Slavomír Čéplö,[1] Ján Bátora, Adam Benkato, Jiří Milička, Christophe Pereira, Petr Zemánek

Mutual Intelligibility of Spoken Maltese, Libyan Arabic and Tunisian Arabic Functionally Tested: A Pilot Study

Abstract: This paper presents the results of a project designed to functionally test the mutual intelligibility of spoken Maltese, Tunisian Arabic and Benghazi Libyan Arabic. We compiled an audio-based intelligibility test consisting of three components: a word test where the respondents were asked to perform a semantic classification task with 11 semantic categories; a sentence test where the task was to provide a translation of a sentence into the respondent's native language and a text test where a short text was listened to twice and the respondents were asked to answer 8 multiple-choice questions. We collected data from 24 respondents in Malta, Tunis and Benghazi which we analyzed to determine that there exists asymmetric mutual intelligibility between the two mainstream varieties of Maġribī Arabic and Maltese, where speakers of Tunisian and Libyan Arabic are able to understand about 40% of what is being said to them in Maltese, whereas that ratio is about 30% for speakers of Maltese exposed to either variety of Arabic. Additionally, we found that Tunisian Arabic has the highest level of mutual intelligibility with either of the other two varieties. Combining the intelligibility scores with comparative linguistic data, we were able to sketch out the phonological variables involved in enabling and inhibiting mutual intelligibility for all three varieties of Arabic and set stage for further research into the subject.

Keywords: Maltese, Tunisian Arabic, Libyan Arabic, mutual intelligibility, functional testing

1 Introduction

In Neo-Arabic dialectology, the concept of mutual intelligibility is often haphazardly invoked – whether in positive (Ryding 2005: 6) or negative terms (Abu-Haidar 2000: 93) – to conveniently illustrate various claims about the nature of the complex linguistic landscape that is Arabic and the relationship between its varieties. As one of those varieties, Maltese is also subjected to the same treatment, where the claims range from total lack of mutual intelligibility with any variety of Arabic (Owens 2010: 117) to anecdotal evidence asserting the ability of speakers of Arabic (usually Tunisian Neo-Arabic, cf. Chaouachi 2014: 127) to understand it nearly perfectly.

It is therefore remarkable that to date, there has been no rigorous study conducted aiming to investigate the mutual intelligibility of Neo-Arabic varieties at any level, even more so considering the fact that various methodologies have been successfully used for the very purpose in a number of other scenarios (such as Tang and van Heuven [2009] for topolects of Chinese or Delsing and Lundin-Åkesson [2005] for Scandinavian Germanic languages). This paper is the product of a field study which sought to remedy this omission. The study was conceived as focusing primarily on determining to what degree Maltese as an outlier and heavily contact-influenced variety of Arabic is

[1] **Corresponding author: Slavomír Čéplö**, Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague, Panská 890/7, 110 00 Prague 1, Czech Republic, E-mail: bulbul@bulbul.sk.
Ján Bátora, Sonic Studio s.r.o., Ružová dolina 6, 821 08 Bratislava, Slovakia, E-mail: jan.batora@sonic.sk.
Adam Benkato, Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22/23, 10117 Berlin, Germany, E-mail: benkato@gmail.com.
Jiří Milička, Institute of Comparative Linguistics, Faculty of Arts, Charles University in Prague, Celetná 20, 116 42 Prague 1, Czech Republic, E-mail: milicka@centrum.cz.
Christophe Pereira, INALCO - Institut National des Langues et Civilisations Orientales 65, Rue des Grands Moulins, 75013 Paris, France, E-mail: christophe.pereira@inalco.fr.
Petr Zemánek, Institute of Comparative Linguistics, Faculty of Arts, Charles University in Prague, Celetná 20, 116 42 Prague 1, Czech Republic, E-mail: petr.zemanek@ff.cuni.cz.
Names of co-authors are listed alphabetically, the contribution by individual authors is as follows: AB prepared Libyan test data and conducted field research in Benghazi, CP prepared Tunisian test data and conducted field research in Tunisia, JB designed and wrote the test application LingTest, JM and PZ assisted with the analysis and interpretation of data and SČ designed the experiment, assisted with the design of LingTest, prepared Maltese data, conducted field research in Malta, analyzed and interpreted the results and wrote the paper. As such, SČ assumes full responsibility for any and all errors.

mutually intelligible with mainstream Arabic dialects of the same subgroup – Tunisian Arabic and Benghazi Libyan Arabic – with the secondary purpose of establishing a standard tool and data kit for the functional testing of mutual intelligibility of all varieties of Arabic.

2 Varieties involved

All three varieties included in this project are branches of Maġribī (or North African) Arabic which is – along with Egyptian Arabic, Sudanese Arabic, Levantine Arabic, Arabic of the Arabian Peninsula and Mesopotamian Arabic – one of the major dialectal subgroupings of Neo-Arabic (Fischer and Jastrow 1980, Corriente and Ángeles 2008). The membership of Tunisian and Libyan Arabic in this groups is uncontroversial, that of Maltese, however, is somewhat complicated for reasons ranging from linguistic (such as certain Levantine Arabic features, cf. Fabri [2010]) through sociolinguistic (e.g. the use of Latin script) all the way to political. Primarily, however, the issue seems to be the use of the term "dialect" in Arabic dialectology. To avoid any controversy, we will therefore for the purposes of this paper define an "Arabic dialect" or "Arabic variety" as a "Semitic language which evolved from post-hijra Neo-Arabic tied to a particular geographical region" and trust that no one will object if Maltese is included in that definition. The membership of Maltese in the Maġribī Arabic dialectal group is then made clear with a review of its linguistic features (Corriente and Ángeles 2008: 381).

While the question of the position of all three dialects within the taxonomy of Neo-Arabic is easily settled, the same cannot be said of their synchronic relationship within the Maġribī branch of Neo-Arabic, an issue closely tied to that of mutual intelligibility. The general assumption holds that the closer the synchronic linguistic distance between two languages, the more easily will their speakers understand each other. This, however, is not an uncontroversial proposition, if only because the definition of linguistic distance is far from clear cut. In case of the three varieties in question, the issue is further complicated by the lack of data which would cover all three dialects: while there has been plenty published on Tunisian Arabic and Maltese, major descriptive works on Benghazi Libyan Arabic are over 70 years old (e.g. Panetta 1943) and new studies of the dialect have only begun to appear in print (Benkato 2014). Some progress has been made, such as the recent study by Hammett (2012) which examines the position of Maltese within Maġribī Arabic using the Cohen-Caubet-Roth dialectological questionnaire (Cohen et al. 2000) and determines that in linguistic terms, Maltese is closest to the dialect of Sousse and the Judeo-Arabic dialects of Tunis. The absolute as well as relative position of all three varieties within the North African dialectal subgroup, however, remains an open question. We hope to provide a partial answer by examining both the mutual intelligibility of these three dialects of Arabic as well as providing an analysis of the linguistic determinants of their mutual intelligibility (or lack thereof) and thus an overview of the synchronic relationship between them.

3 Test composition

3.1 Preliminaries

Gooskens (2013) provides a comprehensive overview of various existing methodologies developed to measure the mutual intelligibility of related varieties of a number of languages. While varied in approach and purpose, these methodologies essentially fall into two camps: opinion testing and functional testing. In opinion testing, respondents are asked to provide their impression of how well they understood speakers or speech samples provided. In functional testing, comprehension is

measured using some type of objective criteria. Having examined these methodologies, surveyed a number of previous studies and considered various practical issues, we decided to model our test after the functional test employed by Tang and van Heuven (2009) in their groundbreaking study of the mutual intelligibility of topolects of Chinese. We did so for a number of reasons: first, we wanted to conduct a functional test which Tang and van Heuven (2009) determined to be a more reliable way of testing mutual intelligibility than opinion tests. Second, we wanted to perform at least two types of tests to compare the results and evaluate their usefulness; in such a situation, word test and sentence test are the most obvious options for which Tang and van Heuven (2009) have established – to our mind – a very successful precedent in a linguistic landscape similar to that of Arabic. And last but not least, we decided to use Tang and van Heuven (2009) as a model for logistical reasons. The most important practical issue facing us was that of writing: Tunisian and Libyan Arabic are written (when used in writing at all) in both Arabic and Latin script without any standardized orthography while Maltese is written in Latin script only, using a number of idiosyncratic digraphs and diacritics. This naturally immediately ruled out the use of a written test and this is where the audio-only input procedure used by Tang and van Heuven (2009) seemed most reasonable and practical.

After some preliminary testing, we decided to perform a few modifications to the test procedure, the chief among being the addition of a text test to the word and sentence tests and the exclusion of a listener's native variety from testing: where in Tang and van Heuven's test involving 15 varieties of Chinese each respondent tested all the 15 varieties, in our test comprising 3 varieties of Neo-Arabic, each respondent only tested the two varieties that were not his own native variety.

3.2 Word test

In the preparation of the word test, we closely followed the procedure used by Tang and van Heuven (2009). We selected 160 words divided into 11 semantic categories (for the full list of word test items and categories, see Appendix A). There were three primary criteria for the selection of words: 1) high-frequency, 2) low neighborhood density (i.e. none of the words should be too similar to another one on the list) and 3) unambiguous identification of the semantic category the word belongs to. The combination of these criteria made it necessary for us to reach beyond the limited scope of wordlists used for similar purposes such as the Swadesh list, which we did by including everyday words describing shapes and properties of objects, household items, clothing and emotions. The application of criteria two and three also prompted the expansion of semantic categories from Tang and van Heuven's (2009) 10 to our 11. We excluded the "Verbs of action/things people do" category used by Tang and van Heuven (2009: 716) because of the salient nature of Neo-Arabic verbal morphology: since there is no equivalent of an infinitive in Arabic, the verbs would have to be presented in the 3rd person masculine singular perfect or the 3rd person masculine singular imperfect, both of which tend to have a rather conspicuous structure even in the first stem, doubly so in the derived stems. Consequently, we decided to distribute the verbs across categories and so 5 of the 11 categories contain at least one verb. The verbs are presented in the 3rd person masculine singular imperfect to increase the length of the audio input except for item W084C06 where the usage in both Tunisian and Libyan Arabic prefers the passive participle. At least one of the Sicilian-Italian borrowings typical for Maltese was also included in 10 of the 11 categories. For the purposes of analysis, words were sorted according to alphabetically arranged category and each word was assigned a code consisting of the letter W followed by a sequential three-digit number and the letter C (for category) followed by a two-

digit category number (see the list in Appendix A) resulting in each of the 160 words being given a unique code in the range W001C01 to W160C11.

3.3 Sentence test

As with the word test, we also set out to replicate the methodology used by Tang and van Heuven (2009) in the design of the sentence test. Soon, however, a number of concerns emerged. Tang and van Heuven chose the English SPIN test (Kalikow et al. 1977) as the basis for their sentence test. The SPIN test consists of two sets of sentences where the listeners' task is to correctly identify the last word. In one set of sentences, that word is easily inferred from the content, in the other, it is not. The fundamental principle of the SPIN test lies in comparing the word identification rate for high-predictability sentences with that of low-predictability sentences. Tang and van Heuven, however, opted to use only the high-predictability set, thus casting some doubt on the justification for the use of the SPIN test, especially considering its cultural bias and the resulting choice of vocabulary.[2] Additionally, with only one data point provided, the SPIN test would be best described as a "word in context" test and thus the question arose whether in the context of Neo-Arabic varieties, there is any significant difference between the word test and this type of sentence test.

Having considered all of that, we decided to stick with the general methodology, but opted to adopt a slightly different approach and model the sentence test after the Bamford-Kowal-Bench Standard Sentence Test (BKB-R). This test (already used for a similar purpose by Bent and Bradlow [2003]) consists of simple sentences of no more than 8 words, each with three or four keywords (both content and functional words). The respondents' task is to write down what they heard and the response is evaluated based on all the keywords. In the conditions of our study, this would essentially be a translation test and we implemented it as such. Based on the list in Bent and Bradlow (2003), we compiled a list of 60 simple sentences (mostly declaratives, but also some questions and imperatives) and each sentence was assigned 3 or 4 keywords for a total of 219 keywords. The keywords consisted of selected items from the word test supplemented by functional words (pronouns and prepositions) and a number of common verbs ('to bring', 'to ask', 'to reply' etc.).

In the original test design (which was ultimately not implemented), the sentences were divided into 8 categories based on isoglosses distinguishing Maltese from mainstream Arabic dialects such as merger and ultimate loss of [ġ][3] and [ʕ] (in Maltese, the sounds [ġ] in Tunisian Arabic *ġīra* 'envy' and [ʕ] in *ʕarūsa* 'bride' have first merged into a single sound represented in the orthography by *għ* giving *għira* and *għarusa*), strong *imāla* (i.e. raising of [ā] to [ī] as in the first vowel in Maltese *jiekol* [yīkol] 'he eats' vs. Tunisian Arabic *yēkəl* and Libyan Arabic *yākəl*) and Sicilian Italian borrowings (such as Maltese *missier* 'father' vs. Tunisian Arabic *bu* and Libyan Arabic *bāṭ*). In each sentence in each category, one keyword (termed "targeted keyword") represented that isogloss and was to be translated with a cognate with the purpose of determining to what extent these uniquely Maltese linguistic developments inhibited intelligibility with more mainstream varieties of Maġribī Arabic. Ultimately, however, this proved to be unrealistic as a fluent translation often could not accommodate the selected word without sounding too literal or stilted and, consequently, the concept was abandoned. It survives

[2] The sentence set includes items such as "The king wore a golden crown", "The farmer baled the hay" and "Cut the bacon into strips" which pose some difficulty in their transfer to different cultural contexts.

[3] In the tests and in this paper, we use square brackets to provide a phonetic transcription for sounds and Maltese words which are normally written in standard Maltese orthography. For the transcription, we use the DIN 31635 standard with the following modifications: for reasons of legibility and ease of computational processing, [ʕ] (IPA number 145) is used for the voiced pharyngeal fricative and [ʔ] (IPA number 113) is used for the glottal stop.

in the final test design in the category numbers consisting of the letter C and a sequential two-digit number which are added to the sentence codes made up of the letter S followed by a three-digit number resulting in each sentence being given a unique code in the range S001C01 to S060C08 where the sequence C01-C08 stands for one of the abandoned categories. For the full list of sentence test items, see Appendix B.

3.4 Text test

Recorded Text Tests (RTT) have been a standard tool for determining mutual intelligibility of closely related varieties for some time now, favored especially in the analysis of the relationship between unwritten languages by SIL (e.g. Casad 1974). The procedure commonly involves playing each text twice where the second replay is interrupted at intervals to ask a context-relevant question and record the answer. Confident in their field-tested utility even despite certain criticisms (chiefly the issue of to what extent they really test language comprehension instead of text comprehension and whether answering questions is a good measure of comprehension at all, cf. Bouwer 2007: 264-265]), we decided to incorporate a text test into our test suite, however, not without some reservations. Primarily, our concern was that with the typical length of a text test at 1 to 3 minutes, the interruptions required for asking questions and the comparatively long periods necessary to record them would break the respondent's concentration and ultimately turn the procedure into another sentence test. Having experimented with a number of technical solutions to that problem, we ultimately decided to implement the text test as a multiple-choice answer test of the type used in language learning, such as the TOEFL® Listening Comprehension test.[4] We selected two texts for their relatively simple vocabulary and low memory load, one from a test used for a listening exam at a Maltese primary school[5] and one from a beginner textbook of Maltese (Vella 1994: 144). For the former, we used the test's original 8 questions adding one option to bring the total of choices to four; for the latter, we added 8 questions with four options each.

4 Test delivery

4.1 Material preparation

All the test items were first compiled in English and then translations and recordings were made into each of the three varieties. A single male native speaker was selected for each variety; all were born and raised until the age of 18 in their respective country and region. For Maltese, the recordings were done in Malta. Recordings into Tunisian Arabic were done in Paris, while the speaker of Libyan Arabic was recorded in London.

In case of Maltese, the translations were done beforehand, proofread and recorded on a PC using a standard desktop microphone. For the other two varieties, the translations (including those of the questions for the text test) were done on the fly during the recording performed using a Zoom H2 Handy Recorder (Libyan Arabic) and the default recording application in iPhone / Nokia 8 (Tunisian Arabic). A small number of inevitable issues resulting from this process was fixed in retakes for

[4] Cf. http://www.ets.org/toefl/ibt/about/content/ (accessed on 4 October 2015).
[5] Kulleġġ San Ġorġ Prezza in Ħamrun. The test was given to 4th Form pupils in 2013 and is also available online at http://sgpc.skola.edu.mt/resources/hyprimary2013/Yr%204%20Malti%20Smigh%20HY%20Exam%202013%20Ghalliema.pdf (accessed on 4 October 2015).

Libyan Arabic. Due to lack of time, a few minor issues in Tunisian Arabic recordings remained unresolved.[6]

The resulting WAV files (channels: stereo, codec: PCM, sample rate: 44000, bit depth: 24) were processed with the help of Adobe Audition CS6: first, any residual noise was removed using the Capture Noise Print / Noise Reduction functionality and then the volume was normalized to -3 dB. Additionally, an audio cue consisting of a 0.7 second level tone followed by a 0.1 second silence was prepended to each word. Initial testing suggested that respondents found it difficult to even realize that what they had just heard was a word, as the average length of a word recording was under a second. By adding the audio cue and thus extending the total length of word input to approximately 2 seconds, we resolved the issue. In the final step, the edited WAV files were cut into individual component files (160 words, 60 sentences and 2 texts) which were then converted to M4A using iTunes and prepared for delivery.

4.2 LingTest

Very early in the test preparation stage, we became aware of the practical challenges with regard to its administration, from the rather complex issue of randomization down to the simple matter of how to present input and record the response. Having considered the available options, we decided to make full use of modern technology and employ a touchscreen device with a custom testing software. As the device and platform, we selected the Apple iPad Mini 1st generation with iOS 7 for its compactness, reliability and user-friendliness and paired the device with Koss SB/45 headphones. For the actual software solution, we designed an application called LingTest which was used to administer and evaluate the tests. In what follows, we will briefly describe the functionality of the application and its use in testing. [7]

LingTest was designed as a modular application with data as independent of the functionality as possible. The data is imported into the application in the form of a ZIP archive with audio as M4A files and textual (instructions, category names with associated images, questions etc.) and structural information (languages, test components etc.) in descriptive XML files. For this project, each package contained one set of data per each variety tested, so in Malta, the package contained data for Libyan Arabic and Tunisian Arabic.

The application itself consists of six parts: admin screen, respondent info screen, word test, sentence test, text test and evaluation module. Admin screen and evaluation module are used to prepare and evaluate the test and are only accessible to the person administering the test. The admin screen (see Figure 1) contains a list of imported packages with package information (languages, test components, number of elements in components etc.); a menu item to select the language in which the test will be administered (currently English, Czech, Maltese and Modern Standard Arabic), selection buttons with number entry fields and a "Start test" button.

[6] See the comments in Appendix A.

[7] A detailed description of the application, including its inner workings and data structures, will be published by the present authors as "Introducing LingTest: A Field-Friendly Application for the Functional Testing of Mutual Intelligibility of Related Varieties".

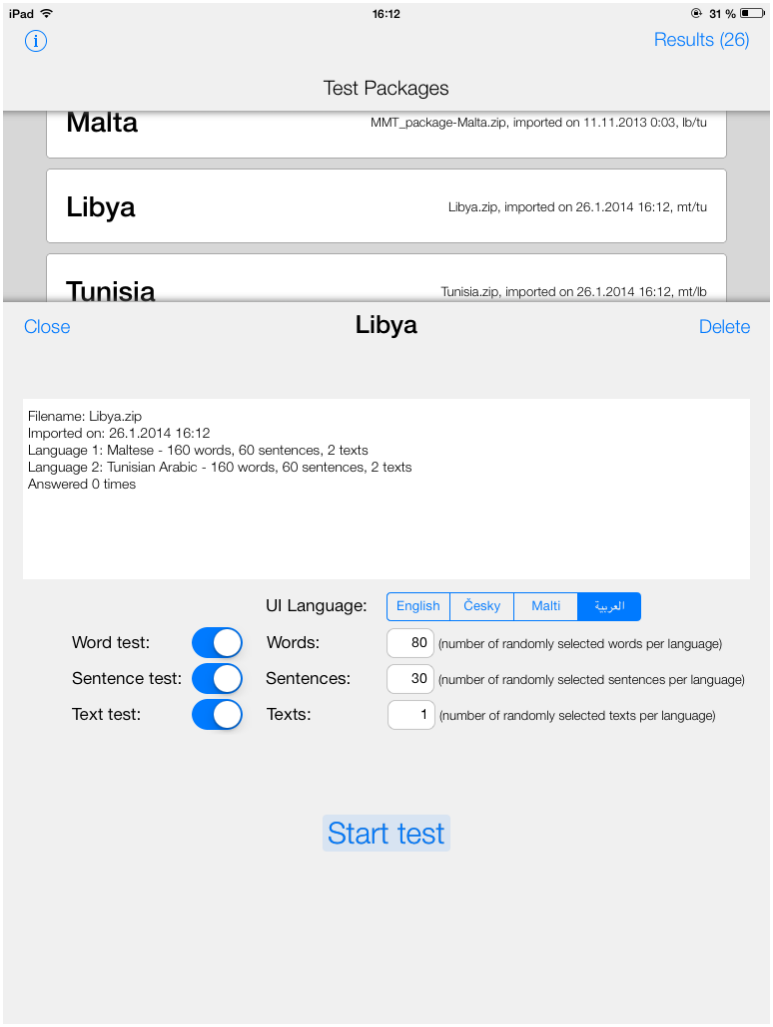

Figure 1: The admin screen of LingTest

The selection buttons with number entry fields enable the user to customize the test by a) selecting one or more from the three available test components (word test, sentence test and text test) and b) by selecting the number of items in each of the components. The latter setting is used to select a randomized subset of test items in case the full set is too extensive. In this project, all three components were selected and the defaults for the number of items were set at one half of the total number of items (i.e. 80 words, 30 sentences and 1 text) per each variety tested. Upon tapping the "Start test" button, the application uses the Objective-C function *arc4random()*[8] to randomly select and order the specified number of items for each test component and variety in a Latin square matrix. This was done to eliminate any priming effects by ensuring that each item is played only once during a single test. Additionally, the application recorded which items have been used and once a test was successfully completed, stored that information to make sure that only those items not yet tested would be selected for the next round. With the default settings, two respondents were required to test every item in the test (i.e. the full data set) exactly once.

[8] See https://developer.apple.com/library/ios/documentation/System/Conceptual/ManPages_iPhoneOS/man3/arc4random.3.html (accessed on October 4 2015) for a detailed description of the function.

Once the test starts, the respondent is first presented with the respondent information screen where they are asked to provide some basic demographic data, including age, education, place of residence in the last 5 years and native language (including that of each parent). Once the information is filled out and confirmed, the actual test starts. There is no time limit on any component or question, so the respondents take as long as they like.

Each component begins with an introductory screen describing the task at hand and providing a feature to test the audio volume. The word test introductory screen contains a brief description of the semantic categorization task along with four samples of lexical items and their respective categories. When the respondent is ready to begin, they press the "Next" button and the answer screen appears where, for each word, the audio is played (see Figure 2, left). The respondent's task is to select the correct semantic category by tapping one of 11 icons representing that category as both text and a simple black-and-white image and then tap "Next" to proceed to the next word (which the respondent can only do when one of the icons was selected). After the last word, a screen appears notifying the respondent of the conclusion of the word test and the application proceeds to the sentence test.

For the sentence test, the procedure is much the same, except the respondent is instructed to provide a translation of what they just heard with the actual instruction being "Write down what you've just heard in your language". To do that, they have the option of using the keyboard (see Figure 2, right) or writing freehand (i.e. drawing the letters with their finger on a specifically designated portion of the screen). When the respondents are satisfied with their answer, they tap the "Next" button and the application proceeds with the next sentence until all sentences are played for each variety tested and the conclusion screen appears.

Figure 2: Word test (left) and sentence test (right) screens of LingTest

In the text test, each text is played twice while the screen displays a running timer. Once the text has finished playing, 8 questions (with four choices each) appear one after another on the screen. The respondent is asked to select one correct answer and then tap "Next". When the last text finishes playing, a "Thank you" screen is displayed. Upon tapping on it, the evaluation data is saved (including test selection data) and the admin screen displays again.

4.3 Evaluation

The admin screen contains a link named "Results" which opens the evaluation screen (see Figure 3). This contains a list of all completed tests, ordered by packages. When an item on the list is tapped, the evaluation record appears which consists of an overview of respondent data and the answers for all test components. For both the word and the text test, the answers are evaluated automatically: the descriptive XML files in the test package include correct answers and once a test has been completed, the correct answers will appear marked by a green check mark on the evaluation record.

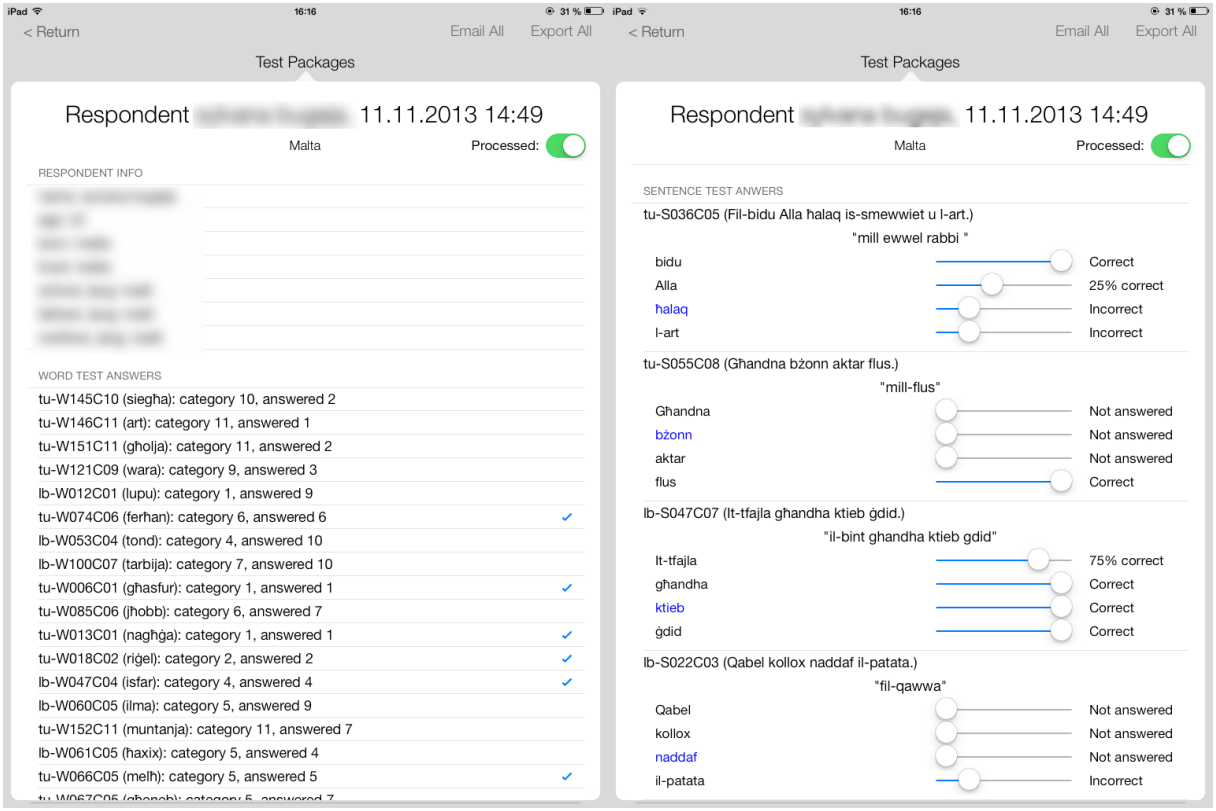

Figure 3: Word test evaluation (left) and sentence test evaluation (right) screens of LingTest

The answers for the sentence test need to be evaluated manually. For that, each respondent's answer is displayed on the screen next to the correct answer and the list of the assigned keywords with four sliders with five options: "Not answered", "25% correct", "50% correct", "75% correct" and "Correct". The detailed evaluation instructions can be found in Appendix C.

The primary authors AB, SČ and CP were in charge of conducting the testing in each of the three countries and they were also responsible for evaluating the sentence test. The application of the evaluation instructions (see Appendix C) was not always entirely straightforward and while the

authors made every effort to diligently evaluate each answer, in case of doubt, a false negative was deemed preferable to a false positive.

5 Results

5.1 Extraction and analysis

The results of the evaluations were exported from LingTest as XML files. Relevant data was extracted into CSV files using Perl scripts and then analyzed and visualized with R. All the raw data (including the Perl and R scripts) is available at www.bulbul.sk/GAUK58313.

5.2 Respondent information

Beginning in Malta, we set out to record as many responses as possible within a period of a month and then proceed to record the same number in the remaining countries. In total, 24 responses were collected from each of the three countries obtaining a total of 12 full data sets.[9] Table 1 below provides a summary of the respondents' demographic data.

Table 1: Respondent information by country

Country[10]	**Age**		**N females**	**Education**
	Mean	**SD**		
Malta	25.17	9.68	18	2.83
Libya	26.75	9.48	2	2.83
Tunisia	21.79	2.38	18	3.00

Age = mean and standard deviation of age in years. N females = number of female respondents (out of 24). Education (highest level attained): 0 = none, 1 = elementary, 2 = secondary, 3 = university.

In Malta and Tunisia, respondents were primarily recruited from among university students. In Libya, respondents came largely from the same age group and same educational background, but varying current employment status. None of the three groups of respondents had come into any extensive contact with any of the other two varieties. One respondent in Malta reported some work-related exposure to Arabic, but upon closer examination, it was determined that their knowledge did not go beyond the very basic conversational vocabulary which would not interfere with the test.

5.3 Word test[11]

[9] In Malta and Tunisia, the actual number of respondents interviewed was 26 and 27, respectively, but due to issues of technical nature, only 24 responses for each country were usable. In Malta, two respondents were recorded using an early version of LingTest in which the randomization functionality was not implemented correctly. In Tunisia, response 1 was a test run after which LingTest was not properly reset. This forced us to discard the full data set, i.e. response 1 and response 2. Response 27 was without a pair and thus discarded as well.

[10] In what follows, we will use the term "country" as a shorthand for "listener variety". For brevity's sake, we will use codes in the form of XX/YY where XX indicates the listener variety (MT = Maltese, LB = Libyan Arabic and TU = Tunisian Arabic) and YY the variety tested.

[11] Due to an error in the LingTest package used to administer the test in Malta, a small correction had to be made in the word data: categories 8-11were labelled incorrectly in the descriptive XML files and thus while the correct icon and description were presented to the respondent, the wrong label was recorded in the results and the evaluation. Consequently, a manual correction had to be made to the results data by relabeling the categories in answers as follows: 8>11, 9>8, 10>9 and 11>10. Both sets of CSV files are available in the raw data package.

Table 2 summarizes the results of the word as the mean of scores for all 24 respondents where we first calculated the mean of correctly answered questions for each respondent and then computed the mean of all 24 respondents per every country/variety combination, Figure 4 provides a bar plot with confidence intervals obtained using boostrap resampling of those means.[12] Note that with the lowest score at 22, the p-value for the binomial probability for this outcome is well below 0.001,[13] indicating that this and all the other results are extremely unlikely to have been achieved by guessing alone.

Table 2: Correctly assigned words (mean for all respondents, in %)

Country / Language	Maltese	Libyan Arabic	Tunisian Arabic
Malta	x	38.13%	37.14%
Libya	44.32%	x	73.07%
Tunisia	45.00%	79.58%	x

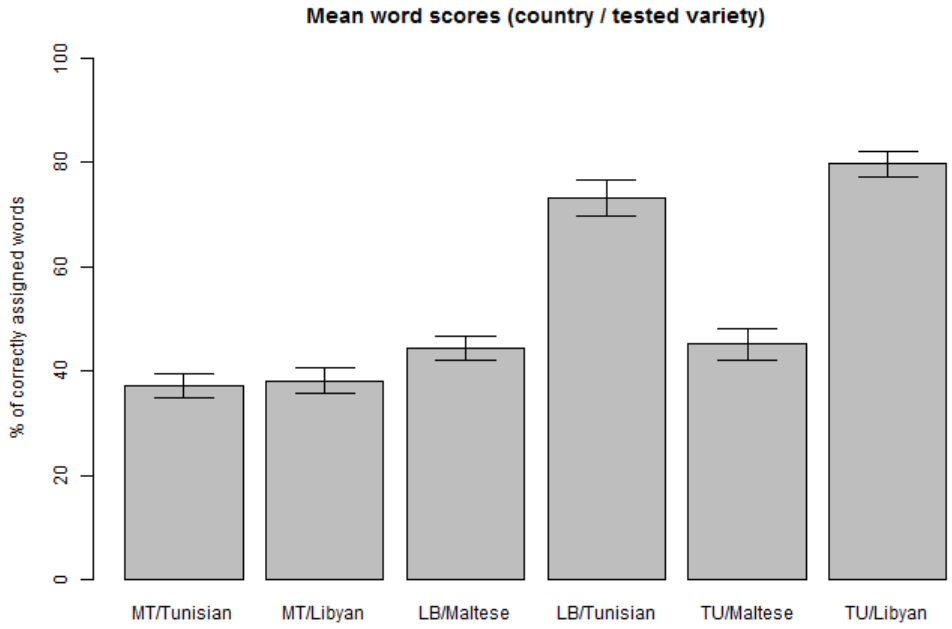

Figure 4: Correctly assigned words with confidence intervals obtained by bootstrap resampling

It is interesting to observe that there is no statistically significant difference[14] between the degree to which speakers of Maltese were able to identify isolated words in either of the other two varieties. Likewise, speakers of Tunisian and Benghazi understood their Maltese counterparts roughly at the

[12] Calculated in R using the function boot() with 1,000,000 replications (cf. Canty and Ripley 2014 and Davison and Hinkley 1997).
[13] Calculated in R using the function binom.test() with 22 successes on 80 trials and probability of success on a single trial at 0.09 for $p = 1.488 \times 10^{-6}$ resulting in the rejection of the null hypothesis (that the results were achieved by random guessing).
[14] In what follows, the comparison of two sets of data was calculated on the full set of data per respondent (24 data points per language pair) using the R function t.test() to perform a paired two-tailed Welch's t-test with 95% confidence interval. The normality of distribution required for the T-test was verified using the R implementation of the Shapiro-Wilk normality test (the R function shapiro.test()) and an inspection of Q-Q plots (using the R function qqnorm()).
In this case, for speakers of Maltese exposed to both mainstream varieties of Arabic, the p-value was 0.52 and consequently, the null hypothesis (that the results for Tunisian Arabic and Libyan Arabic are the same) cannot be rejected.

same rate,[15] although they were slightly better at it, suggesting the asymmetrical nature of mutual intelligibility between both Tunisian and Libyan Arabic on one hand and Maltese on the other.[16] And finally, the difference between the mutual intelligibility of the two mainstream varieties of Maġribī Arabic was statistically significant,[17] suggesting that the rate at which speakers of Tunisian Arabic understand Libyan Arabic is higher than that of speakers of Libyan Arabic exposed to Tunisian Arabic.

5.4 Sentence test

For the results of the sentence test, evaluation scores for each keyword were converted to percentages whereby evaluation scores "Not answered" and "Incorrect" scores were conflated to 0%, the "Correct" score was translated to 100% and evaluation scores 25%-50%-75% were assigned weights and converted to 10%-25%-85% to better reflect their contribution to the overall comprehension of the sentence: in other words, since a single keyword scoring 25% or 50% can impede the comprehension of the entire sentence, we penalized those scores. Arithmetic mean of the entire set of evaluation scores was calculated for each sentence to provide a total correctness score (TCS) of the sentence. These were then grouped into three categories: "sentence understood" for TCS 100%-85%, "sentence partially understood" for TCS 84%-45% and "sentence not understood" for sentences with TCS below 45%. The primary reason for this is the comparison of our methodology to that of Tang and van Heuven (2009): in their methodology, only a single word had to be correct for the sentence to be considered understood. In our methodology, 3 or 4 words had to score at 75% (85% with weights) for the sentence to be deemed fully understood and we therefore wanted to maintain a distinction between such fully understood sentences and sentences where – as it often happened – 3 of the 4 keywords scored 100%, but the remaining one scored 0%. At the same time, we wanted to identify sentences that were not understood at all, hence the need for three categories.

We then calculated mean TCS values for each respondent to obtain a set of 24 data points; table 3 provides the mean of those values per country and variety; Figure 5 plots the same data with confidence intervals obtained by bootstrap resampling.[18]

Table 3: Mean TCS score for the sentence test (for all respondents, in %)

Country / Language	Maltese	Libyan Arabic	Tunisian Arabic
Malta	x	23.86%	33.39%
Libya	28.90%	x	70.16%
Tunisia	32.18%	67.80%	x

[15] The p-value obtained using the same procedure as above for speakers of both mainstream varieties exposed to Maltese is 0.7 indicating that the null hypothesis (that the results for the two varieties are the same) cannot be rejected.
[16] Following the same procedure as above, we obtained p-values of 0.0002 for the mutual intelligibility of Libyan Arabic and Maltese and 0.0005 for the mutual intelligibility of Tunisian Arabic and Maltese indicating that in both cases, the null hypothesis (that the results for both directions are the same) must be rejected.
[17] The p-value obtained using the same procedure as above for the mutual intelligibility of Tunisian and Libyan Arabic is 0.008 showing that the null hypothesis (that the results for both directions are the same) must be rejected.
[18] Calculated in R using the function boot() with 1,000,000 replications (cf. Canty and Ripley 2014 and Davison and Hinkley 1997).

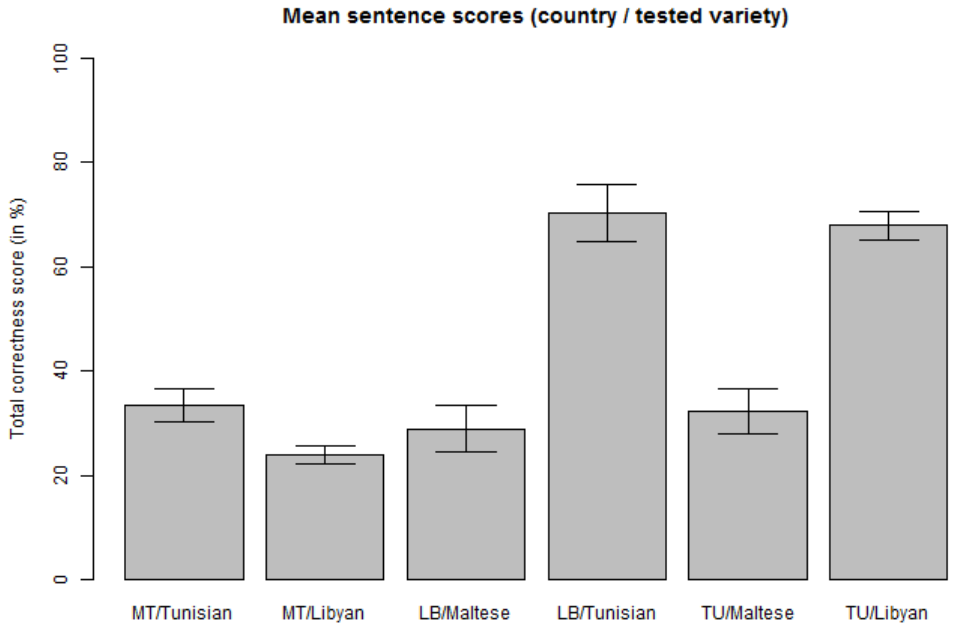

Figure 5: Mean total correctness scores (TCS) with confidence intervals obtained by bootstrap resampling

A more accessible overview of the results is perhaps provided by averaging the number of sentences the respondent understood fully (i.e. those with TCS => 85%).

Table 4: Fully understood sentences (mean for all respondents, absolute figures out of 30)

Country / Language	Maltese	Libyan Arabic	Tunisian Arabic
Malta	x	2.4	3.2
Libya	5.1	x	16.0
Tunisia	5.1	11.5	x

The asymmetrical nature of mutual intelligibility of Maltese and the two mainstream Arabic dialects noted in reference with the word test is once again apparent, but only for Libyan Arabic,[19] and it is even more obvious when considering only sentences with TCS =>85%. This is unsurprising, as this time, there is a statistically significant difference between how well the two mainstream varieties of Maġribī Arabic are understood in Malta, with Tunisian comprehended better than Libyan Arabic.[20] On the other hand, there is no statistically significant difference in the intelligibility of Maltese to speakers of either mainstream Maġribī dialect according to either measure[21] and same is true for their mutual

[19] The p-value for the mutual intelligibility of Maltese and Tunisian Arabic obtained as per procedure described above is 0.07 indicating that the null hypothesis (that the results for both directions are the same) cannot be rejected. On the other hand, the p-value of the test of mutual intelligibility data for Maltese and Libyan Arabic is 0.05 indicating that in this case, the null hypothesis (that the results for both directions are the same) can be rejected with 95% confidence.

[20] For speakers of Maltese exposed to either of the remaining two varieties, the p-value calculated using the procedure above was 2.191×10^{-5} and consequently, the null hypothesis (that the results for both pairs of varieties are the same) must be rejected.

[21] The p-value obtained by the same procedure as above using the TCS data for speakers of Tunisian Arabic and speakers of Libyan Arabic exposed to Maltese is 0.362 indicating that the null hypothesis (that the results for both pairs of varieties are

intelligibility, at least when it comes to TCS.[22] However, when considering only fully understood sentences, we observe that speakers of Libyan Arabic are much better at understanding their counterparts in Tunisia than the other way around.

Our test suite offered the respondents an option of indicating they haven't understood anything. Table 5 below summarizes the average number of such responses per respondent.

Table 5: Answer not attempted (total / average out of 30 per respondent)

Country / Language	Maltese	Libyan Arabic	Tunisian Arabic
Malta	x	271 / 11.3	214 / 8.9
Libya	350 / 14.6	x	94 / 3.9
Tunisia	334 / 13.9	46 / 1.9	x

While no accurate measure, this data provides a rough picture of how much confidence the respondents had in their ability to understand the tested variety. It is interesting to note that just as there was no significant difference in how well speakers of the two mainstream dialects understood Maltese, there is no difference in the way their speakers approached the task, i.e. speakers of Libyan Arabic display just as much confidence (or lack thereof) in their ability to understand Maltese as their Tunisian counterparts. On the other hand, the confidence with which speakers of Maltese translated Tunisian and Libyan Arabic mirrors the results obtained by TCS scores indicating that in Malta, Tunisian Arabic is both perceived as being easier to understand and actually understood better than Libyan Arabic.

In conclusion, two methodological asides: as we noted above, the test performed by Tang and van Heuven only required one correct word for the answer to be judged correct whereas in our test, there were three or four keywords which all had to be answered correctly for the sentence to be deemed understood. In the preparation stage, we worried that with only one data point analyzed, the sentence-intelligibility test as implemented in the form used by Tang and van Heuven would essentially duplicate the word test. Having performed some informal preliminary testing on a small set of sentences using both the SPIN and the BKB-R test, we determined that the SPIN test would not provide an accurate assessment of the mutual intelligibility of sentences in our context and opted therefore to use the BKB-R test. This conclusion is supported by the final respondent data, more specifically, a comparison of figures for sentences with TCS => 85% (i.e. sentences deemed fully understood in our test) and all those where the last keyword was given the 85% or 100% score (i.e. correctly understood sentences according to methodology employed by Tang and van Heuven [2009]) in Table 6. The large number of what we consider false positives (i.e. sentences where the last keyword was translated correctly, but the rest of the keywords were not) for all country/language combinations shows that at least for Neo-Arabic varieties, the BKB-R test is a more accurate measure of actual comprehension than the SPIN test.

the same) cannot be rejected. This is also borne out by the fact that the average number of fully understood Maltese sentences is the same for both pairs.

[22] The p-value obtained using the procedure above with the TCS data for speakers of Tunisian Arabic and speakers of Libyan Arabic exposed to the other variety is 0.5 indicating that the null hypothesis (that the results for both pairs of varieties are the same) cannot be rejected.

Table 6: Fully understood sentences (mean for all respondents)
Our methodology (TCS => 85%) / SPIN test according to Tang and van Heuven (2009)

Country / Language	Maltese	Libyan Arabic	Tunisian Arabic
Malta	x	2.4 / 7.3	3.2 / 11.6
Libya	5.1 / 8	x	16.0 / 20.6
Tunisia	5.1 / 11.7	11.5 / 20.3	x

And finally, the application LingTest allowed the respondents to record their responses either using a keyboard or writing freehand (i.e. by moving their finger across a dedicated portion of the screen). It is remarkable (and not only from the point of view of graphical user interface design) that in Malta and Libya, only a handful of people selected the freehand option – 2 in Malta (with 1 and 5 sentence responses) and 3 in Libya (with 2 respondents only providing 1 answer each in this manner and 1 respondent giving 7). In contrast, in Tunisia, 7 respondents chose to write freehand, 6 of whom provided most of their translations in this way for a total of 263 responses.

5.5 Text test

Table 7 summarizes the results of the text test as percentages of correct answers (out of 8) to the multiple-choice questions. Figure 6 provides a bar plot of the results with confidence intervals obtained using boostrap resampling of means for all respondents.[23]

Table 7: Correctly answered questions (mean for all respondents, in %)

Country / Language	Maltese	Libyan Arabic	Tunisian Arabic
Malta	x	48.96%	47.40%
Libya	48.44%	x	76.04%
Tunisia	55.73%	81.25%	x

[23] Calculated in R using the function boot() with 1,000,000 replications (cf. Canty and Ripley 2014 and Davison and Hinkley 1997).

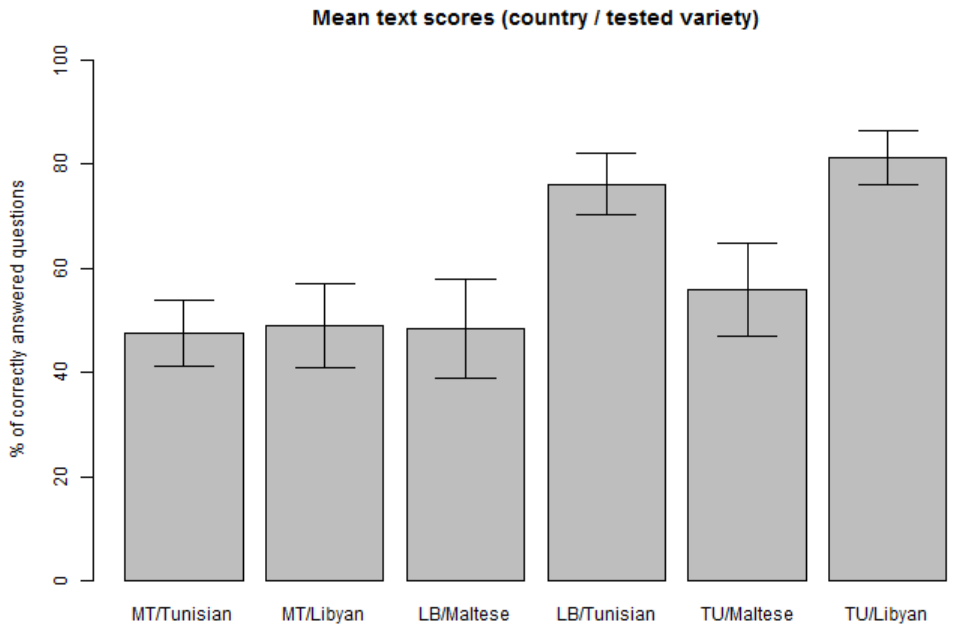

Figure 6: Correctly answered questions with confidence intervals obtained by bootstrap resampling

Both the wide confidence intervals and the binomial probability[24] indicate the low reliability of the text test as implemented in this project, raising questions regarding its utility in its present form. Nevertheless, some relatively clear trends can be observed and so, for example, one can note that the mutual intelligibility of the two mainstream varieties of Maġribī Arabic is higher than that of either of these varieties with Maltese. On the other hand, there is no statistically significant difference between the performance of speakers of Tunisian Arabic and that of their counterparts in Benghazi while speakers of Maltese once again show no preference for either of the mainstream Maġribī dialects. Consequently and, in contrast to the other two tests, the asymmetrical nature of the mutual intelligibility between Maltese and Libyan Arabic is nearly completely gone, with both groups of respondents performing nearly identically and same holds true of the mutual intelligibility between Tunisian and Libyan Arabic.

It is interesting to note that for all countries and variety combinations (save Tunisia with Libyan Arabic), there was a statistically significant gap in the scores for the two texts (see Table 8 below). This shows that despite comparable levels of vocabulary (one text is taken from an elementary school reading comprehension, the other is from a beginners' textbook), text T002 was much easier to comprehend than text T001. It is our hypothesis that this was due to the salient nature of the narrative in T002 which provided plenty of cognitive anchors. T001, on the other hand, was somewhat repetitive in nature (e.g. there were three groups of protagonists, all dogs), which may have increased recognition effort and memory load.

[24] The lowest (rounded) average score is 4 correct answers out of 8 (4 successes on 8 trials with a probability of 25% on a single trial) which translates to a (non-cumulative) p-value of 0.08. Consequently, the null hypothesis (that the results were achieved by random guessing) cannot be rejected, especially seeing as in all three countries, the lowest score for any listener variety was 1 correct answer out of 8. In other words, we cannot be certain that the results were not achieved by guessing alone, hence the low reliability of the text test as a whole.

Table 8: Average of correctly answered question for either text (in %) with significance test p-values[25]

Country / Language	Maltese T001 / T002	p-value	Libyan Arabic T001 / T002	p-value	Tunisian Arabic T001 / T002	p-value
Malta	x	x	35.42% / 62.5%	0.001	39.58% / 55.2%	0.06
Libya	38.54% / 58.33%	0.02	x	x	67.7% / 84.38%	0.02
Tunisia	40.63% / 70.83%	0.002	79.17% / 83.33%	0.5	x	x

5.6 Correlation between results for individual test components

Having examined the intelligibility data for the individual components, we now turn to the issue of the relationship between them. In other words, the question we ask is whether the respondent's performance in one test component can predict how well they will do in another. To answer it, we plotted the 24 sets respondent data for each test component in the form of a scatterplot matrix and calculated the Pearson correlation between individual components (see Figs. 7-9 below).

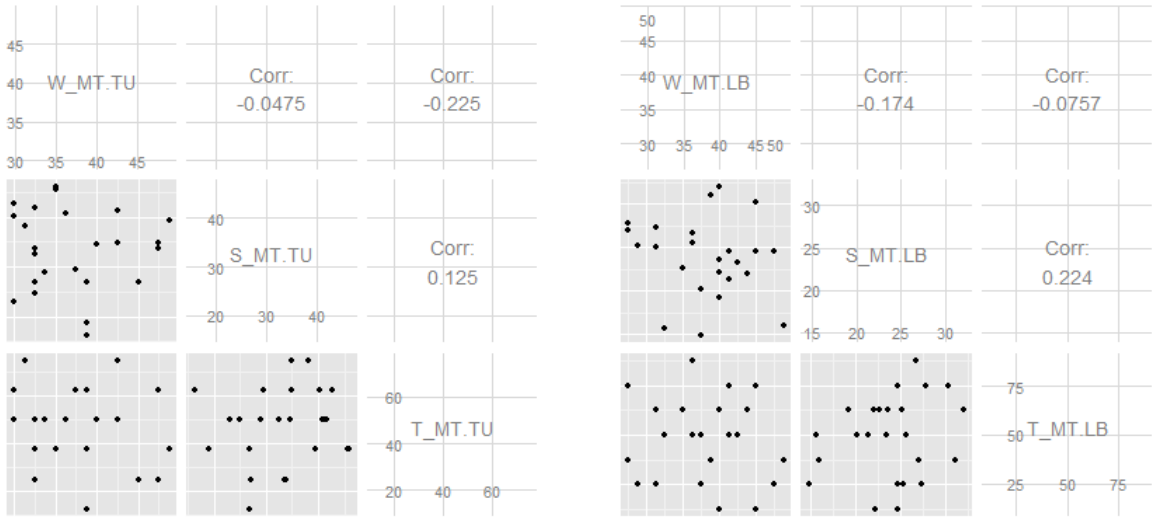

Figure 7: Scatterplot matrices of correlation data for all three test components (Word, Sentence and Text) administered to speakers of Maltese (MT) in Tunisian Arabic (TU, left) and Libyan Arabic (LB, right).

[25] Calculated on the full set of data per respondent (24 data points) using R function t.test() to perform a paired two-tailed Welch's t-test with 95% confidence interval to determine whether the null hypothesis (that the average performance of respondents is the same for both texts) should be rejected (if p-value is lower than 0.05).

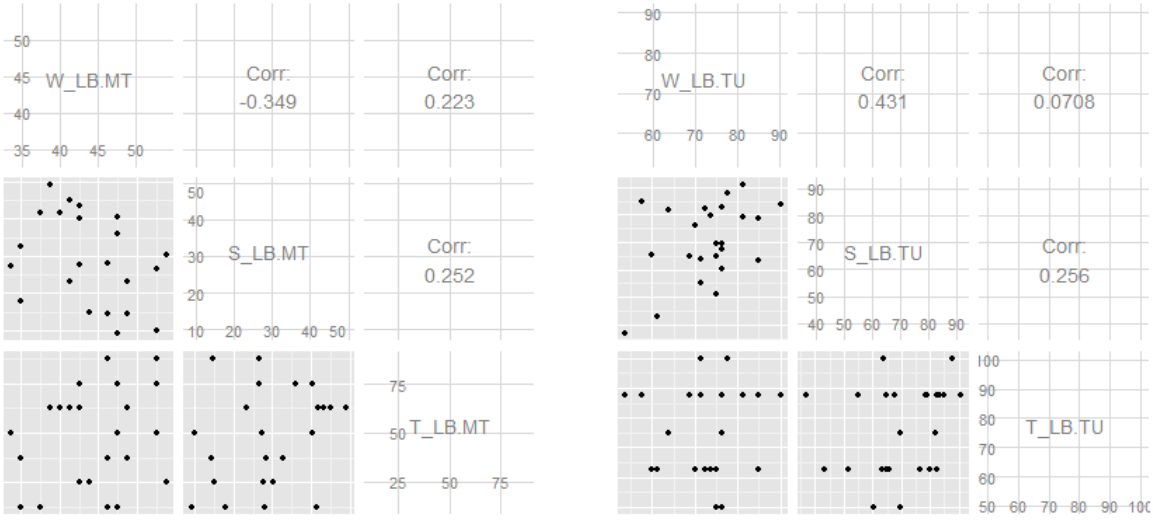

Figure 8: Scatterplot matrices of correlation data for all three test components (Word, Sentence and Text) administered to speakers of Libyan Arabic (LB) in Maltese (MT, left) and Tunisian Arabic (TU, right).

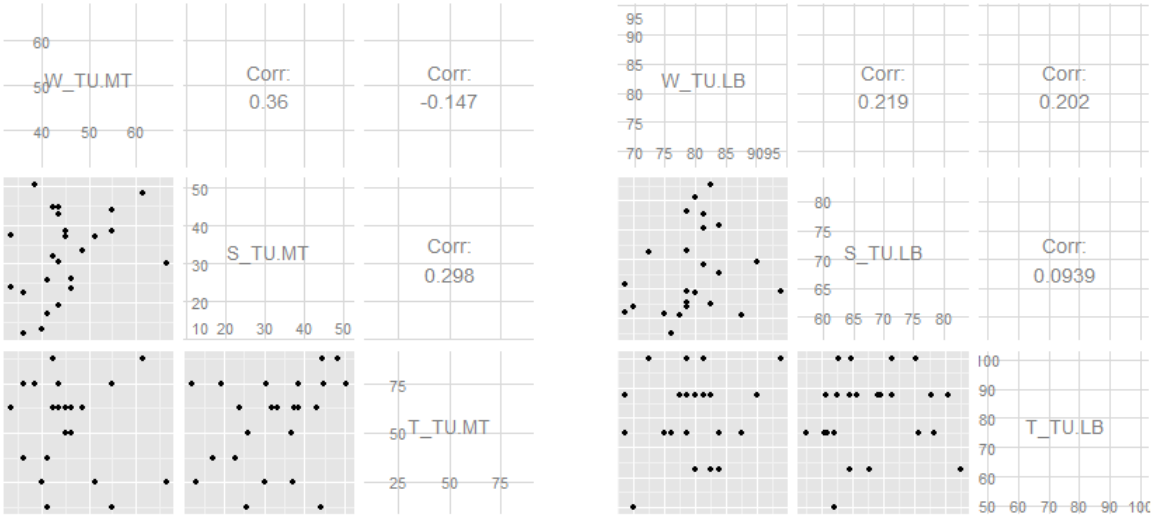

Figure 9: Scatterplot matrices of correlation data for all three test components (Word, Sentence and Text) administered to speakers of Tunisian Arabic (TU) in Maltese (MT, left) and Libyan Arabic (LB, right).

As the graphs show, the strength of correlations between word test results and sentence test results is generally low and for speakers of Maltese testing both Tunisian and Libyan Arabic, the relationship is negligible (-0.0475 and -0.174, respectively). Interestingly, there is a moderate negative relationship between the results of both tests for speakers of Libyan Arabic exposed to Maltese (-0.349), but a strong positive relationship for the same group of respondents testing Tunisian Arabic (0.431). For speakers of Tunisian Arabic, on the other hand, there is positive relationship between the results of the word test and the sentence test for both Libyan Arabic and Maltese – in fact, the relationship is stronger with Maltese (0.36) than with Libyan Arabic (0.219). This is surprising considering the generally high level of mutual intelligibility between Tunisian and Libyan Arabic and would lead to believe that a respondent's performance in either test is not a good predictor of their performance in the other. This in turn seems to support the conclusion reached by Tang and van Heuven (2009: 722)

that the word test itself is not sufficient to determine the level of mutual intelligibility – after all, the two tests constitute two significantly different tasks cognitively. Additionally, however, these data may shed further light on the consistency of results: in other words, the fact that for speakers of Tunisian Arabic there is a moderate positive relationship between both tests for both tested languages would confirm our findings that on the whole, speakers of Tunisian Arabic are better at understanding the other two varieties than vice versa.

As for the remaining combination of tests, the situation is comparable to that with word and sentence tests with the exception that the correlation between the results of the sentence test and those of text test for all country/variety combinations is predominantly positive. However, with the low reliability of the text test data, these figures do not mean much.

6 Determinants of intelligibility

6.1 Methodology

While a more thorough analysis of the factors influencing the mutual intelligibility of the three varieties studied would require a different test design, it is nevertheless possible to use the intelligibility data to roughly sketch out the linguistic variables involved, particularly the phonological ones. The word test data is especially suitable for this purpose, so we first categorized the items in the word test into cognates, secondary cognates (i.e. false friends) and non-cognates. This produced three lists of cognates with 77 cognates in the MT-TU pair, 85 cognates in the MT-LB pair and 106 cognates in the TU-LB pair. Then for the cognates in each pair, we established a list of features that set them apart (see the full list in table 9). These features are conceptualized as isoglosses split into two categories – those involving consonants and those involving vowels – and may not always be unidirectional (e.g. the presence of feature V1:vowel-schwa does not necessarily mean that where one variety always has a vowel, the other always has a schwa) and regular (such as vowel quantity or quality).

Table 9: Full list of isoglosses

Isoglosses	**Comments**
no change	
C1:intertendal-normal	Involves the pairs [d]/[d̪] and [t]/[t̪]
C2:devoiced-voiced	Word-final devoicing of stops in Maltese
C3:reflexes of qaf	Different developments of Classical Arabic [q]
C4:0-ghayn	Loss of [ʕ] in Maltese
C5:0-h	Loss of [h] in Maltese
C6:loss of gemination	
C7:additional morphology	Presence of absence of features such as fused definite article, infixed –yy- and feminine suffixes -a/ -t (e.g. W060C05)
C8:pharyngealized-normal	Loss of pharyngealization in stops in Tunisian Arabic and Maltese
C9:reflexes of gim	Different realizations of Classical Arabic [ǧ]
C10: merger of kh	Merger of [ḥ] and [ẖ] in Maltese
V1:vowel-schwa	Vowel reduction to [ə] or its complete elision
V2:quality	Changes in vowel quality, including imāla
V3:quantity	Changes in vowel quantity
V4:diphthong-vowel	Monophthongization of diphthongs and vice-versa
V5:0-epenthetic vowel	Epenthetic vowel [i] or [u] in Libyan Arabic

We added these to the respective entries to the CSV export of the results where responses for each country and target language combination consist of the respondent code, target language and the word code where each of the features was marked as 0 (absent) or 1 (present). Table 10 provides an overview of the structure of the CSV files created:

Table 10: Sample of data file

Respondent	Language	Code	Correct	MT_LB	C1:intertendal-normal	C2:devoiced-voiced	V1:vowel-schwa	V2:quality	V3:quantity
xmlanswer.pkg16.10.xml	MT	W064C05	correct	cognate	0	0	0	0	0
xmlanswer.pkg16.10.xml	MT	W106C08	incorrect	secondary cognate	0	0	0	0	0
xmlanswer.pkg16.10.xml	MT	W068C05	correct	cognate	0	0	0	1	0
xmlanswer.pkg16.10.xml	MT	W039C03	correct	non-cognate	0	0	0	0	0
xmlanswer.pkg16.10.xml	MT	W019C02	incorrect	non-cognate	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…

The data in the CSV files was then imported into R and used to analyze the relationship between the features and the scores. For that purpose, we opted to use a logistic mixed effects model (the R library *lme4*) with the score (the "Correct" column above) as the modelled binary dependent variable and the features as fixed effects. We selected this particular method because it allows us to include two random effects to account for the unavoidable unpredictability of human respondents in these scenarios. We added two such random variables, one per respondent and one per word (the "Code" column above), the latter because each respondent only tested one half of the words. We then used the R functions *scale()* to standardize the data and applied the following R code to analyze which of the fixed effects (i.e. linguistic features) influence the intelligibility of – in this particular case – Tunisian Arabic to speakers of Maltese:

```
mod.MT_TU.MIX <- glmer(Correct ~ no.change + C1.intertendal.normal + C2.devoiced.voiced + C3.reflexes.of.qaf + C4.0.ghayn + C5.0.h + C6.loss.of.gemination + C7.additional.morphology + C8.pharyngealized.normal + C9.reflexes.of.gim + C10.merger.of.kh + V1.vowel.schwa + V2.quality + V3.quantity + V4.diphthong.vowel + V5.0.epenthetic.vowel + (1|Respondent) + (1|Code), family="binomial", scaled_cogsMT_TU_lr)
```

Note that in this analysis, each feature is treated independently, i.e. we only consider the effect the feature has on its own and not in interaction with other features. Having performed extensive testing, we determined that this type of model is generally preferable to one where certain features interact, such as changes in vowel quality with the absence of pharyngealized consonants in Maltese. Nevertheless, there were some interactions that were found to be significant and we will highlight them as necessary.

We built six such basic full models, one per each speaker's language / tested language combination, with the purpose of determining which of the features have an effect on mutual intelligibility. As the primary form of diagnostics (in addition to the usual tests for normalcy etc.), we conducted an analysis of the predictive performance of each model using the R function *somers2()*

which determines the correlation between values predicted by the model and the actual data.[26] The function produces two measures on the 0-1 scale, the concordance index C and Somer's Dxy rank correlation. With the C index scores ranging from 0.89 to 0.94 and Dxy scores between 0.79 and 0.89, we deemed each model's fit good enough to provide a reasonably accurate picture of the variables involved, assuming a certain degree of caution in interpreting them is exercised. As the next step, we applied the R function *drop1()* to the full model to remove features one by one while assessing whether removing this feature has any effect on the fit of the model. We used the function's option *test="chisq"* to test whether each reduced model is different from the full model and thus to obtain a list of features that impact mutual intelligibility of the two varieties at a statistically significant level. In the analysis below, the p-values for the features are taken from the chi square test and we will analyze those features found to influence mutual intelligibility of the varieties involved in their context, i.e. in comparison with their total absolute and relative scores.

Before we proceed, a word of caution: the data and our analysis presented here are far from the complete picture: first, we only focus on the word intelligibility data as sentence-level analysis is much more complex, involving not only the phonology of words, but also differences suprasegmental features, morphology, syntax and phraseology and thus necessitating a different approach, one for which the methodology perhaps does not yet exist. Secondly, there are some indications that changes to the coda of a syllable or the end of the word are less likely to affect mutual intelligibility. Additionally, coding of the features was informed synchronically and thus some of the choices involved could very well be questioned. In this light, the conclusions outlined below should not be viewed as anything else than a rough estimate and an impetus to further targeted research into the linguistic factors influencing the mutual intelligibility of Arabic dialects.

6.2 Linguistic determinants of mutual intelligibility of Maltese and Tunisian Arabic

Table 11 provides a summary of features with statistically significant effect on intelligibility between Maltese and Tunisian Arabic.

Table 11: Isoglosses affecting mutual intelligibility of Maltese and Tunisian Arabic

Feature	**MT_TU**	**TU_MT**
	p-value	p-value
no.change		<0.1
C1.intertendal.normal		<0.05
C4.0.ghayn		<0.05
C5.0.h		<0.001
C7.additional.morphology	<0.05	
C10.merger.of.kh	<0.05	
V3.quantity	<0.01	<0.05
V4.diphthong.vowel	<0.001	<0.1
V2.quality:C4.0.ghayn	<0.01	

[26] http://www.inside-r.org/packages/cran/Hmisc/docs/somers2, accessed on 4 October 2015. For comments on the general procedure involving the testing and intepretation of mixed effect models, see http://glmm.wikidot.com/faq (accessed on 4 October 2015).

It is interesting to note that some of the most salient isoglosses seem to play no role at all, such as the typical Maltese devoicing of final stops or the issue of reflexes of Old Arabic *qāf* (glottal stop in Maltese, uvular stop [q] in our Tunisian Arabic recordings). One could speculate on the role of intra- and inter-dialectal variation here: there still are dialects of Maltese with a (usually voiceless) velar stop as the reflex of Old Arabic qāf.[27] It is therefore likely that the exposure to such variation makes it easier for speakers of Maltese to make sense of the Tunisian dialect which uses it. As for the other direction, the matters are a little more complicated: while the realization of Old Arabic *qāf* as a glottal stop is uncommon in either Tunis or Libya (Bahloul 2005: 252-253), it is a feature of other Arabic dialects, most prominently that of Cairo Egyptian Arabic (Fischer and Jastrow 1980: 208-209). This particular variety of Egyptian Arabic is a prestigious one and is often heard in popular music, movies and TV shows outside of Egypt. As such, its use of the glottal stop as a reflex of *qāf* is not entirely unfamiliar to speakers of Tunisian Arabic and may aid them in making sense of Maltese.

As for the features that do influence, it is surprising to see that the "no change" feature only has a significant effect for speakers of Tunisian Arabic exposed to Maltese. One would expect that the fact that both words sound the same would be strongly correlated with high scores for both speaker/listener pairs (as is the case for the TU_LB and LB_TU pairs), however, of the five items in this category (Maltese W008C01 [ḥūta], W017C02 [rās],W051C04 [twīl], W123C09 [barra] and W150C11 [šatt]), this is only true for the first two. For the third item, the scores are low in both directions (4 for MT_TU, 3 for TU_MT) and for the fourth and fifth item, speakers of Maltese were much better at understanding their Tunisian counterparts (with scores of 11 and 10, respectively) than the other way around (3 and 1). Why this is so we cannot answer yet, but one possibility is the phonetic detail in the realization of the vowel [a] in both varieties (cf. Gooskens, Van Bezooijen and Van Heuven 2015).[28]

Of the consonant changes, the loss of interdentals (isogloss C1), [ʕ] (isogloss C4) and [h] (isogloss C5) in Maltese poses a significant problem for speakers of Tunisian Arabic where all these consonant were retained. This does not apply to the converse direction where the non-phonemic status of the interdentals and [h] in Maltese does not pose any additional problems for its speakers in understanding Tunisian Arabic. On the surface, it appears that same would be true of [ʕ], however, changes in vowel quality which often accompany the loss of [ʕ], were found to interact with it at a statistically significant level. In other words, it is not the absence of [ʕ] on its own that makes understanding Maltese more difficult for speakers of Tunisian Arabic, but rather the combination of this development with changes in vowel quality. Interestingly, this does not work in the opposite direction where only changes in the morphological makeup of a word were found to impede the understanding of Tunisian Arabic to speakers of Maltese.

Technically, one more consonant change appears as significant and that is the merger of [ḥ] and [ḫ] in Maltese. The closer examination of the items involved reveals that this is most likely due to two outliers, word item W144C10 (MT [il-ḥarīfa], TU *ḫrīf*), with scores of 1 (for MT_TU) and 0 (TU_MT) and word item W110C08 (MT [mḥadda], TU *mḫadda*) with scores 12 (for MT_TU) and 1 (TU_MT). The former could be explained by an interplay of factors (additional morphology in Maltese, itself a significant factor), but it cannot be verified by the model and, more importantly, no such explanation can be offered for the latter. Since mutual intelligibility of the remaining three words does not seem to be affected by this isogloss and no other significant interactions of other features with this one were found, it appears that the merger of [ḥ] and [ḫ] as such doesn't affect the mutual

[27] Such as those of Cottonera and parts of Gozo (Aquilina 1961: 148).
[28] We are grateful to reviewer 1 for pointing out this possibility and the reference.

intelligibility of Maltese and Tunisian Arabic at all.

And finally, two vowel changes have a significant effect on the mutual intelligibility of Maltese and Tunisian Arabic: changes in vowel quantity and monophthongization of diphthongs (almost exclusively in the MT > TU direction). The latter is a clear-cut case, evident also from the comparison of results for Tunisian Arabic (which has a long vowel where Maltese has a diphthong) and Libyan Arabic (which, like Maltese, preserves the Old Arabic diphthong): W064C05 (MT [zeyt]) where for TU *zīt*, speakers of Maltese scored 0, but they scored 12 for LB *zeyt*, or W130C10 (MT [leyl]) with TU *līl* scoring 1 and LB *leyl* scoring 6. Changes in vowel quantity, although often accompanied by changes in vowel quality, do not interact with them – in other words, a change in vowel quantity on its own is enough to have an effect on intelligibility of a particular word.

6.3 Linguistic determinants of mutual intelligibility of Maltese and Libyan Arabic

Table 12 lists the statistically significant isoglosses that pose a challenge for the mutual intelligibility of Maltese and Libyan Arabic.

Table 12: Isoglosses affecting mutual intelligibility of Maltese and Libyan Arabic

Features	**MT_LB**	**LB_MT**
	p-value	p-value
C1.intertendal.normal	<0.05	<0.01
C5.0.h	<0.01	<0.01
C7.additional.morphology	<0.01	<0.01
C10.merger.of.kh	<0.05	
V3.quantity		<0.05

These results are similar to those for Maltese and Tunisian Arabic, especially when it comes to the role of the random effects and the loss of [h] in Maltese as well as the merger of [ḥ] and [ḫ] which is likewise explainable by the role of word item W144C10 as an outlier. The puzzling absence of the "no change" feature as a significant effect can also be encountered here, however, this time it may be explained by relative dearth of data as for this pair, the category only included three items. Once again, an important part of the real story is in what is absent: the realization of [ʕ] plays no role and neither do reflexes of Old Arabic *qāf* ([g] in Libyan Arabic). Additionally, unlike both Tunisian Arabic and Maltese, Libyan Arabic has retained pharyngealized consonants, yet this particular isogloss also plays no significant role in the mutual understanding between speakers of Maltese and Libyan Arabic. In light of this, it surprising to see that another major isogloss, that involving interdental fricatives and dental stops, does have a significant effect in both directions. This is most likely due to the nature of the phonological phenomena involved – stops vs. fricatives is a more salient contrast than the absence of a secondary articulation phenomenon such as pharyngealization – rather than the interaction with other features, such as changes in vowel quality which often accompany the loss of pharyngealization in Maltese (not found to have a significant effect). And finally, the additional morphological phenomena in Libyan Arabic (such as the diminutive infix [-eyy] in W052C04 LB *gṣeyyir* or W055C04 LB *irgeyyig*) and, conversely, their absence in Maltese constitute a significant obstacle to mutual intelligibility of the two varieties of Arabic.

In terms of vowels, the fact that these two varieties are similar in their retentions and innovations largely explains the absence of vowel features with significant effect on mutual

intelligibility between the two varieties. Only speakers of Libyan Arabic seem to have some difficulty comprehending words where the vowel quantity is different from what they are used to.

6.4 Linguistic determinants of mutual intelligibility of Tunisian and Libyan Arabic

Table 13 provides an overview of the features with statistically significant influence on mutual intelligibility of Tunisian and Libyan Arabic.

Table 13: Isoglosses affecting mutual intelligibility of Tunisian Arabic and Libyan Arabic

Features	**TU_LB**	**LB_TU**
	p-value	p-value
no.change	<0.05	<0.05
C7.additional.morphology	<0.01	<0.01
C8.pharyngealized.normal	<0.1	
V2.quality	<0.01	<0.05
V3.quantity		<0.01
V4.diphthong.vowel	<0.1	<0.001

Here caution in interpreting the model data is even more warranted than for the other two pairs: with the high intelligibility rates going in either direction (79.58% for TU_LB and 73.07% for LB_TU), linguistic features play a much smaller role. In other words, speakers of Tunisian and Libyan Arabic understand each other well enough that any failure in mutual intelligibility is much more likely to be caused by a random factor than by a particular isogloss. That being said, the table above paints a picture quite similar to that for the other two pairs of dialects: once again, the additional morphological phenomena found in Libyan Arabic present an obstacle, as does the monophthongization of diphthongs and changes in vowel quality and, for speakers of Libyan Arabic, in vowel quantity as well.

7 Conclusion

To roughly summarize our findings, we might observe that when it comes to the most basic everyday language as reflected in our data sets, speakers of Maltese are able to understand less than a third of what is being said to them in either Tunisian or Benghazi Libyan Arabic with Tunisian Arabic having a slightly higher chance to be understood in Malta than Libyan Arabic. In turn, Maltese is easier to understand for speakers of both mainstream Arabic dialects, with speakers of Tunisian doing slightly better than speakers of Libyan Arabic. In comparison, speakers of Libyan Arabic and speakers of Tunisian Arabic understand about two thirds of what is being said to them where, once again, speakers of Tunisian Arabic are slightly better at understanding their counterparts in Benghazi than the other way around. These results suggest that the anectodally supported idea of Tunisian Arabic's central position within Maġribī Arabic may not be wholly unfounded. Further research into the mutual intelligibility of North African varieties of Arabic as well as their relationship, especially using modern dialectometrical methods, is highly recommended.

In general methodological terms, this pilot has provided a wealth of experience and learning potential for any further iterations which will be able to avoid this study's major problems such as respondent selection or the exclusion of the listener's native variety from the test. As for test design, the study has confirmed the utility of both word and sentence tests, the latter preferably implemented

as a Bamford-Kowal-Bench Standard Sentence Test and a translation task. The inclusion of a text test in the standard mutual intelligibility testing toolkit, on the other hand, has not proven to be advantageous for our purposes and if implemented, greater care should be taken in the text and scoring scheme selection. The inclusion of some form of opinion testing – trivial to implement – should also be considered for follow-up studies, especially when outlier or minority varieties are involved. From a technical standpoint, the application LingTest developed for the purpose of the study has shown to be a tremendous asset in the field. More functionality, such as the ability to record answers (whether in audio or video form) and further improvement of its robustness and versatility would enhance its utility in various types of linguistic field research scenarios.

And finally, a rough analysis of the isoglosses affecting mutual intelligibility of the three varieties under study revealed some interesting insights, such as the lack of any role of reflexes of qāf or pharyngealized consonants and, conversely, the confounding effect of the lack of [h] in Maltese and of monophthongization of diphthongs where it occurs. In general, changes affecting vowels are more likely to affect mutual comprehension than those involving consonants. This is noteworthy not only because studies such as e.g. Gooskens et al. 2008 have found the opposite, but also for typological reasons: unlike the languages examined by Gooskens et al. (2008), the three varieties of Arabic we studied all exhibit root and pattern morphology. Recent studies have found that in both Maġribī Arabic (Schluter 2013) and Maltese (Ussishkin et al. 2015), the root plays a role in lexical acces. One would therefore assume that in cognates, the root would facilitate the intelligibility of the word and any changes to it would impede it. And this is in fact largely what we have found, particularly in the case of Maltese where the consonant system has undergone significant changes as compared to mainstream Maġribī Arabic, such as loss of [ʕ] (isogloss C4), loss of [h] (isogloss C5) and merger of [ḥ] and [ẖ] (isogloss C10). Both C4 and C5 have been found to negatively affect the intelligibility of Maltese words to speakers of Tunisian Arabic and same is true of C5 for speakers of Libyan Arabic. Speakers of Maltese had, in turn, trouble understanding words where the other two varieties of Arabic preserved the contrast between [ḥ] and [ẖ]. The absence of significant effects for the other consonantal isoglosses could be then explained either by interdialectal variation and the listeners' ability to deal with it, or as allomorphic variation which has been found not to impact root-facilitated lexical access (Boudelaa and Marslen-Wilson 2015: 976). However, the variation between the pairs [d]/[ḏ] and [t]/[ṯ] (isogloss C1) which we found to be significant in both Maltese and Tunisian Arabic (though only in one direction) and Maltese and Libyan Arabic complicates the picture. As for the role of the vowels, one possible explanation is that it is not actually the root, but rather the consonant and vowel pattern that plays the predominant role in lexical access and thus in mutual intelligibility (cf. Boudelaa and Marslen-Wilson 2015: 976). At present, however, there is very little data to support this hypothesis and only further studies into both mutual intelligibility and lexical processes can provide an answer.

Acknowledgments

The research underlying this paper was conducted at the Faculty of Arts of the Charles University and supported by the Grant Agency of Charles University in Prague, project no. GAUK 58313/2013.

The authors also gratefully acknowledge the assistance of Albert Gatt (University of Malta) offered during the field research and the preparation of the paper, as well as the feedback provided by Michael Cooperson (UCLA), Lameen Souag (CNRS) and the two anonymous reviewers incorporated into the final version of this paper. A preliminary version of this paper was presented at the *Methods in Dialectology XV* conference (University of Groningen, August 11th – August 15th, 2014) in the

special session *Bridging the gap between ELF and receptive multilingualism*, focusing on the technical aspects of the application LingTest. We gratefully acknowledge the helpful comments of the participants in that session. All errors remain ours.

References:

Abu-Haidar, Farida. 1992. Shifting boundaries: the effect of MSA on dialect convergence in Baghdad. In Ellen Broselow, Mushira Eid & John McCarthy (eds.), *Perspectives on Arabic Linguistics IV: Papers from the Fourth Annual Symposium on Arabic Linguistics*, 91-106. Amsterdam & Philadelphia: John Benjamins.

Anderson, Heidi. 2005. *Intelligibility Testing (RTT) between Mendankwe and Nkwen*. Dallas: SIL. http://www.sil.org/system/files/reapdata/11/85/53/118553469617548701614160758635455515473/silesr2005_002.pdf (accessed 4 October 2015)

Aquilina, Joseph. 1961. *Papers in Maltese Linguistics*. Malta: The University of Malta.

Bahloul, Maher. 2005. The Qaaf Across Arabic Dialects. In Elabbas Benmamoun (ed.), *Perspectives on Arabic Linguistics XIX: Papers from the Nineteenth Annual Symposium on Arabic Linguistics, Urbana, Illinois, April 2005*, 247-266. Amsterdam & Philadelphia: John Benjamins.

Benkato, Adam. 2014. The Arabic Dialect of Benghazi, Libya: Historical and Comparative Notes. *Zeitschrift für Arabische Linguistik* 59. 56-102.

Bent, Tess & Ann R. Bradlow. 2003. The interlanguage speech intelligibility benefit. *Journal of Acoustic Society of America* 114(3). 1600-1610.

Bouwer, Leoni. 2007. Intercomprehension and mutual intelligibility among southern Malagasy languages. *Language Matters: Studies in the Languages of Africa.* 38(2). 253-274.

Canty, A. and Ripley, B., 2014. *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-11.

Casad, Eugene H. 1974. *Dialect intelligibility testing*. Dallas: SIL. http://www.sil.org/system/files/reapdata/85/37/26/85372638740600013916377001436207275325/10386.pdf (accessed 4 October 2015)

Chaouachi, Kamal. 2014. *La culture orale commune à Malte et à la Tunisie*. Paris: L'Harmattan.

Corriente, Federico and Ángeles Vicente. 2008. *Manual de dialectológia neoárabe*. Zaragoza: Instituto de Estudios Islámicos y del Oriente Próximo.

Cohen, David, Dominique Caubet and Arlette Roth. 2000. Un questionnaire différentiel pour la dialectologie arabe. *Oriente Moderno* 19(1). 1-23.

Davison, A. C. & D. V. Hinkley. 1997. *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.

Delsing, Lars-Olof and Katarina Lundin-Åkesson. 2005. *Håller språket ihop Norden?* Kopenhagen: Nordiska ministerrådet.

Fabri, Ray. 2010. Maltese. *Revue Belge de Philologie et d'Histoire* 88(3). 791-816.

Fischer, Wolfdietrich and Otto Jastrow. 1980. *Handbuch der arabischen Dialekte*. Wiesbaden: Harrassowitz.

Gooskens, Charlotte. 2013. Experimental methods for measuring intelligibility of closely related language varieties. In Robert Bayley, Richard Cameron & Ceil Lucas (eds.), *The Oxford Handbook of Sociolinguistics*, 195-213. Oxford: Oxford University Press.

Gooskens, Charlotte, Wilbert Heeringa & Karin Beijering. 2008. Phonetic and lexical predictors of intelligibility. *International Journal of Humanities and Arts Computing* 2(1-2). 63-81.

Gooskens, Charlotte, Renée van Bezooijen & Vincent J van Heuven. 2015. Mutual intelligibility of Dutch-German cognates by children: The devil is in the detail. *Linguistics* 53(2). 255-283.

Hammett, Sandra. 2012. *Maltese within the maghrebi dialectological framework.* Exeter: University of Exeter MA thesis.

Kalikow, D. N., K. N. Stevens & L. L. Elliott. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America* 61(5). 1337-1351.

Owens, Jonathan. 2010. What is a Language? Review of Bernard Comrie, Ray Fabri, Elizabeth Hume, Manwel Mifsud, Thomas Stolz & Martine Vanhove (eds.), Introducing Maltese Linguistics. Selected papers from the 1st International Conference on Maltese Linguistics, Bremen, 18–20 October 2007, 2009. xi, 422 pages. Studies in Language Companion Series 113. Amsterdam. *Journal of Language Contact – VARIA 3*. 103-118.

Panetta, Ester. 1943. *L'arabo parlato a Bengasi. Testi con traduzione e note. Grammatica.* Rome: Ministero dell'Africa italiana.

Tang, Chaoju & Vincent J. van Heuven. 2009. Mutual intelligibility of Chinese dialects experimentally tested. *Lingua* 119. 709-732.

Ryding, Karin C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge: Cambridge University Press.

Schluter, Kevin. 2013. *Hearing words without structure: Subliminal speech priming and the organization of the Moroccan Arabic lexicon.*Tuscon: University of Arizona PhD thesis.

Vella, Joseph. 2004. *Learn Maltese. Why Not?* Valetta: Valetta Publishing.

Ussishkin, Adam, Colin Reimer Dawson, Andrew Wedel & Kevin Schluter. 2015. Auditory masked priming in Maltese spoken word recognition. *Language, Cognition and Neuroscience* 30(9). 1096-1115.

Appendix A: Word test data

Word code	Word (English)	Category	Maltese (standard orthography and transcription)	Tunisian	Benghazi
W001C01	dog	Animals	kelb [kelp]	kɛlb	kelb
W002C01	horse	Animals	żiemel [ziemel]	ḥsān	ḥṣān
W003C01	rabbit	Animals	fenek [fenek]	ʔarnɛb	arnab
W004C01	cat	Animals	qattus [ʔattūs]	qattūs	gaṭṭūs
W005C01	mouse	Animals	ġurdien [ǧurdin]	fār	fār
W006C01	bird	Animals	għasfur [asfūr]	ʕasfūr	ʕaṣfūr
W007C01	pig	Animals	ħanżir [ḥanzīr]	ḥallūf	ḥəllūf
W008C01	fish	Animals	ħuta [ḥūta]	ḥūta	ḥūṭa
W009C01	spider	Animals	brimba [brimba]	rtīla	ʕankabūt
W010C01	fly	Animals	dubbiena [dubbiena]	ḏəbbēna	ḏəbbāna
W011C01	fox	Animals	volpi [volpi]	ṯaʕləb	ṯaʕlab
W012C01	wolf	Animals	lupu [lupu]	ḏīb	ḏīb
W013C01	sheep	Animals	nagħġa [nāǧa]	ʕallūš	ḥowlī
W014C01	donkey	Animals	ħmar [ḥmār]	bhīm	ḥumāṛ
W015C02	body	Body parts	ġisem [ǧisem]	bdɛn	žisim
W016C02	hand	Body parts	id [īt]	yɛdd	yad
W017C02	head	Body parts	ras [rās]	rās	ṛās
W018C02	leg	Body parts	riġel [riǧel]	sēq	krāʕ
W019C02	foot	Body parts	sieq [sīʔ]	sēq	krāʕ[29]
W020C02	hair	Body parts	xagħar [šār]	šʕār	šəʕar
W021C02	face	Body parts	wiċċ [wičč]	wužh	wəžih
W022C02	eye	Body parts	għajn [ayn]	ʕīn	ʕeyn
W023C02	blood	Body parts	demm [dem]	dɛmm	dəmm
W024C02	ear	Body parts	widna [widna]	wuḏɛn	wuḏin
W025C02	neck	Body parts	għonq [onʔ]	raqba	ruguba
W026C02	tooth	Body parts	snien [snīn]	sənna	sinn
W027C02	finger	Body parts	saba' [saba]	sboʕ	ṣobəʕ
W028C02	mouth	Body parts	fomm [fom]	fumm	fəmm
W029C02	heart	Body parts	qalb [ʔalp]	qalb	gəlib
W030C03	shirt	Clothing and jewelry	qmis [ʔmīs]	sūrīya	ṣūriya
W031C03	pants (trousers)	Clothing and jewelry	qalziet [ʔalcīt]	sɛrwēl	sirwāl
W032C03	dress	Clothing and jewelry	libsa [lipsa]	rūba	gufṭān
W033C03	shoes	Clothing and jewelry	żarbun [zarbūn]	sabbāt	kindara
W034C03	belt	Clothing and jewelry	ċinturin [činturīn]	sɛbta	seyr
W035C03	ring	Clothing and jewelry	ċurkett [čurkett]	ḫātəm	ḫātim
W036C03	earring	Clothing and jewelry	misluta [mislūta]	ballūta	dandūla
W037C03	scarf	Clothing and jewelry	xalpa [šalpa]	kāškōl	šāl

[29] Same translation for items W018C02 and W019C02 was provided for Tunisian and Libyan Arabic.

W038C03	cloak	Clothing and jewelry	mantar [mantār]	barnūs	kābūṭ
W039C03	pocket	Clothing and jewelry	but [būt]	žīb	žeyb
W040C03	gold	Clothing and jewelry	deheb [dēp]	ḏhɛb	ḏahab
W041C03	silver	Clothing and jewelry	fidda [fidda]	fəḏḏa	fuḏḏa
W042C03	wear	Clothing and jewelry	jilbes [yilbes]	ḥwɛ̄yəž[30]	yelbes
W043C04	white	Colors, shapes and properties	abjad [abyat]	abyəḏ	abyaḏ
W044C04	black	Colors, shapes and properties	iswed [iswet]	ɛkḥɛl	aswud
W045C04	green	Colors, shapes and properties	aħdar [aḥdar]	aḫḏər	aḫḏar
W046C04	red	Colors, shapes and properties	aħmar [aḥmar]	aḥmər	aḥmar
W047C04	yellow	Colors, shapes and properties	isfar [isfar]	asfər	aṣfar
W048C04	brown	Colors, shapes and properties	kannella [kannella]	šoklāti	gahwī
W049C04	dark	Colors, shapes and properties	skur [skūr]	ġāmaq	azrag
W050C04	blue	Colors, shapes and properties	blu [blu]	azraq	azrag[31]
W051C04	long	Colors, shapes and properties	twil [twīl]	twīl	ṭəwīl
W052C04	short	Colors, shapes and properties	qasir [ʔasīr]	qsīr	gṣeyyir
W053C04	round	Colors, shapes and properties	tond [tont]	mdawwər	mdowwər
W054C04	narrow	Colors, shapes and properties	dejjaq [deyyaʔ]	ḏeyyəq	ḏeyyig
W055C04	thin	Colors, shapes and properties	rqiq [rʔīʔ]	žwɛyyəd	irgeyyig
W056C04	wide	Colors, shapes and properties	wiesa' [wīsaʔ]	wɛ̄faʕ	ʕarīḏ
W057C04	heavy	Colors, shapes and properties	tqil [tʔīl]	rzīn	ṯigīl
W058C04	light	Colors, shapes and properties	ħafif [ḥafīf]	fɛ̄taḥ	ḫəfīf
W059C05	bread	Eating and drinking	ħobz [ḥops]	ḫubz	ḫubza
W060C05	water	Eating and drinking	ilma [ilma]	mɛ̄	mməyya
W061C05	vegetables	Eating and drinking	ħaxix [ḥašīš]	ḫɔḏra	ḫuḏra
W062C05	meat	Eating and drinking	laħam [laḥam]	lḥam	ləḥam
W063C05	fruits	Eating and drinking	frott [frott]	ġalla	fākiha
W064C05	oil	Eating and drinking	żejt [zeyt]	zīt	zeyt
W065C05	cheese	Eating and drinking	ġobon [ǧobon]	žbən	žibna
W066C05	salt	Eating and drinking	melħ [melḥ]	mɛlḥ	miliḥ
W067C05	grapes	Eating and drinking	għeneb [ēnep]	ʕnəb	ʕinab
W068C05	wine	Eating and drinking	inbid [inbīt]	šrāb	nəbīt
W069C05	he drinks	Eating and drinking	jixrob [yišrop]	yušrob	yešrəb

[30] The Tunisian translation actually reads "clothes". This had no effect on the scores and the term was excluded from modeling.
[31] Same translation for both W049C04 and W050C04 was provided for Tunisian and Libyan Arabic.

W070C05	he eats	Eating and drinking	jiekol [yīkol]	yɛ̄kəl	yākəl
W071C05	egg	Eating and drinking	bajda [bayda]	ʕḏəm	daḥī
W072C06	angry	Emotions	irrabjat [irrabyāt]	mətġaššəš	ragīla
W073C06	sad	Emotions	imdejjaq [imdeyyaʔ]	ḥzīn	zaʕlān
W074C06	happy	Emotions	ferħan [ferḥān]	farḥān	farḥān
W075C06	tired	Emotions	għajjien [ayyīn]	tɛ̄ʕəb	taʕbān
W076C06	love	Emotions	imħabba [imḥabba]	ḥobb	ḥubb
W077C06	fear	Emotions	biża' [biza]	ḫūf	ḫowf
W078C06	patient	Emotions	paċenzjuż [pačencyūs]	sābər	ṣəbūr
W079C06	ashamed	Emotions	mistħi [mistḥi]	ḥāšəm	mitḥaššim
W080C06	crazy	Emotions	miġnun [miǧnūn]	mɛhbūl	mažnūn
W081C06	hope	Emotions	tama [tāma]	āmal	mutaʔammil
W082C06	envy	Emotions	għira [eyra]	ġīra	ġayūr
W083C06	proud	Emotions	kburi [gbūri]	farḥān	fəḫūr
W084C06	he worries	Emotions	jinkwieta [yinkwīta]	mətqallaq	mašġūl
W085C06	he loves	Emotions	jħobb [yḥopp]	iḥɛbb	īḥebb
W086C07	human being	Family and other people	bniedem [bnīdem]	ʕabd	insān
W087C07	family	Family and other people	familja [familya]	ʕīla	ʕāʔila
W088C07	people	Family and other people	nies [nīs]	ʕbɛ̄d	nās
W089C07	mother	Family and other people	ommi [ommi]	ʔumm	umm
W090C07	father	Family and other people	missier [missīr]	bu	bāṯ
W091C07	brother	Family and other people	ħija [ḥiya]	ḫu	ḫū
W092C07	sister	Family and other people	oħti [oḥti]	oḫt	əḫit
W093C07	bride	Family and other people	għarusa [arūsa]	ʕarūsa	ʕarūs
W094C07	cousin	Family and other people	kuġin [kuǧīn]	wuld ʕamm	qarīb
W095C07	aunt	Family and other people	zija [cīya]	ʕamma	ʕamma
W096C07	uncle	Family and other people	ziju [cīyu]	ʕamm	ʕamm
W097C07	married	Family and other people	miżżewweġ [mizzewweč]	mʕarrəs	mizowwəž
W098C07	woman, wife	Family and other people	mara [mara]	mart	wəliya
W099C07	man, husband	Family and other people	raġel [rāǧel]	rāžəl	rāžul
W100C07	baby	Family and other people	tarbija [tarbīya]	sġīr	ʕāyl
W101C07	was born	Family and other people	twieled [twīlet]	tūləd	wətəled
W102C08	door	In the house	bieb [bīp]	bɛ̄b	bāb
W103C08	window	In the house	tieqa [tīʔa]	šubbɛ̄k	rōšen
W104C08	roof	In the house	saqaf [saʔaf]	sqaf	sṭāḥ
W105C08	floor	In the house	qiegħ [ʔīḫ]	qāʕa	arḍ
W106C08	room	In the house	kamra [kamra]	bīt	dāṛ
W107C08	table	In the house	mejda [meyda]	tāwla	ṭāwla
W108C08	chair	In the house	siġġu [siǧǧu]	korsi	kirsī
W109C08	bed	In the house	sodda [sodda]	farš	sərīr
W110C08	pillow	In the house	mħadda [mḥadda]	mḫadda	məxədda

W111C08	carpet	In the house	tapit [tapīt]	zarbēya	farša
W112C08	stairs, staircase	In the house	taraġ [tarač]	drūž	drūž
W113C08	key	In the house	muftieħ [muftīḫ]	məftēḥ	miftāḥ
W114C09	here	Orientation in space	hawn [awn]	hūni	hena
W115C09	there	Orientation in space	hemm [hemm]	ġādi	ġādī
W116C09	left	Orientation in space	lemin [lemīn]	īsār	yeṣār
W117C09	right	Orientation in space	xellug [šelluk] [32]	īmīn	yemīn
W118C09	above	Orientation in space	fuq [fuʔ]	fūq	fowg
W119C09	below	Orientation in space	isfel [isfel]	taḥt	taḥit
W120C09	in front of	Orientation in space	quddiem [ʔuddīm]	qoddēm	giddām
W121C09	behind	Orientation in space	wara [wara]	wurā	wərā
W122C09	inside	Orientation in space	ġewwa [ǧewwa]	fi wost	žowwa
W123C09	outside	Orientation in space	barra [barra]	l-barra	bərra
W124C09	north	Orientation in space	tramuntana [tramuntāna]	šmēl	šamāl
W125C09	east	Orientation in space	lvant [lvant]	žanūb[33]	šarg
W126C09	west	Orientation in space	punent [punent]	ġarb	ġarəb
W127C10	time	Time	ħin [ḥīn]	waqt	wagit
W128C10	day	Time	jum [yūm]	nhār	yōm
W129C10	month	Time	xahar [šār]	šhər	šəhar
W130C10	night	Time	lejl [leyl]	līl	leyl
W131C10	daytime	Time	binhar [binār]	nhār	yōm
W132C10	year	Time	sena [sena]	ʕām	sana
W133C10	today	Time	illum [illum]	l-yūm	el-yūm
W134C10	yesterday	Time	ilbieraħ [ilbīraḥ]	l-bɛ̄rɛḥ	ams
W135C10	tomorrow	Time	għada [āda]	ġodwa	bukra
W136C10	in the morning	Time	filgħodu [filōdu]	f əs-sbɛ̄ḥ	fi l-ṣobəḥ
W137C10	in the evening	Time	filgħaxija [filašīya]	f əl-līl	fi l-ʕašiya
W138C10	now	Time	issa [issa]	tawwa	towwa
W139C10	always	Time	dejjem [deyyem]	dīma	dīma
W140C10	never	Time	qatt [ʔatt]	žɛ̄mla	māʕomraš
W141C10	summer	Time	is-sajf [is-sayf]	sīf	ṣeyf
W142C10	winter	Time	ix-xitwa [iš-šitwa]	štē	šitā
W143C10	spring	Time	ir-rebbiegħa [ir-rebīa]	rbīʕ	rəbīʕ
W144C10	autumn	Time	il-ħariſa [il-ḥarīfa]	ḫrīf	ḫərīf
W145C10	hour	Time	siegħa [sīa]	sɛ̄ʕa	sāʕa
W146C11	earth, ground	World around us	art [art]	aṛḏ	aṛḏ
W147C11	world	World around us	dinja [dinya]	dənya	ʕālam
W148C11	sky	World around us	sema [sema]	smɛ̄	səmā
W149C11	sea	World around us	baħar [baḥar]	bḥar	bəḥar
W150C11	beach	World around us	xatt [šatt]	šatt	šəṭṭ

[32] Items W116C09 and W117C09 were swapped in Maltese. This had no effect on the scores and the appropriate correction was made for the modeling.
[33] The Tunisian translation actually reads "south". This had no effect on the scores and the term was excluded from modeling.

W151C11	hill	World around us	għolja [ōlya]	žbəl	žibel
W152C11	mountain	World around us	muntanja [muntanya]	žbəl	žibel[34]
W153C11	village	World around us	raħal [raḥal]	qarya	qərya
W154C11	city	World around us	belt [belt]	mdīna	medīna
W155C11	street, road	World around us	triq [triʔ]	šɛ̄raʕ	šāriʕ
W156C11	square	World around us	pjazza [pyaca]	batḥa	sāḥa
W157C11	field	World around us	għalqa [ālʔa]	arḏ	məzraʕa
W158C11	island	World around us	gżira [gzīra]	žazīra	žəzīra
W159C11	sun	World around us	xemx [šemš]	šəms	šams
W160C11	moon	World around us	qamar [ʔamar]	gamra	gəmar

Appendix B: Sentence test data

Sentence code	Sentence English	Sentence Maltese
S001C01	Wash your hands with soap.	Aħsel idejk bis-sapun.
S002C01	My brother went to England to find work.	Ħija mar l-Ingilterra biex ifittex xogħol.
S003C01	My son has a small dog.	Ibni għandu kelb żgħir.
S004C01	There is no rose without thorns.	M'hemmx warda mingħajr xewk.
S005C01	He found all the doors locked.	Sab il-bibien magħluqin kollha.
S006C01	His face was red with anger.	Wiċċu kien aħmar bil-għadab.
S007C02	How many children do you have?	Kemm għandek tfal?
S008C02	The bride is waiting in front of the church.	L-għarusa qed tistenna quddiem il-knisja.
S009C02	The young people are dancing without clothes.	Iż-żgħażagħ jiżfnu mingħajr ħwejjeġ.
S010C02	Why don't you come with us?	Għax ma tiġix magħna?
S011C02	They lived there for four years.	Huma damu jgħixu hemm erba' snin.
S012C02	They stole her bag.	Serqulha l-basket tagħha.
S013C02	Children are listening to the teacher.	It-tfal qed jisimgħu lill-għalliem.
S014C02	This one costs forty-seven.	Dan jiswa seba' u erbgħin.
S015C03	The doctor comes to see you at home.	It-tabib jiġi jarak f'darek.
S016C03	The boy broke his leg.	It-tifel kiser siequ.
S017C03	The men brought a long ladder.	L-irġiel ġabu sellum twil.
S018C03	There was a lot of trash on the beach.	Fix-xatt kien hemm ħafna żibel.
S019C03	The sick recover from their illness.	Il-morda jfiqu mill-mard tagħhom.
S020C03	The tree casts a shadow on the building	Is-siġra titfa' dell fuq il-bini.
S021C03	Every time they see him, they laugh at him	Kull meta jarawh, jidħku bih.
S022C03	First, clean the potatoes.	Qabel kollox naddaf il-patata.
S023C04	The cat sleeps in the middle of the road.	Il-qattus rieqed f'nofs it-triq.
S024C04	In summer, many festivals take place.	Fis-sajf isiru ħafna festi.
S025C04	Let's go before the night arrives.	Ejja nimxu qabel jidlam.
S026C04	The fishermen take the fish to the market.	Is-sajjieda jieħdu l-ħut is-suq.

[34] Same translation for both W151C11 and W152C11 was provided for Tunisian and Libyan Arabic.

S027C04	People fast during Lent/Ramadan.	In-nies isumu matul ir-Randan.
S028C04	Look how pretty it is!	Ara kemm hi sabiħa!
S029C05	The two women entered the shop.	Iż-żewġ nisa daħlu fil-ħanut.
S030C05	The birds are dying from heat.	L-għasafar imutu bis-sħana.
S031C05	I've never heard this story before.	Din il-ħrafa qatt ma smajtha qabel.
S032C05	They came to give him the last goodbye.	Ġew biex jagħtuh l-aħħar tislima.
S033C05	The girls are eating bread with oil.	Ix-xbejbiet jieklu l-ħobż biż-żejt.
S034C05	What news have you brought us?	X'aħbar ġibtilna?
S035C05	There is black smoke coming from the window.	Mit-tieqa ħiereġ duħħan iswed.
S036C05	In the beginning, God created heaven and earth.	Fil-bidu Alla ħalaq is-smewwiet u l-art.
S037C06	Everyone loves his mother.	Kulħadd iħobb lil ommu.
S038C06	Do not add more salt!	Iżżidx aktar melħ!
S039C06	A leaf flies on the wind.	Werqa ttir mar-riħ.
S040C06	I feel strong pain in my chest.	Inħoss uġigħ qawwi f'sidri.
S041C06	He was sitting with his back against a wall.	Kien bilqiegħda b'dahru mal-ħajt.
S042C06	Do you (sg.) remember this thing?	Tiftakarha din il-ħaġa?
S043C06	She looked at me with a smile.	Ħarset lejja bi tbissima.
S044C06	He appears to be lost in his thoughts.	Jidher mitluf fi ħsibijietu.
S045C07	They began standing up, one after another	Bdew iqumu wieħed wara l-ieħor.
S046C07	The knife is on the table.	Is-sikkina qiegħda fuq il-mejda.
S047C07	The girl has a new book.	It-tfajla għandha ktieb ġdid.
S048C07	Today ends time of Lent/Ramadan.	Illum tmiem żmien ir-Randan.
S049C07	Some workers came out when they heard what happened.	Xi ħaddiema ħarġu meta semgħu x'ġara.
S050C07	The foreigner speaks to us in our language.	Il-barrani jkellimna b'ilsienna.
S051C07	The horse is walking and the old man is riding.	Iż-żiemel miexi u x-xiħ riekeb.
S052C07	Strong rain fell yesterday.	Ilbieraħ niżlet xita qawwija.
S053C08	I opened the door with a key.	Ftaħt il-bieb biċ-ċavetta.
S054C08	Our neighbors bought a new car.	Il-ġirien xtraw karrozza ġdida.
S055C08	We have need for more money.	Għandna bżonn aktar flus.
S056C08	Thanks to you that you came.	Grazzi lilek talli ġejt.
S057C08	Everything is ready to begin the game.	Kollox lest biex tibda l-logħba.
S058C08	Is it true or not?	Dan veru jew le?
S059C08	Every time I ask him, he doesn't reply to me.	Kull darba li nistaqsih, ma jirrispondinix.
S060C08	This may not be used.	Din ma tistax tintuża.

Sentence code	**Sentence Tunisian Arabic**	**Sentence Libyan Arabic**
S001C01	aġsəl īdīk b əs-sābūn	əġsil īdeyk biṣṣābūn
S002C01	ḫūya mše l anglətɛrra bɛ̄š yalqa ḫədma	ḫūya ʕədda li briṭānya īdowwər ʕali šoġəl
S003C01	wuldi ʕandu kalb sġīr	wulidī ʕinda kelb ṣəġeyyir
S004C01	ma fammɛ̄š warda blɛ̄š šūk	māfīšī warid bilā wərəg
S005C01	lqa l-bībɛ̄n kull msakkrīn	ligā l-bībān killhin msəkkərāt

S006C01	wəžhu aḥmar b əl-ġušš	wəža kān ḥəmər mi l-ġəḏəb
S007C02	qaddēš ʕandək s-sġār	kam ʕindak ʕeyl
S008C02	l-ʕarūsa təstanna quddēm əl-knīsēya	el-ʕarūs itrāžī giddām el-kinīsa
S009C02	š-šɛbēb yəštḥu blēš ḥwēyž	eš-šəbāb yirigṣu minġeyr dibeš
S010C02	ʕalēš ma žītš mʕāna	kannak mātži maʕāna
S011C02	ʕandu ġādi arbʕa snīn	lhum ʕāyišīn ġādī arbaʕ sinīn
S012C02	sərqu lha s-sāk mtaḥḥa	sirgū šənṭitha
S013C02	s-sġār yesmʕu f əl-muʕalləm	l-aṭfāl yesməʕū fi kəlām el-ustāḏ
S014C02	hēḏēya sūmu sabʕa w arbaʕīn	haḏi ḥəgghā sabaʕ u arbaʕīn
S015C03	t-tbīb žēy bēš išūfək f əd-dār	ed-doktor ḥaīžī išūfak fi l-ḥowš
S016C03	t-tfəl kassər sēqu	el-ʕāyl kəssər krāʕa
S017C03	r-rāžəl šre sallūm twīl	er-rāžul žāb sellūm ṭawīl
S018C03	kēn famma barša zɛbla f əš-šatt	kān fī wsəḥ wāžid ʕa l-šəṭṭ
S019C03	əl-morḍa qāʕdīn yebrāw m əl-marḏ mtaḥḥum	l-imruḍa bidow iṣəḥḥū mi l-mərəḏ imtāḥḥum
S020C03	əš-šɛžra mḍɛlla ʕ al-bānya	eḏ-ḍull imtāʕ eš-šužura ʕa l-mabnā
S021C03	wīn nšūfu naḏḥak aʕlīh	kull mā īšūfū yaḍaḥkū ʕaley
S022C03	awwəl ḥāža naḏḏəf əl-bātāta	fi l-awwəl nəḍḏəf l-bəṭāṭa
S023C04	l-qattūs rēqəd f wost ət-trīq	el-gaṭṭūsa rāgda fi noṣṣ eš-šāriʕ
S024C04	f sīf famma barša mahrajēnēt	fi ṣ-ṣeyf fi ḥefalāt wāžid
S025C04	hɛyya nəmšīw qbəl ma itīḥ əl-līl	hayya nʕəddū gəbəl mā tḍəlləm
S026C04	s-sayyēda hēzzīn l-ḥūt l əs-sūq	el-ḥəwāta yāḫəðu fi l-ḥūt li s-sūg
S027C04	n-nēs isūmu fi rumḍān	en-nās itṣīm fi ramaḍān
S028C04	šūf qaddēš məzyēna	baḥḥit keyf simḥa
S029C05	zūz nse daḫlu l əl-ḥānūt	l-wəlīteyn ḫəššen li d-dukkān
S030C05	l-ʕasāfər qāʕdīn imūtu m əs-sḫēna	el-ʕaṣāfīr īmūten mi l-ḥamu
S031C05	ʕomri ma smaʕt la-ḥkēya hēdi qbəl	māʕomrīš səmaʕt el-qiṣṣa haḏi min gəbəl
S032C05	žēw bēš iwaddʕūh	žow beyš īgūlūla maʕa səlāma li l-āḫir mərra
S033C05	l-bnēt qāʕdīn yēklu f əl-ḫobəz b əz-zīt	el-bənāt yākəlan fi l-ḫubza bi z-zeyt
S034C05	šnuwwa l-ḫbār lli žəbthum əlna	šin el-aḫbār lī žibthin linna
S035C05	famma duḫḫān akḥəl qāʕd iḫrəž m əš-šubbēk	fī dəḫḫān iswud ṭāləʕ mi l-rōšen
S036C05	m əl-awwəl rəbbi ḫləq sme w ul-arḏ	fi l-awwəl rəbbī ḫələg əs-simmā w əl-arḏ
S037C06	n-nēs əl-kull iḥabbu ummēthum	kill wāḥid īḥebb umma
S038C06	ma tzīdš melḥ	mātzīdš miliḥ akṯar
S039C06	warqa tāyra f ər-rīḥ	wurga ṭṭīr fi l-howā
S040C06	nḥəss fi barša wužīʕa fi sədri	nḥiss fi wəžəʕ gowwī fī ṣədrī
S041C06	kān qāʕd u ḏahru mʕa ḥīt sġīr	kān mgaʕmiz w ḍəhara ʕa l-sās
S042C06	tfakkər š-šēy hēḏa	təḍəkkər haḏi
S043C06	ḫazrət li u hīya tətbassəm	baḥḥətat fiya bibtisāmha
S044C06	ḏāhər fīh ḏāyəʕ fīha	ībān inna howa rāyiḥ fī afkāra
S045C07	bdēw iwāqfu b əl-wēḥɛd b əl-wēḥɛd	bidow īṣəbbū wāḥid bi l-wāḥid
S046C07	s-səkkīna fūq ət-tāwla	el-mūs ʕa ṭ-ṭāwla
S047C07	lə-bnēya ʕandha karrāsa ždīda	el-bint ʕandha kitāb žədīd
S048C07	l-yūm yūfa rumḍān	el-yūm yikmil wəgit ramaḍān

S049C07	l-ḫaddēma žēw ki samʕu bəlli sār	wāḥdīn yištəġəlū ṭəlʕū baʕd mā simʕū šin ṣār
S050C07	l-barrāni yaḥki mʕēna b luġətna	el-ažnabī yidwīna bī luġitna
S051C07	lə-ḥsān yəmši u rāžəl kbīr rēkəb aʕlīh	l-əḥṣān yimšī wa r-rāžul l-kibīr īsūg fīh
S052C07	šte qwīya sēbət əl-bērɛḥ	məṭərit bil-guwwa āms
S053C08	ḫallīt əl-bēb b əl-məftēḥ	fitaḥt el-bāb bi-miftāḥ
S054C08	žīrēnna šrēw karhba ždīda	žārna šərā sayyāra žədīda
S055C08	ḫāšətna b akṯər flūs	nibbū filūs uḫra
S056C08	yaʕtīk saḥḥa ki žīt	šukrān lak ʕala žeyytak
S057C08	kull šēy ḫāḏər bēš tabda l-laʕba	kull ḫāža wātiya beyš nebdū el-geym
S058C08	b əl-mən žədd wa lɛ	ṣaḥ wəla lā
S059C08	kull marra nasʔalu ma ižāwəbnīš	kull mā nesʔela māirəddš ʕaleya
S060C08	ma lāzəmš yistaʕməl	haḏi rāhī mā tinišġəlš

Appendix C: Evaluation instructions for the sentence test

Assign the following categories to the answers in the sentence test:

Not answered: No answer. (Analyzed as "incorrect")
Incorrect: Incorrect answer. (Analyzed as "incorrect")
25% correct: Not the correct lexical item, but identified root or stem or gave a false friend. (Analyzed as "incorrect")
50% correct: Partial synonym used or something is missing, e.g. when Maltese 'xiħ' is translated as 'راجل كبير' and only 'raġel' is given by the respondent. (Analyzed as "partially correct")
75% correct: Partial synonym or equivalent used, correct lexical item, incorrect morphology. (Analyzed as "correct")
100% correct: Full synonym or correct lexical item used, correct morphology. (Analyzed as "correct")

Remarks:
1. If the answer is "x" or "ו", mark all items as "Not answered".
2. If only a partial answer is provided, it might not be easy to determine which items were not answered. In such case, do your best to guess as I did above. It doesn't really matter for the purpose of final analysis (both "not answered" and "incorrect" will be analyzed as "incorrect"), but we want to get a realistic picture of situations where the respondent doesn't have a clue (i.e. "not answered").
3. Since translations can differ in the lexical choice, evaluate based on the translation, not the original. For example, S045C07 MT has "wieħed wara l-ieħor", but both LB and TU have "واحد واحد". If the MT respondent gives "wieħed wieħed", evaluate as 100% correct.
Another example: S060C08 MT has "Din ma tistax tintuża", but TU only has "ma lāzəmš yistaʕməl". A keyword DEMONSTRATIVE has been added to the test package to enable you to correctly evaluate the answer should a TU respondent be able to catch and translate the initial "din".