Towards a corpus of spoken Maltese: Korpus tal-Malti Mitkellem - KMM

Alexandra Vella, Sarah Agius, Aiden Williams, Claudia Borg
University of Malta
alexandra.vella@um.edu.mt, agius.sarah@gmail.com, aiden.williams@um.edu.mt, claudia.borg@um.edu.mt

Abstract

This paper presents the rationale for a "dedicated" corpus of spoken Maltese, *Korpus tal-Malti Mitkellem*, *KMM*, 'Corpus of Spoken Maltese', based on the concept of a gold-standard Core collection. The Core collection is designed to cater to as wide a variety of user needs as possible whilst respecting basic principles governing corpus design, such as representativeness and balance, and delivering high quality in terms of both audio quality and annotations. An overview is provided of the composition of the current Core corpus of around 20 hours of data and of the human annotation effort involved. We also carry out a small qualitative analysis of the output of a Maltese ASR system and compare it to the human annotators' output. Initial results are promising, showing that the ASR is robust enough to generate first-pass texts for annotators to work on, thus reducing the human effort, and consequently, the cost involved.

Keywords: spoken Maltese, speech corpus, low-resource

1. Introduction

Spoken corpora are a necessary resource for any language in the digital age. There are different types of speech corpora. Some audio collections simply provide transcripts of the data whilst other collections provide time-aligned transcriptions or annotation at different levels of structure, possibly also with some kind of metadata. Speech data acquisition has been referred to as the "underestimated challenge" (Niebuhr and Michaud, 2015). In the case of low-resource languages such as Maltese, the task of developing such a corpus can be even more challenging.

This paper aims to present the design and compilation of a spoken corpus for Maltese. A brief overview of the language technology support for Maltese is provided in Section 2. A discussion on the nature of a spoken corpus and/or speech corpora follows in Section 3, setting the scene for discussion of the design concept of a spoken corpus for Maltese intended to cater to as wide a variety of needs as possible. Section 4 outlines work done so far on compilation and annotation of a Core collection, spoken corpus for Maltese. A preliminary evaluation of the annotation efforts involved is also carried out. Section 5 concludes the paper.

2. Language technology for Maltese

The situation of language technology support for Maltese described by Rosner and Joachimsen (2012), though far from rosy, was already beginning to improve in so far as text corpora were concerned: the same could not be said for speech at the time. The situation has improved substantially since then. Rosner and Borg (2023) report on improvements whilst also pointing out gaps with respect to tools, resources and support, highlighting speech technologies for Maltese as an area in need of improvement, whilst Skowron et al. (2023) report on work carried out by Mena et al. (2020) which showed that data augmentation and pooling methods can be used effectively in the context of a scarcity of training data, as is the case for Maltese.

2.1. Overview of resources for Maltese

The main digital resource so far continues to be the Maltese Language Resource Server, MLRS[1], described by Rosner and Joachimsen (2012, p. 75) as "an extensible computational infrastructure in the form of a server providing the basic functionality to enable access over the web to available corpora, some services, and a rudimentary system to facilitate the submission of contributions". In turn, the main resources included in the MLRS are a number of corpora, foremost amongst which is the *Korpus Malti*. Gatt and Čéplö (2013) give a brief overview of the data in the corpus at the time of writing: together these consisted of a total of around 250 million tokens. A newer version of the *Korpus Malti*, v4.0 has been published since then (Micallef et al., 2022), with more careful curation of the data sources from which the corpus is collected. The current size involves around 500 million tokens and further development is in progress. The authors acknowledge that one of the challenges of the opportunistic nature of these (mainly text) corpora is that achieving representativeness and balance, two of the important mainstays of any corpus is particularly difficult, see e.g. Sinclair (2005) and summary of basic principles for designing cor-

[1] http://mlrs.research.um.edu.mt

20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

pora in Knight and Adolphs (2022). Recent efforts have also resulted in a National Language Technology Platform (Cortis et al., 2021; Vasiļevskis et al., 2022) which is now online and focuses mainly on Machine Translation[2].

A number of other resources, including corpora, are included in the MLRS, (Rosner and Joachimsen, 2012) but we will move on here to examine the situation with corpora of spoken data.

2.2. What about Speech?

As mentioned above, the greater part of the *Korpus Malti* collections, including the newer versions, involve text rather than speech. In fact, the greater part of the spoken data included in the *Korpus Malti* corpora involve presumably modified or cleaned up versions of the written records of parliamentary debates put together by the parliament scribes. A small amount of data involving speeches is also available. The parliamentary speech data in particular is far-removed from everyday conversational speech in its stylistic features (see Vella, Magro, and Chetcuti, 2015), containing, amongst other features, frequent repetitions for rhetorical effect and greater than usual use of both silent and filled pauses, not to mention different use of lexical and morphosyntactic structures. Moreover, whilst the audio recordings corresponding to the parliamentary data transcripts are available from the parliamentary archives[3], a fair amount of searching is required to link specific texts to their audio counterparts.

Nevertheless, four main types of spoken data are available to date:

- *Korpus Malti* v3.0 – transcripts of c. 43,400,000 + 50,000,000 (*bulbulistan*) tokens involving "parliamentary debates" data and an additional 18,000 tokens involving "speeches" (Gatt and Čéplö, 2013).
- *CommonLanguage* – 1h of open access audio recordings (Sinisetty et al., 2021).
- *Common Voice* – currently includes 18h of crowdsourced audio recordings of read sentences together with the text prompts for these – in this version 50% of these data have been validated (Mena et al., 2020).
- *MASRI-Headset* – 8h audio recordings of speech read from text prompts and additional spontaneous speech with transcriptions "suitable for training ASR systems" (Mena et al., 2020, p. 6382).

The "spoken" component in *Korpus Malti* stands out from the remainder of the spoken data in that they were "spoken" first (even if parliamentary interventions may sometimes be wholly or partly scripted), and transcribed later. Of course, since recordings of these data are available, they constitute a large (and continually increasing) amount of data. However, such data can in no way be considered representative of naturally occurring speech for reasons mentioned earlier.

The other corpora listed above share one element, which is that they all involve read speech: speakers were recorded whilst reading text presented to them in the form of some kind of prompt. This means that these data come with a ready text to accompany the recordings.

Apart from the above, a number of smaller spoken (e.g. *MalToBI*[4]) and/or multimodal corpora (e.g. *MAMCO*[5]) annotated in line with conventions and standards developed over the years in the context of various projects, are also available, as are some collections of other data harvested from miscellaneous sources – the latter have not yet been transcribed.

To date however, there has been no attempt to collect a dedicated spoken corpus for Maltese which will cater to as wide a variety of user needs as possible, in other words, not just to the technological community but also to speech scientists and the public at large. It is in this context that the work on the *Korpus tal-Malti Mitkellem*, *KMM*, 'Corpus of Spoken Maltese', is being carried out. We delve further into the requirements of a spoken corpus in Section 3 below.

3. Spoken corpora

By definition, a spoken corpus consists minimally of two elements:

- **audio** (and, increasingly, *video*) **recordings** involving spoken language; and
- an accompanying text or **transcript** allowing for searchability of the content of the spoken data.

For most purposes, these two elements are sufficient for use as training data, as is the case when developing speech technologies of different sorts (e.g. automatic speech recognition and text-to-speech systems).

A third element is often also included in spoken corpora:

[2] https://traduzzjoni.mt/

[3] https://parlament.mt/en/menues/reference-material/archives/media-archive

[4] https://hdl.handle.net/21.11129/0000-000B-D309-D

[5] https://sites.google.com/view/mamcocorpus/home

- **time-aligned transcriptions** allowing recordings to be searched in relation to the text.

For use by those carrying out research on speech of different sorts, a fourth element is also needed:

- **metadata** on the speakers and the recordings.

It is a well-known fact that spoken corpora, unlike text corpora, continue to be under-represented e.g. whilst 133 spoken corpora are included in CLARIN[6], these corpora represent only 15 languages. In cases where National corpora contain a spoken component, such as in the case of the British National Corpus (BNC)[7], this is more often than not relatively small. In fact, remedying this lacuna led to the development of the *Spoken BNC2014* (Love et al., 2017). Amongst the reasons for this lacuna is the fact that spoken corpora are much more time-consuming to design, collect, process, extend and maintain in terms of human effort, and therefore expensive in financial terms when compared to text corpora. For this reason, as Knight and Adolphs (2022, p. 25) say, "the issue of cost-effectiveness requires consideration, in particular, weighing up the advantages between capturing large amounts of data (in terms of time, number of encounters or discourse contexts), the amount of detail added in transcriptions and annotations and the nature of analyses that the data may support relative to the cost of carrying out these tasks".

3.1. Desiderata for a new spoken corpus of Maltese

For a large number of reasons, including the costliness involved in pre-processing spoken data for inclusion in a corpus, the opportunistic approach to corpus collection does not lend itself well to spoken corpora. At some level of course, if all that is required is speech without the minimal requirement of the accompanying text, then accepting any data which comes along would not be a problem. In practice however, what we want is a corpus which will be seen as a kind of proxy for present-day Maltese and, for this reason, trying to respect the principles of representativeness and balance in particular, whilst also juggling with the notion of size to the extent possible for reasons mentioned above, is important. One further consideration is that, for reasons to do with the high cost of constructing a spoken corpus, we would like the corpus to serve as a reference point for as wide a variety of user needs as possible, as mentioned earlier.

3.2. The design of the Korpus tal-Malti Mitkellem, KMM

The design concept for the *Korpus tal-Malti Mitkellem*, *KMM*, centres around the idea of a dedicated corpus of spoken Maltese consisting of a Core collection which can be extended in principled ways, together with Satellite material of two sorts: Donated material from corpora collected and annotated in the context of other projects (see e.g. 2.2 above) and material Harvested from different sources (e.g. online programmes, podcasts, radio and television shows, etc.), transcribed and annotated following established conventions and standards. We will focus in this paper on what we are referring to as the Core collection.

The considerations outlined above on the extremely time-consuming, and therefore expensive task involved in constructing spoken corpora bring with them some important consequences for representativeness, balance and homogeneity[8] as set out in Knight and Adolphs (2022).

The work on the *KMM* started with an attempt at outlining a number of discourse events or text-types which would form the basis of data collection from speakers for the Core collection. These text-types are described briefly below.

3.2.1. Text-types in the Core collection

The design of the material for the Core collection is intended to allow for speech involving a variety of text-types ranging from easy-to-process read data (easy in the sense that a text of the audio data collected is available *a priori*) to unscripted data. Although we do intend to collect samples of free speech, it is expected that the majority of the unscripted data will be that requiring speakers to either respond to a prompt of some sort, or participate in some type of task-oriented activity.

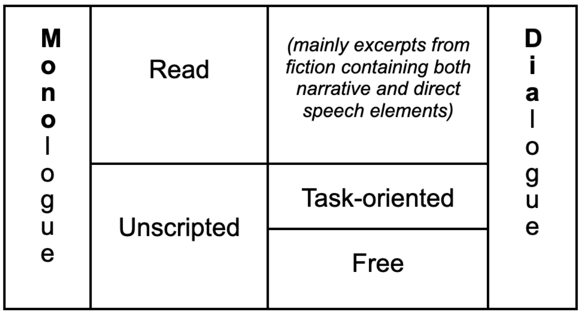
Speakers who volunteer to participate in the project are asked whether they would be willing to engage in a number of speaking tasks alone (monologue, **Mono**) or in conversation with a speaking partner (dialogue, **Dia**). Speakers will not be required to complete the full battery of tasks although efforts will be made to collect as much data as possible from each participant. The general design of the Core collection in terms of types of data is shown in Figure 1.

The reading (**Mono**) task used in the collection of Core data is based on excerpts from novels written by contemporary authors and normally includes a small amount of direct speech. All necessary permissions to use these excerpts have been obtained. The read data comes with the advantage of a small amount of data from different speakers

[6] https://www.clarin.eu/resource-families/spoken-corpora

[7] http://www.natcorp.ox.ac.uk/

[8] See Cavaglià (2002).



Monologue	Read	*(mainly excerpts from fiction containing both narrative and direct speech elements)*	Dialogue
	Unscripted	Task-oriented	
		Free	

Figure 1: Different types of data for the Core collection.

which will allow for direct comparison. All read data is **Mono** data. All other **Mono** tasks are unscripted and include explaining a favourite recipe, retelling the stories which unfold in a number of Disney (non-verbal) shorts, talking about a specified topic, responding to a picture or a Mind Map, and finally, free speech.

The only task which was restricted to pairs involves use of a Map Task (Anderson et al., 1991). This task requires speakers to work collaboratively to complete an information gap activity. **Dia** data from this task is highly task-oriented, and generates speech which is natural mainly in that engaging in the task takes the speakers' minds off the fact that they are being recorded. The task-oriented nature of the Map Task needs to be kept in mind. The battery of **Dia** tasks complements the **Mono** tasks described above. As already mentioned, we also intend to collect samples of free speech wherever possible.

In summary, it is expected that the Core collection will consist of data involving some read speech, whilst prioritising unscripted data. It will be roughly divided between **Mono** and **Dia** speech. Whilst we will seek to include data which is ecologically valid in being as close as possible to naturally occurring speech, an important *proviso* needs to be made at this point: this is that recording speakers as they converse (**Dia** data), and more so whilst they simply perform a task solo (**Mono** data), is **not** totally natural. Data collected via tasks such as those involved here is certainly more ecologically valid than read speech. Nevertheless, opportunities to collect instances of truly free speech should be optimised whenever they present themselves.

3.2.2. Participants

Since the aim is to continue to extend the Core collection of the spoken corpus, participants will be recruited periodically via social media and using a friends-of-friends approach depending on the availability of funds. Participants will be required to complete a language background questionnaire as well as to read an information sheet explaining the aims of the project and sign a consent form. Select information from the language background questionnaire will form the basis of the (anonymised) metadata (see also 4.1.3) which will accompany recordings. Such metadata is yet to be compiled for the collection so far.

3.2.3. Recording

Recordings have been and will continue to take place in sound proofed premises wherever possible, although recordings made by participants on their own devices may also be accepted as long as they were made in a quiet environment and saved in lossless .wav format. External microphones will be used wherever possible, with speakers in **Dia** tasks each using their own mono-directional microphones to allow for separate channel recordings which make for easier annotation. In order for the Core collection of the spoken corpus to have as wide a reach as possible, and given that good acoustic quality is important to some expected users of the spoken corpus, the Core collection will prioritise best quality recordings.

3.3. Annotation of the corpus

A number of annotation training sessions were held in order to train annotators in the use of the *SPeech ANnotation Guidelines for Maltese*, *SPAN* (Vella et al., 2010). Primary annotation of the spoken data collected was carried out by these annotators following the established conventions and standards.

In this latter part of the project, a recently developed speech-to-text model for Maltese (Williams et al., 2023, described below in 3.3.2) is being tested to generate a first-pass text of the spoken data which can then be manually corrected by our human annotators.

3.3.1. Human Annotation

In the case of the read data, annotators were provided with the text and required to listen carefully to the recording and add to the text elements in the actual textual rendering produced by the speaker. The original text was corrected in order to reflect use of any normal disfluencies (deletions/insertions, false starts, hesitations, fillers etc.) in the actual reading.

Annotation of the unscripted data, **Mono** as well as **Dia**, was carried out in PRAAT (Boersma and Weenink, 2001), with one tier per speaker being used in the case of the **Dia** data. Annotators were required to insert boundaries at pauses in the speech (not necessarily the same as syntactic boundaries) and to then fill in the text for each interpausal stretch of speech. Once again, all instances of normal disfluencies, as well as a few other elements, were indicated in the transcript following the established *SPAN Guidelines*. Apart from the conventions mentioned above, the *SPAN Guidelines* also allow for elements such as variant pronunciations, ungrammatical forms, non-Maltese words etc. to be flagged up, e.g. *&puluzija* ‘police’, where the dictionary entry in Aquilina (1990) is *pulizija*, **fil-fuq* rather than *il-fuq* ‘above’, and /ratings/ respectively.

The reason for following (and propagating use of) the *SPAN Guidelines* (Vella et al., 2010) is that they allow for consistency across different annotators and corpora. Whilst it was not feasible to organise second-pass annotations of all the data, spot-checking of these is currently in progress, also with a view to identifying (the more accurate and efficient) annotators with the intention of recruiting these to contribute to the next phase of the project.

3.3.2. Automatic Annotation

The model used to generate automatic first-pass texts of the spoken data involves a recently developed version of Wav2Vec2 trained on Maltese data (Williams et al., 2023). Wav2Vec2 is an automatic speech recognition (ASR) model composed of 3 main components: a CNN feature extractor, transformer blocks, and a quantization module (Baevski et al., 2020). XLS-R is an instance of Wav2Vec2 pre-trained on 436k hours of speech from 128 different languages. In this work we are using the 2 billion parameter XLS-R which was fine-tuned on 50 hours of Maltese speech (Williams et al., 2023). We will refer to this model as Wav2Vec2MT.

The rest of this paper provides a discussion of aspects of the ongoing work of corpus collection, including by taking a look at ways in which the heavy-duty annotation work can be improved.

4. Work done to date and evaluation

As stated above, the Core collection is a work in progress. However, an overview of its state-of-play and a first evaluation of the annotation effort which has been undertaken is discussed below.

4.1. The Core collection to date

This section presents an overview of the composition of the Core collection in relation to the design concept presented in Section 3. The overview presented below is based on a Core collection of 20 hours collected and processed, but not fully annotated, at the time of writing.

4.1.1. Text-type composition

The text-type design for the Core collection of this new corpus involved collecting less of the read speech type of data. This is in line with the aim of prioritising data involving more naturally occurring speech represented by different types of unscripted data, and in this way moving towards better respecting the principle of representativeness.

Read vs Unscripted		
Read	10,820	15%
Unscripted	60,891	85%
Mono vs Dia		
Mono	34,387	48%
Dia	37,324	52%
Text-type		
Recipe	1,768	3%
Retell	4,857	8%
Topics	10,972	18%
Pictures	16,186	27%
Mind Map	20,130	33%
Map Task	2,627	4%
Free	4,351	7%

Table 1: Overview of the data distribution, by text-type, in seconds, and as a percentage, of the total.

As can be seen from Table 1, the skew towards read data has been redressed with the majority of the data (85%) consisting of unscripted data. There is also a good balance between **Mono** and **Dia** data in the collection so far. There is a reasonable balance in the data involving the different text-types although more work is needed to increase the amount of data involving the less-controlled of these, in particular free speech.

4.1.2. Demographically representative and balanced corpus

The aim at the start of this project was to collect data from as demographically balanced as possible a sample of speakers of Standard Maltese representing a cross-section of Maltese society in terms of sex, age, and to a lesser extent, level of education. A balanced distribution of speakers coming from different localities is also one of the desiderata, the idea being to take note of gaps as we continue to populate the Core part of the corpus and to seek to fill these gaps in future data collection attempts.

As can be seen from Table 2, the balance of speakers in terms of sex, age and education is reasonable. Efforts will be made to recruit more females as we progress with the data collection in order to achieve a better balance of female vs male speakers. We will also try to recruit more speakers who have not attended tertiary education studies.

Moreover, in the complex linguistic context of Malta (Vella, 2013), speakers are often not only bilingual but possibly also bidialectal. Homogeneity is a third criteria often invoked in discussions on the standards to be aimed at when compiling

Distribution by sex		
Male	23	62%
Female	14	38%
Distribution by age group		
20-29	10	27%
30-39	8	22%
40-49	7	19%
50-59	4	11%
>60	8	22%
Distribution by level of education		
Post-tertiary	24	65%
Tertiary	8	22%
Pre-University	4	11%
Secondary	1	3%

Table 2: Distribution of speakers in terms of sex, age group and education in the current Core collection.

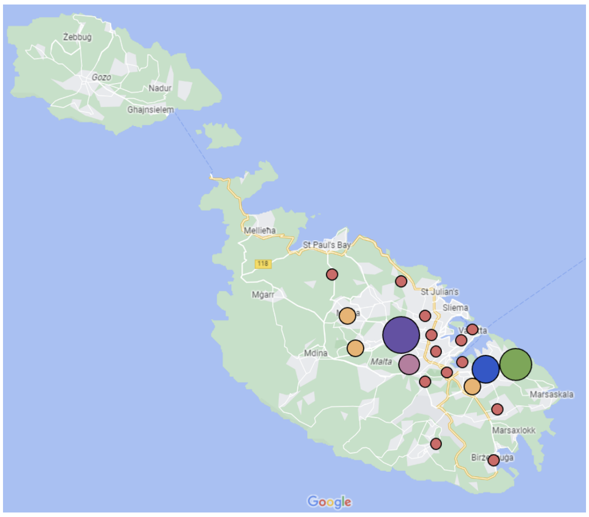


Figure 2: Distribution of speakers in the Core collection (bigger dots indicate a larger number of speakers).

a corpus, see Sinclair (2005), as well as Cavaglià (2002)). In the interest of homogeneity, but also since cross-linguistic influence between the languages and/or language varieties or dialects in a speaker's repertoire can be expected, data collection efforts have focussed on speech likely to be categorised as Standard Maltese. The idea was to recruit speakers whose spoken Maltese was least likely to be heavily influenced by English on the one hand, or by a dialect other than Standard Maltese on the other. Maltese rather than English dominant speakers, and mono-dialectal rather than bidialectal speakers, were prioritised. The responses provided by participants in their completion of the language background questionnaire served as a soft filter for following up with potential participants.

As can be seen from Figure 2, data collected so far involves speakers coming from a good spread of localities in Malta, although not from Gozo and some of the areas more prone to use of a dialect alongside Standard Maltese, for reasons discussed above. Speakers coming from Birkirkara, a locality in Malta with a very high population density[9], represent the largest group having contributed to the Core collection to date.

4.1.3. Labelling and organisation of data

In order to ensure anonymity, participants were assigned a code and all recordings were coded by task type and participant code e.g. the examples in 4.2.2 below were taken from an audio recording bearing the filename Mono_Task_Pics03_11M: this indicates that the data involves a Mono unscripted recording of a male speaker, 11M, responding to a trio of pictures coded as Pics03. The corresponding annotation files include the extension .txt and/or .TextGrid. Metadata information on the speakers and the quality of the recordings is yet to be compiled.

4.2. Evaluation of the annotation effort

As indicated earlier in Section 3, one aim of the Core collection of the *KMM* is that it be a representative and balanced collection of spoken texts with corresponding plain text transcripts and/or time-aligned transcriptions. In this section, we provide a first evaluation of the annotation effort carried out.

4.2.1. Human annotation

Calculating the amount of time involved in the human annotation effort is not a straightforward exercise for all sorts of reasons including that different annotators were involved, with different levels of experience and ability to work to the *Guidelines* etc. We have in fact carried out a spot-check exercise aimed at identifying the more accurate as well as efficient annotators for internal purposes. A full evaluation of the human annotation effort is yet to be carried out. Clearly, manual annotation by human annotators is at best fluctuating. Nevertheless, as a result of the work carried out to date, some fine-tuning to the instructions given to annotators has been agreed and these instructions will be put together in a protocol for annotators. This is expected to increase the consistency of the output of the human annotators.

In the meantime, in an attempt to speed up the annotation process, some of the data was run through the Wav2Vec2MT model introduced in 3.3.2. The text output of the model was then corrected by the human annotators. A preliminary evaluation of the

[9] https://www.citypopulation.de/en/malta/cities/

usefulness of this model for the purposes of speeding up the annotation process is presented below, starting with a comparison of some examples of machine output to the human output in 4.2.2 below.

4.2.2. Error rate analysis of machine as compared to human annotation

A brief examination of the gold-standard human annotation (**H** for Human) compared to the automatic output of the Wav2Vec2MT model (**M**, for Machine) can serve to throw light on the nature of the errors being made by the ASR model. One element which we mention here but will not take up any further, is that of the human annotators' transcription of "fillers" in particular, but also other normal disfluency elements which occur in speech. The *SPAN Guidelines* give clear instructions on the annotation of such elements, e.g. *'* indicates a deleted segment as in *'ed* for *qed* in (1), *ee* and *em* in (3) and (5) indicate specific fillers, whilst [ ] indicates the insertion of segments which do not show up in standard orthography such as the [i] in (4). A number of examples are discussed below (sourcefile: Mono_Task_Pics03_11M) to give a taste of the issues encountered by the Wav2Vec2MT model. In these examples, **M** elements which differ from their **H** counterparts are shown in **red** and **blue** respectively. Fillers and other normal disfluency markers are indicated in green.

(1)
H - ee pjuttost mitluq em **eżawrit għajjien**
'FILLER rather shabby FILLER exhausted tired'
M - pjuttost mitluq **eeżawrit ħajjien**
(2)
H - fl-ewwel+ ee+ ritratt **għan'na raġel**
'In the first FILLER photo we have a man'
M - fl-ewwel ritratt **tan-nar aġiel**
(3)
H - Għaliex taħseb li **'**ed **jistennew** in-nies ta' dawn l-istampi?
'Why do you think that they are waiting, the people in these pictures?'
M - għaliex taħseb li qed **istennew** in-nies ta' dawn l-istampi
(4)
H - f'dan il-mument **kien 'ed jesperjenza** [i]l-ġmiel tan-natura
'In this moment he was experiencing the beauty of nature'
M - f'dan il-mument **kienet esperjenza** l-ġmiel tan-natura
(5)
H - X'taħseb li se **jiġri** wara dan il-mument+ maqbud fl-istampi?
'What do you think will happen after this moment captured in the pictures?
M - x'taħseb li se **jiġu** wara dan il-mument maqbud fl-istampi

The first issue relates to errors in the identification of specific segments and/or to segmentation. (1) involves an example of the former. The system gives a <ħ>, normally realised as a [h] for <għ>, often not vocalised but "pronounced" with the sound of the following vowel, in this case an [ɐj]. The context here is interesting since final voiceless stops in Maltese are often quite heavily aspirated so that the final /t/ in *eżawrit* is realised as [t^h]: the aspiration here seems to have been interpreted by the model as consisting of a [h] and mis-segmented as the first sound of the following word, thus giving the non-word **ħajjien* for *għajjien*. A further error occurs here, which is that the filler before *eżawrit* has been integrated into the first part of *eżawrit*, thus giving **eeżawrit*.

Two further instances arising from mis-segmentation but resulting also from misinterpretation by the model of the connected speech version of *għandna* pronounced as [ˈɐnnɐ] i.e. without a medial [d], can be noted here. The segment sequence here is: [rɪtrɐtt ɐnnɐ rɐːdʒɛl] (spaces here are used simply to show where the different words end and begin). The model's reinterpretation of this sequence gave: [rɪtrɐt tɐn nɐr ɐːdʒɛl], and hence, *tan-nar *aġiel*.

The three remaining examples illustrated here involve errors in the verbs. Such errors are clearly more problematic in that they result in a loss of meaning. In (3) *qed jistennew* is rendered as *qed *istennew* even though there is a clear transition from a [j] to an [ɪ] in the acoustic signal. The glottal stop /ʔ/ at the beginning of *qed* in (4), as in (3), is not pronounced by the speaker (this is indicated by means of the ' in the **H** annotation). This has no negative impact on the **M** output. What does have an impact in (4) is omission of the final /t/ in *qed* (final stops are devoiced in Maltese), with the result that the complex verb sequence /kɪːn ʔɛt jɛspɛrjɛntsɐ/ pronounced without the glottal stop, caused the system to reinterpret the acoustic sequence as [kɪːnɛt] and then replace the second element in the verb sequence by the noun [ɛspɛrjɛntsɐ]. The final example (5), involves the **M** output of *jiġu* [jɪdʒʊ] rather than *jiġri* [jɪdʒrɪ]. An examination of what is actually happening shows that there is some element of coalescence between the final /ɪ/ of *jiġri* and the /w/ at the beginning of *wara*. Interpretation of 'r' is not straightforward since the pronunciation of this segment in Maltese (as in other languages) is so variable. For example, affricated 'r's have been noted by Vella and Grech (personal communication), and such affrication may have caused the 'r' segment to be lost to the model as a result of the frication at the end of the preceding /dʒ/, hence giving *jiġu* for *jiġri*. Wav2Vec2MT does not include

Type	# files	hh:mm:ss	WER%	CER%
Dia	8	01:30:31	71.44	50.36

Table 3: % Word and Character Error Rates for the **Dia** files.

Type	# files	hh:mm:ss	WER%	CER%
Rec	12	00:20:35	32.64	11.47
Spon	11	00:45:32	32.71	12.98
Pics	12	00:48:26	30.74	11.20
Retell	12	00:49:05	22.68	8.25

Table 4: % Word and Character Error Rates for the **Mono** files, by text-type.

a Maltese language model (Williams et al., 2023) and therefore, such errors, resulting from interpretations of the spoken output mainly based on the acoustic model, are not surprising.

4.2.3. Human vs machine annotation comparison

We were interested in analysing the output of the ASR model made available by Williams et al. (2023) when compared to the human annotations. We use Word and Character Error Rates (WER & CER), expressed as percentages for this purpose. The *SPAN Guidelines* annotation labels for elements such as “fillers” mentioned in 3.3.1 were removed as the ASR model has not yet been trained on data including such elements. The texts were compared on the basis of alphabetic strings, with only the <'> (apostrophe) and <-> (dash) characters left in the annotator text. Boundary placement does not feature in this analysis.

The results of the analysis carried out are shown in Table 3 for the **Dia** data and Table 4 for the **Mono** data, with results provided separately by text-type. Focusing on the WER, we can see that the different text-types impacted the ASR model’s ability to transcribe the speech accurately to different extents. The error rate for the **Dia** data is relatively high at 71.44%. Considering that in collecting the **Dia** data, we used separate microphones for the two speakers, there was an expectation that the WER would not be as high as is the case. The distance between the two microphones may not have been enough to prevent the speech of the two speakers being caught by both microphones, especially in instances, not infrequent in Maltese, of overlap (Paggio and Vella, 2014). It is likely that overlapping speech was one reason that the ASR model performed relatively worse in the case of the **Dia** data as compared to **Mono** data.

The error rates for the **Mono** data, by contrast, are much lower overall than those for the **Dia** data, aligning quite well with results reported in Williams et al. (2023). It is interesting that the model performed better on the retelling task as compared to the other tasks, and also that the results for the recipe description and the spontaneous **Mono** data are very similar. In view of the discussion earlier of the issue of prioritising the collection of data which is as close as possible to naturally occurring speech, a closer examination of the data collected as a function of the different tasks would be worth carrying out.

Overall, however, these results suggest that the fine-tuned XLS-R model is indeed quite robust. It is clear that the output of the ASR model, at least in the case of the **Mono** data, is already good enough to allow for a pipeline similar to that reported in Vella et al. (2024) for Maltese English using YouTube captioning as a starting point, to be put in place for use in the annotation of spoken data collected in the context of this project.

5. Conclusion

In this work, we presented the design concept (Subsection 3.2) for a "dedicated" corpus of spoken Maltese centred around the idea of a representative and balanced Core collection. We believe that this is a necessary step to ensure that the corpus can serve multiple purposes. The aim is to supplement this part of the corpus with data from both Donated and Harvested collections whilst also continuing to expand the Core collection. In order to ensure that the spoken corpus does cater to the needs of a wide variety of users (referred to at the end of Subsection 2.2), the Core part of the corpus should aspire to the highest standards possible in all respects, including in terms of the quality of both the audio data and the annotations.

The objective of the annotation element is that the material for the Core collection will include accompanying transcripts and, for at least part of the corpus, time-aligned transcriptions, that are as accurate as possible. Applying the Wav2Vec2MT model (Williams et al., 2023) to provide an initial transcription has given positive initial results and merits further investigation. From the corpus linguistic perspective, care needs to continue to be taken to ensure representativeness and balance as the corpus grows, as well as to fine-tune the protocol for annotators in order to improve efficiency. From the computational perspective, we could investigate the fine-tuning of an XLS-R model specifically on this corpus, in a similar way to Williams et al. (2023). We would expect it to provide better transcriptions, especially in relation to boundary placement. In an ideal scenario, similar to Fallgren et al. (2019), we would expect such a tool to aid human annotators, thus reducing the time and effort required to produce high-quality annotations to accompany recordings. As the corpus grows to include a greater amount of Maltese dialogue data, we plan to experiment with diarization techniques. In the meantime, we continue to work towards improving the CER/WER for standard ASR, still a challenge given the low amount of data available for Maltese.

6. Acknowledgements

This work was funded by the Malta Digital Innovation Authority (MDIA) under the National AI Strategy Initiative. We also acknowledge support from LT-Bridge Project (GA 952194). We would also like to thank all speakers who participated in this study: without their active contribution a spoken corpus would not be possible.

6.1. Ethical considerations and limitations

Participants and annotators were paid for their work. Those who participated in recordings were offered payment at an average rate of €10 per hour. Annotators were paid at an hourly rate depending on their qualifications (€10 in the case of those with an undergraduate degree, €12 in the case of those with a postgraduate degree or a Maltese proofreading certificate). Ethical approval from the ethics board of the University of Malta was obtained prior to the start of data collection.

7. Bibliographical References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC map task corpus. *Language and Speech*, 34(4):351–366.

Joseph Aquilina. 1987. *Maltese-English Dictionary Vol. I, A-L*. Midsea Books, Valletta, Malta.

Joseph Aquilina. 1990. *Maltese-English Dictionary Vol. II, M-Z*. Midsea Books, Valletta, Malta.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Paul Boersma and David Weenink. 2001. PRAAT, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Gabriela Cavaglià. 2002. Measuring corpus homogeneity using a range of measures for inter-document distance. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands - Spain. European Language Resources Association.

Keith Cortis, Judie Attard, and Donatienne Spiteri. 2021. Malta National Language Technology Platform: A vision for enhancing Malta's official languages using Machine Translation. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 12–19, Online (Virtual Mode). INCOMA Ltd.

Per Fallgren, Zofia Malisz, and Jens Edlund. 2019. How to annotate 100 hours in 45 minutes. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 341–345. ISCA. QC 20200310.

Albert Gatt and Slavomír Čéplö. 2013. Digital corpora and other electronic resources for Maltese. In *Proceedings of the International Conference on Corpus Linguistics*, pages 96–97. UCREL, Lancaster, UK.

Dawn Knight and Svenja Adolphs. 2022. Building a spoken corpus: what are the basics? In Anne O'Keeffe and Michael J. McCarthy, editors, *The Routledge Handbook of Corpus Linguistics*, pages 21–34. Routledge.

Robbie Love, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3):319–344.

Carlos Mena, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, Amanda Muscat, and Ian Padovani. 2020. MASRI-HEADSET: A Maltese Corpus for Speech Recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6381–6388, Marseille, France. European Language Resources Association.

Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natu-*

ral Language Processing, pages 90–101, Hybrid. Association for Computational Linguistics.

Oliver Niebuhr and Alexis Michaud. 2015. Speech data acquisition: the underestimated challenge. *KALIPHO - Kieler Arbeiten zur Linguistik und Phonetik*, 3:1–42.

Patrizia Paggio and Alexandra Vella. 2014. Overlaps in Maltese conversational and task-oriented dialogues. In *Proceedings of the First European Symposium on Multimodal Communication*. Linköping Electronic Conference Proceedings.

Michael Rosner and Claudia Borg. 2023. *Language Report Maltese*. In: Georg Rehm and Andy Way, editors, European Language Equality: A Strategic Agenda for Digital Language Equality. Springer International Publishing, Cham.

Michael Rosner and Jan Joachimsen. 2012. *Il-Lingwa Maltija Fl-Era Diġitali – The Maltese Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at `http://www.meta-net.eu/whitepapers`.

John Sinclair. 2005. *Corpus and Text: Basic prinicipiles*, AHDS guides to good practice. Oxbow Books.

Ganesh Sinisetty, Pavlo Ruban, Oleksandr Dymov, and Mirco Ravanelli. 2021. Commonlanguage.

Marcin Skowron, Gerhard Backfried, Eva Navas, Aivars Bērziņš, Joachim Van den Bogaert, Franciska de Jong, Andrea DeMarco, Inma Hernáez, Marek Kováč, Peter Polák, Johan Rohdin, Michael Rosner, Jon Sanchez, Ibon Saratxaga, and Petr Schwarz. 2023. *Deep Dive Speech Technology*, pages 289–312. Springer International Publishing, Cham.

Artūrs Vasiļevskis, Jānis Ziediņš, Marko Tadić, Željka Motika, Mark Fishel, Hrafn Loftsson, Jón Gu, Claudia Borg, Keith Cortis, Judie Attard, and Donatienne Spiteri. 2022. National language technology platform (NLTP): overall view. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 345–346, Ghent, Belgium. European Association for Machine Translation.

Alexandra Vella. 2013. Languages and language varieties in Malta. *International Journal of Bilingual Education and Bilingualism*, 16(5):532–552.

Alexandra Vella, Flavia Chetcuti, Sarah Grech, and Michael Spagnol. 2010. Integrating annotated spoken Maltese data into corpora of written Maltese. In *Proceedings of the Workshop on Language Resources and Human Language Technologies for Semitic Languages*, pages 83–90. Seventh Conference on International Language Resources and Evaluation (LREC 2010).

Alexandra Vella and Paulseph-John Farrugia. 2006. MalToBI - Building an annotated corpus of spoken Maltese. In *Speech Prosody, Dresden*.

Alexandra Vella and Sarah Grech. 2022. What can a corpus tell us about phonetic and phonological variation? In Anne O'Keeffe and Michael J. McCarthy, editors, *The Routledge Handbook of Corpus Linguistics*, pages 281–295. Routledge.

Alexandra Vella, Sarah Grech, Ian Padovani, and Maria-Christina Micallef. 2024. Resources and tools for pre-processing speech data in a lesser-known variety of English. In *Proceedings of the Twentieth International Congress of Phonetic Sciences*. Guarant International.

Alexandra Vella, Elgar Paul Magro, and Flavia Chetcuti. 2015. Cohesion phenomena in Maltese parliamentary debates. In *Proceedings of the Fifth International Conference on Maltese Linguistics*.

Aiden Williams, Andrea DeMarco, and Claudia Borg. 2023. The applicability of Wav2Vec2 and Whisper for low-resource Maltese ASR. In *Second Annual Meeting of the Special Interest Group on Under-resourced Languages, a satellite Workshop of Interspeech 2023*.