz.

Arnold, W. (1991b). *Das Neuwestaramäische: IV. Orale Literatur aus Maʿlūla.* Otto Harrassowitz.

Arnold, W. (2011). Western Neo-Aramaic. In S. Weninger, G. Khan, M. P. Streck, & J. C. E. Watson (Eds.), *The Semitic languages. An international handbook* (pp. 685–696). De Gruyter Mouton.

Arnold, W. (2019). *Das Neuwestaramäische: VI. Wörterbuch*. Harrassowitz.

Bergsträsser, G. (1915). *Neuaramäische Märchen und andere Texte aus Maʿlūla*. F.A. Brockhaus.

Bergsträsser, G. (1933). *Phonogramme im neuaramäischen Dialekt von Malula. Satzdruck und Satzmelodie.* Verlag der Bayerischen Akademie der Wissenschaften.

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/

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2021). *Ethnologue: Languages of the world* (24th ed.). SIL International. Online version: http://www.ethnologue.com

Eid, G. (in prep.). *The phonology of Maaloula Aramaic* [Doctoral dissertation]. Heinrich-Heine-Universität Düsseldorf.

Eid, G., & Plag, I. (2021). Syllable structure and syllabification in Maaloula Aramaic. *Submitted*.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Linguistic Data Consortium.

Godfrey, J. J., & Holliman, E. C. (1993). *Switchboard-1 Release 2 LDC97S62*. Linguistic Data Consortium.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)* (pp. 517–520). San Francisco, CA.

Heinrichs, W. (1990). Introduction. In W. Heinrichs (Ed.), *Studies in Neo-Aramaic* (pp. ix–xvii). Scholars Press.

Khan, G., & Noorlander, P. M. (Eds.). (2021). *Studies in the grammar and lexicon of Neo-Aramaic*. Open Book Publishers. https://doi.org/10.11647/OBP.0209

Kisler, T., Reichel, U. D., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*(September 2017), 326–347.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Routledge.

Moseley, C. (Ed.). (2010). *Atlas of the World's Languages in Danger*. (3rd ed.). UNESCO Publishing. Online version: http://www.unesco.org/languages-atlas/

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). *Buckeye Corpus of Conversational Speech (2nd release)*. Department of Psychology, Ohio State University (Distributor). [www.buckeyecorpus.osu.edu]

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/.

Reich, S. (1937). *Études sur les villages araméens de l'Anti-Liban*. Institut Français de Damas.

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.

Spitaler, A. (1957). Neue Materialien zum aramäischen Dialekt von Maʿlūla. *Zeitschrift Der Deutschen Morgenländischen Gesellschaft*, *107*(2), 299–339.

9. 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