The Speaker and Language Recognition Workshop (Odyssey 2024)
18-21 June 2024, Quebec City, Canada

MAGLIC: THE MAGHREBI LANGUAGE IDENTIFICATION CORPUS

Karen Jones, Kevin Walker, Christopher Caruso, Stephanie Strassel

Linguistic Data Consortium, University of Pennsylvania

karj@ldc.upenn.edu, walkerk@ldc.upenn.edu, carusocr@ldc.upenn.edu, strassel@ldc.upenn.edu

ABSTRACT

The Maghrebi Language Identification Corpus (MAGLIC) is a new resource for language recognition comprising over 600 conversational telephone speech recordings in four North African languages: three Maghrebi Arabic dialects (Algerian, Libyan and Tunisian) plus North African French. Calls from consented subjects were recorded on a custom telephony platform in Tunisia, and call metadata was collected including recording time, anonymized phone number and speaker sex. Segments from collected calls were selected for use as development and test data in the NIST 2022 Language Recognition Evaluation, where results indicate that systems show confusability among the three Maghrebi Arabic dialects, despite an inter-annotator agreement study indicating that humans can distinguish these varieties with high reliability. The MAGLIC corpus will be published in the LDC Catalog, making it broadly available for language recognition and other language-related research, education and technology development.

1. INTRODUCTION

The Maghrebi Language Identification Corpus (MAGLIC) is a new resource for language recognition covering four linguistic varieties commonly spoken in North Africa: three Maghrebi Arabic dialects (Algerian, Libyan and Tunisian Arabic) along with North African French. Native speakers of each variety were recruited as subjects to make telephone calls to friends and family members. Calls were recorded on a custom collection platform located in Tunis, Tunisia. Along with both call sides, call metadata was collected including the date and time of recording, anonymized phone numbers, and self-reported speaker sex for the primary call side. Each call lasted at least 8 minutes, and speakers discussed any topic of their choosing using a single language variety for the duration of the call. A total of 631 calls were collected, roughly equally divided across the four varieties. Each recording was verified for duration and overall recording quality, and native speakers verified language identity for each call.

Portions of the collected recordings were selected for use as development and test data in the NIST 2022 Language Recognition Evaluation [1], whose goal is to advance language identification capabilities, with a special focus in 2022 on low-resource languages spoken in Africa. LRE22 results show relatively high confusability among the Maghrebi Arabic dialects, with lower confusion for North African French. In contrast, inter-annotator agreement studies conducted on the MAGLIC corpus, where auditors judged calls from which NIST extracted their test segments, show that native speakers can distinguish all four linguistic varieties with near-perfect reliability. Given this possible gap between human and system capabilities, the MAGLIC corpus represents an important new data source for language recognition and dialect identification research.

1.1. Languages

The four linguistic varieties included in the MAGLIC corpus are spoken in a region of North Africa traditionally known as the Maghreb. This region is linguistically complex for a variety of interconnected reasons. Diglossia is the norm, in which two varieties of Arabic are used by the same speakers to communicate in different situations. Modern Standard Arabic (MSA) is spoken in formal settings such as official communications and it is the predominant written variety, but it is not typically used by local people to communicate with one another in daily life ; instead, colloquial Maghrebi dialects are used for everyday interactions. In addition to the Arabic diglossic situation, another complicating factor is the prevalence of French, which is widely spoken in Tunisia, Algeria and Morocco due to the region's colonial history. Speakers in the region are typically multilingual, with a high degree of fluency in more than one variety, and codeswitching is very common among MSA, dialectal Arabic and French. Finally, borrowings are frequent in colloquial Maghrebi Arabic dialects, with loan words coming from a variety of languages such as MSA, Berber, English, Spanish, Italian and Turkish [2].

The MAGLIC corpus design required that collected speech be natural and conversational, so MSA was not a focus of collection. Instead, collection targeted the varieties spoken in everyday life: Algerian, Libyan and Tunisian colloquial Arabic, and North African French. While the Algerian, Libyan and Tunisian dialects are mutually intelligible to some degree (with the Tunisian and western Libyan dialects being most similar), they are nonetheless easily distinguishable by Arabic speakers in the region. Similarly, North African French spoken in the region is clearly distinct from both European and West African French.

While MAGLIC required natural conversational speech, it also required each recording to contain a single language. Given the prevalence of codeswitching and multilingualism in the speaker population, it was necessary to provide subjects with clear instructions about limiting their language use to a single variety in a given call. It was also important to audit the resulting recordings with this requirement in mind.

1.2. Related Work

The MAGLIC Corpus makes a significant contribution to the data landscape for Maghrebi language varieties, in particular Libyan Arabic and North African French, where the volume of resources is comparatively scarce.

Data for the NIST Language Recognition Evaluations of 2011, 2015 and 2017 included Maghrebi Arabic telephone recordings [3], [4], [5]. However, although this collection

10.21437/odyssey.2024-13

covered colloquial Arabic spoken in Mauritania, Western Sahara, Morocco, Algeria, Tunisia and Libya, all recordings were classified with a single "Maghrebi Arabic" label without any finer-grained dialect distinctions [6], [7], [8].

Other available multi-dialectal Arabic speech corpora are not specifically focused on Maghrebi dialects ; moreover they are typically designed for speech recognition rather than language and dialect recognition. These include the QASR: QCRI Aljazeera Speech Resource, comprising 2000 hours of broadcast speech that includes dialect labels [9], with a related effort to carefully label collected broadcasts for four major Arabic dialects including one designated as North African [10]. The Massive Arabic Speech Corpus (MASC) contains Arabic speech collected from YouTube and covers multiple Arabic dialects [11].

Other Arabic corpora have focused on single dialects, including some of the varieties present in the MAGLIC corpus. There are several Algerian Arabic corpora including the Algerian Modern Colloquial Arabic Speech Corpus (AMCASC) which includes phone/microphone recordings from 291 males and 44 females [12], the Algerian Arabic Speech Database which contains recordings of prompted utterances from 300 native Algerian Arabic speakers from various regions across the country [13], and KAL AM'DZ, a large speech corpus containing social media, YouTube, online radio and TV recordings in various Algerian dialects [14]. Tunisian Arabic corpora include LDC's Call My Net 2 conversational telephone speech corpus for speaker recognition research [6], which was used in the 2018 NIST Speaker Recognition Evaluation [15]. Other Tunisian data sets include the Spoken Tunisian Arabic Corpus [16], the Tunisian Arabic Railway Interactive Corpus [17], and the OrienTel MCA (Modern Colloquial Arabic) database [18]. Speech resources for Libyan Arabic and North African French in comparison are extremely scarce, and MAGLIC appears to be the only corpus focused specifically on the multiple inter-connected linguistic varieties spoken in the Maghreb.

1.3. MAGLIC in the LRE22 Evaluation

The MAGLIC corpus was one of three corpora developed for the 2022 NIST Language Recognition Evaluation (LRE22). The ongoing LRE evaluation series is designed to assess performance and promote advancements in automatic language recognition in conversational telephone speech (CTS) and broadcast narrowband speech (BNBS). The goal of the evaluation is to determine, for a closed set of languages, whether a system can reliably detect the presence of a target language in a set of test segments. LRE22 had a particular focus on languages spoken in Africa, with 14 linguistic varieties including low resource languages being evaluated. The four languages of the MAGLIC Corpus – Algerian Arabic, Tunisian Arabic, Libyan Arabic and North African French – comprised four of the 14 evaluation languages for LRE22. For each MAGLIC language, between 2400 and 2900 test segments were selected from the corpus, along with 300 segments for system development, with segment durations ranging from 3 to 35 seconds. Additional data was provided for the non-MAGLIC languages, and the evaluation featured both a fixed training condition in which systems could only utilize specified training resources, and an open condition in which there were no such constraints [19].

2. MAGLIC CORPUS DESIGN

2.1. Claque-Based Protocol

To build the MAGLIC corpus, we recruited native speaker subjects known as "claques" to make telephone calls to friends or family members known as "callees". Because all LRE22 test data was drawn from the callee call side, maximizing callee diversity was critical. Therefore, subject recruitment focused on enrolling claques with large social networks. Claques were prohibited from contacting the same callee more than once, such that each callee in the corpus both within and across languages is distinct. While claques themselves were often multilingual, each call made to a callee focused on a single language variety. Claques were responsible for ensuring that:

- Callees were native speakers of the target language variety;
- Callees understood the instruction to speak a single language and to avoid codeswitching for the duration of the conversation;
- Callees did most of the talking during the call;
- Each callee was called by the claque only once; and
- Across all callees for this claque, there was demographic variety in terms of age, gender, sex, education and region.

While claques provided basic demographic information upon enrollment, callees participated anonymously and no demographic information was recorded, though manual auditing of each recording did include a judgment about callee sex. Both callees and claques provided informed consent prior to calls being recorded, and all collection took place under Institutional Review Board oversight. Claques were compensated for each call that satisfied the requirements, and callees received no compensation.

2.2. Corpus Requirements

The MAGLIC corpus design was primarily informed by its intended use as a source of development and test data for LRE22.

All speakers, both callees and claques, were required to be native speakers of the target language (or highly fluent in the case of North African French), and each call was required to consist of speech in a single variety (Algerian Arabic, Libyan Arabic, Tunisian Arabic or North African French), without any dialect mixture, codeswitching or presence of any Modern Standard Arabic speech.

The collection goal was a minimum of 125 calls per language, with each call lasting 8-10 minutes, with at least 3 minutes of speech on the callee call side. Speakers were encouraged to conduct calls in a variety of physical settings, including both noisy and non-noisy environments, but were not strictly required to do so. To enforce the requirement that each call involve a unique speaker pairing, claques were prohibited from calling the same phone number more than 3 times, but a stricter prohibition was infeasible since claques sometimes called multiple members of the same household who shared a phone number.

3. COLLECTION PROCEDURE

3.1. Recording Platform

The MAGLIC collection platform was designed by LDC to support recording of telephony data, and is located in Tunis, Tunisia. It was designed to enable remote access so that collection monitoring, testing and troubleshooting can be performed from a centralized collection control site, while the physical platform remains in Tunisia. The platform's major functions are to provide pre-recorded instructions to enrolled claques and callees as they connect to the system, to obtain consent to be recorded from both parties, to record conversations for a pre-determined duration, and to log call metadata.

Platform hardware includes a control computer that handles the interactive voice recordings and all recording functions. The platform connects to cellular and traditional telephony networks via a GSM gateway and ISDN gateway respectively. Each gateway connects to the Public Switched Telephone Network (PSTN) via an external service provider, and to the platform itself via a SIP connection. Platform software includes Adhearsion (a framework for writing telephony applications in Ruby) and Asterisk which handles network connection configuration and call flow. A database server handles all participant information and call logging. Phone numbers for claques and callees are anonymized prior to being saved as metadata.

3.2. Call Flow

To initiate a call, claques dialed a number that was specific to the language variety designated for the current call. Upon connecting to the platform they listened to a set of recorded prompts for instructions. The prompts came in two versions: Arabic and French, and which version the claque heard depended on which platform number they dialed. Claques used the keypad on their phone to respond to the prompts, providing the following information:

- Their unique, persistent PIN obtained upon enrollment;
- The telephone number of the callee; and
- The number "1" to indicate their consent to have their voice recorded

The telephone platform then dialed out to the callee, who after an introductory message was asked to confirm they had not participated in other MAGLIC calls and to give consent to be recorded. The platform then bridged the call between both parties and recording began. Recording automatically ended at 10 minutes.

Recordings were saved immediately on the platform and then copied to a centralized file server. Each recording was automatically checked upon completion for overall duration and for amount of speech; these simple checks allowed problematic calls to be promptly reported to both project coordinators and to claques.

4. AUDITING

All collected calls were subject to manual auditing to confirm that the call was suitable for inclusion in the final corpus. Auditing focused exclusively on the callee call side, since this was the source of data for LRE22 evaluation.

4.1. Defining Audit Segments

Instead of auditing full calls, which would be cost-prohibitive, speech-dense segments from each callee call side were selected for auditing. The first 30 seconds of the recording was excluded from selection, since the start of the call typically contains minimal speech. For the remainder of the call, a Speech Activity Detector was used [20] to locate segments containing at least 30 seconds of speech within a 30-90 second sliding window. Two speech-dense, non-overlapping segments per call were selected and presented for manual auditing.

4.2. Auditing Approach

Auditing was performed by native (or highly fluent in the case of North African French) speakers of each variety. Auditors received formal training and testing and followed detailed guidelines. Any auditors who had also been MAGLIC claques were prevented from auditing their own calls.

Auditors used a custom web-based auditing user interface that presented audit segments for the specified language one at a time, along with a series of questions to be answered for each segment, as follows:

1. Is there speech throughout most of this segment?
2. How clear is the audio?
3. Is all of this speech in your language?
4. Is all of the speech from a single speaker?
5. What is the speaker's sex?
6. Is the speaker a native speaker of your language?
7. Is this a noisy call?

The user interface required auditors to listen to each segment in its entirety before answering any questions.

Given the prevalence of codeswitching in the Maghrebi region, the auditing guidelines provided detailed instructions on what constituted speech being "in your language". For instance, individual words borrowed from another language but commonly used in the target language were acceptable, while uncommon borrowings or longer stretches of speech in another variety were unacceptable.

4.3. Inter-Auditor Agreement

Auditing assignments for each language consisted of three kinds of segments. Most of the data comprised *target* segments, extracted from calls expected to be in the auditor's own language. Some of the data consisted of *dual* segments, also believed to be in the auditor's language but independently assigned to another auditor for this language, to measure inter-annotator agreement. A smaller number of segments were *distractors*, expected to be from a non-target language. The inclusion of distractor segments was primarily intended to keep the auditors alert and focused, and to identify any auditors who were not sufficiently attentive to the task. Auditors were not aware of each segment's category and simply judged each segment using the same set of questions.

Inter-annotator agreement was nearly perfect for all four linguistic varieties. All dual segments received identical judgments by both auditors, with one exception; for a single segment, one auditor accepted the segment as being "in their language" (Libyan Arabic), while the other auditor labeled the segment as containing insufficient speech to judge.

All distractor segments were judged as "not my language" by all auditors. The results indicate that native speakers can

reliably distinguish Maghrebi linguistic varieties from one another.

5. RESULTS

5.1. Data Yield

A total of 96 claques were enrolled in the MAGLIC study, with a minimum of 21 claques per language. Some enrolled claques never completed any calls, while others dropped out after making only one or two calls. Between 10-19 claques per language were productive in that they completed multiple calls meeting MAGLIC requirements. The 60 productive claques made a total of 631 calls, yielding at least 150 unique callees per language. The 631 MAGLIC collected recordings comprise a total of 110.7 hours of audio. Table 1 shows the number of completed calls per language, the number passing manual audit and the total audio yield per language.

	Collected recordings	Recordings passing manual audit	Total audio duration (hours)
Algerian Arabic	166	162	27.8
Libyan Arabic	158	152	28.5
Tunisian Arabic	157	155	26.6
North African French	150	147	27.8
Total	631	616	110.7

Table 1: MAGLIC Corpus Collection Yield

5.2. Corpus Structure

The MAGLIC corpus includes full telephone recordings, distributed as single channel 8-KHz, 8-bit a-law files. Documentation about the collection protocol is included, along with the guidelines for manual call auditing. The corpus also includes audio metadata (Table 2) and call auditing judgments (Table 3).

Field Name	Field Description
audio_id	6- or 7-digit numeric ID for the audio segment
datetime	yyyy-mm-dd hr:mn:sc = date of audio recording
btime	offset (secs) from recording start to seg start
duration	segment duration (seconds)
file_size	byte count of segment file
file_type	file format (flac)
md5_checksum	checksum of segment file
source_duration	source recording duration (seconds)
source_file	full file name of source recording
source	call-id and channel
segment_type	CTS
origin_info	anonymized phone number

Table 2: MAGLIC Corpus Audio Metadata

Field Name	Field Description
auditor_id	Numeric ID of auditor
audit_type	Target, distractor, confusable
auditor_lang	Language that the auditor is listening for
audio_id	6- or 7-digit numeric ID for the audio segment
language_code	Assumed language of the audio segment based on collection conditions
all_target_lang	Is all of the speech in [language]? (yes, no, no response)
off_target_lang	Auditor's comment if segment is not in their language
mostly_speech	Is there speech throughout most of this segment? (yes, no)
speech_clarity	How clear is the audio? (clear, some unclear, very unclear, no response)
single_speaker	Is all of the speech from a single speaker? (yes, no, unsure, no response)
native_speaker	Is the speaker a native speaker? (yes, no, unsure, no response)
speaker_sex	What is the speaker's sex? (male, female, unsure, no response)
noisy	Is this a noisy call? (yes, no)
kit_uid	audit assignment ID

Table 3: MAGLIC Corpus Audit Metadata

5.3. Evaluation Results on the MAGLIC Corpus

The LRE22 evaluation included 65 distinct system submissions. With respect to the MAGLIC data, results showed that the three Maghrebi Arabic dialects (Tunisian, Algerian and Libyan) were challenging for systems to distinguish from one another, while it was less difficult to distinguish North African French. System performance was also sensitive to segment duration, with shorter test segments resulting in some performance degradation (Lee et al., 2023).

The confusability of Maghrebi Arabic dialects by LRE systems stands in contrast to our findings that human raters can distinguish these varieties with a high degree of consistency, suggesting that there may be headroom for additional technology improvement in future.

6. CONCLUSION

The Maghrebi Language Identification Corpus (MAGLIC) represents an important new resource for language recognition, providing conversational telephone speech for four regional linguistic varieties: Algerian, Libyan and Tunisian dialectal Arabic plus North African French. The MAGLIC corpus was used for both development and test data in the NIST LRE22 Evaluation, with results showing confusability among the three Arabic dialects despite humans being able to distinguish these varieties with perfect accuracy.

The MAGLIC corpus will be made publicly available through publication in the LDC catalog after its use in closed evaluations has concluded.

7. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the work of Craig Greenberg, Yooyoung Lee and Omid Sadjadi from NIST, and Elliot Singer and Trang Nguyen from MIT-LL, who provided input to the corpus design and feedback on the collection itself. We also recognize the contributions of Dr. Mohamed Maamouri, and the many speakers, recruiters and other team members who contributed to the MAGLIC corpus.

8. REFERENCES

1. Lee, Y., Greenberg, C., Godard, E., Butt, A.A., Singer, E., Nguyen, T., Mason, L., Reynolds, D. (2023). The 2022 NIST Language Recognition Evaluation. Proc. INTERSPEECH 2023, 1928-1932, doi: 10.21437/Interspeech.2023-241.
2. Aguadé, J, (2018). The Maghrebi dialects of Arabic, in Clive Holes (ed.), Arabic Historical Dialectology: Linguistic and Sociolinguistic Approaches, Oxford Studies in Diachronic and Historical Linguistics (Oxford; online edn, Oxford Academic, 18 Oct. 2018)
3. NIST, (2011). The 2011 NIST Language Recognition Evaluation Plan (LRE11), https://www.nist.gov/system/files/documents/itl/iad/mig/LRE11_EvalPlan_releasev1.pdf.
4. NIST, (2015). The 2015 NIST Language Recognition Evaluation Plan (LRE15)"https://www.nist.gov/system/files/documents/2016/10/06/lre15_evalplan_v23.pdf
5. Sadjadi, S. O., Kheyrkhah,T.,Tong,A., Greenberg, C.S., Reynolds, D. A., Singer, E., Mason, L. P., and Hernandez-Cordero, J., (2018). The 2017 NIST language recognition evaluation, in Proc. Odyssey, Les Sables d´Olonne, France, June 2018, pp. 82–89.
6. Jones, K., Strassel,, S., Walker, K. and Wright. J., (2020).Call My Net 2: A New Resource for Speaker Recognition. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 6621–6626, Marseille, France. 2020. European Language Resources Association.
7. Strassel, S., Walker, K., Jones, K., Graff, D., Cieri, C., (2012). New resources for recognition of confusable linguistic varieties: the LRE11 corpus, In Odyssey-2012, 202-208.
8. Jones, K., Graff, D., Wright, J., Walker, K. and Strassel, S., (2016) Multi-language speech collection for NIST LRE, in Proc. LREC, Portoroz, Slovenia, May 2016, pp. 4253–4258.
9. Mubarak, H., Hussein, A., Chowdhury, S.A., Ali, A, (2021). QASR: QCRI Aljazeera Speech Resource -- A Large Scale Annotated Arabic Speech Corpus. https://doi.org/10.48550/arXiv.2106.13000
10. Wray, S., Ali, A. (2015). Crowdsource a little to label a lot: labeling a speech corpus of dialectal Arabic. *Proc. Interspeech 2015*, 2824-2828, doi: 10.21437/Interspeech.2015-594
11. Al-Fetyani, M., Al-Barham, M., Abandah, G., Alsharkawi, A., Dawas, M., (2021). MASC: Massive Arabic Speech Corpus, *IEEE Dataport*, doi: https://dx.doi.org/10.21227/e1qb-jv46. August 18, 2021
12. Djellab, M., Amrouche, A., Bouridane, A. et al., (2017). Algerian Modern Colloquial Arabic Speech Corpus (AMCASC): regional accents recognition within complex socio-linguistic environments. *Lang Resources & Evaluation 51, 613–641* (2017). https://doi.org/10.1007/s10579-016-9347-6
13. Droua-Hamdani, G., Selouani, S. A. and Boudraa, M., (2010). Algerian Arabic Speech Database (ALGASD): Corpus design and automatic speech recognition application. *ARABIAN JOURNAL FOR SCIENCE AND ENGINEERING*. (2010) 35. 157-166.
14. Bougrine, S., Chorana, A., Lakhdari, A. and Cherroun, H., (2017). Toward a Web-based Speech Corpus for Algerian Dialectal Arabic Varieties. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 138–146, Valencia, Spain. Association for Computational Linguistics. 2017.
15. Greenberg, C., et al., (2020). 2018 NIST Speaker Recognition Evaluation Test Set" LDC2020S04. Web Download. Philadelphia: Linguistic Data Consortium.
16. Zribi, I., Ellouze, M., Belguith, L. and Blache, P., (2015). Spoken Tunisian Arabic Corpus STAC : Transcription and Annotation. Research in Computing Science. 90. 10.13053/rcs-90-1-9. (2015).
17. Masmoudi, A., Khmekhem, M.E., Yannick, E., Belguith, L.H. and Habash, N. (2014). A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 306-310, Reykjavik, Iceland, May. European Language Resource Association (ELRA).
18. Iskra, D., Siemund, R., Borno, J., Moreno, A., Emam, O., Choukri, K., Gedge, O., Tropf, H., Nogueiras., A., Zitouni, I., Tsopanoglou, A., Fakotakis,. N., (2004). 10 OrienTel - Telephony databases across Northern Africa and the Middle East". http://www.lrec-conf.org/proceedings/lrec2004/pdf/552.pdf
19. Lee, Y. , Greenberg, C. , Mason, L. and Singer, E. (2022). NIST 2022 Language Recognition Evaluation Plan, Language Recognition Evaluation, [online] , https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=935161, https://lre.nist.gov (Accessed February 27, 2023)
20. Ryant, N. (2023) Linguistic Data Consortium Broad Phonetic Class Speech Activity Detector (ldc-bpcsad). Linguistic Data Consortium. https://github.com/Linguistic-Data-Consortium/ldc-bpcsad