Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6381–6388
Marseille, 11–16 May 2020

MASRI-HEADSET: A Maltese Corpus for Speech Recognition

Carlos Mena, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, Amanda Muscat, Ian Padovani
University of Malta
{carlos.hernandez, albert.gatt, andrea.demarco, claudia.borg, lonneke.vanderplas
amanda.muscat.11, ian.padovani.16}@um.edu.mt

Abstract

Maltese, the national language of Malta, is spoken by approximately 500,000 people. Speech processing for Maltese is still in its early stages of development. In this paper, we present the first spoken Maltese corpus designed purposely for Automatic Speech Recognition (ASR). The MASRI-HEADSET corpus was developed by the MASRI project at the University of Malta. It consists of 8 hours of speech paired with text, recorded by using short text snippets in a laboratory environment. The speakers were recruited from different geographical locations all over the Maltese islands, and were roughly evenly distributed by gender. This paper also presents some initial results achieved in baseline experiments for Maltese ASR using Sphinx and Kaldi. The MASRI-HEADSET Corpus is publicly available for research/academic purposes.

Keywords: Maltese, Speech Corpora, Automatic Speech Recognition

1. Introduction

As digital resources and tools for Natural Language Processing and language-enabled interfaces become ever more common in daily use and commercial applications, it has become increasingly important to ensure that all languages are adequately represented in the digital sphere. This concern is evident, for example, in a recent resolution passed by the European Parliament to safeguard *language equality* in the digital age (European Parliament, 2018).

In this regard, so-called ‘under-resourced’ or ‘low-resourced’ languages have been an ongoing concern within some sectors of the NLP community. For example, in a series of white papers published in 2011-12, reviewing the level of digital support for 31 of the languages of the European Union, the METANET Initiative[1] concluded that for a significant number of these languages, support in most areas of NLP was at best fragmentary and in some cases, weak or non-existent.

The METANET papers took the step of determining the level of support for different languages with reference to specific tasks or domains, namely: machine translation, text analysis, speech/language resources and speech processing. In principle, a language could have strong support in a subset of these, with weak or fragmentary support in others. This is reminiscent of the strategy used by Krauwer (2003), who discussed the notion of an ‘under-resourced’ language in terms of a minimal set of language resources (a Basic Language Resource Kit, or BLARK) necessary to undertake further precompetitive research and education. A somewhat more wide-ranging definition was more recently offered by Besacier et al. (2014), in their review of ASR for low-resourced languages, where the criteria included the following:

- Lack of a unique writing system or stable orthography.
- Lack of linguistic expertise.
- Limited presence on the web.
- Lack of electronic resources for speech and language processing.

[1] `http://www.meta-net.eu/`

1.1. The case of Maltese

Maltese, the national language of Malta, was one of the languages which, at the time of the METANET white papers, was ranked as having weak or no support under all four of the headings listed above (Rosner and Joachimsen, 2012). This is in spite of the fact that Maltese does not suffer from the first two of the list of criteria offered by Besacier et al. (2014): the language not only has a long written tradition and a stable orthography (Azzopardi-Alexander and Borg, 2013), but is very well-studied linguistically at all levels, including the morphological (Mifsud, 1995; Hoberman, 2007; Gatt and Fabri, 2018, *inter alia*), syntactic (Fabri, 1993; Čéplö, 2018, *inter alia*), and phonological (Vella, 1994), as well as in terms of its historical development (Brincat, 2011) and typological status (Comrie, 2009). On the other hand, its web presence, while comparatively small compared to that of languages such as English, could be argued to be proportional to the size of its community of speakers.

It is the fourth criterion listed by Besacier et al. (2014) – lack of electronic resources – that was the basis for the conclusions reached by Rosner and Joachimsen (2012). There are numerous factors which could have contributed to this situation. First, Maltese is a language with a small number of speakers (ca. 500,000); this fact makes NLP for Maltese appear less economically advantageous, at least for commercial developers. Second, Malta is officially a bilingual country, with English as the second language. The lack of digital support for Maltese has meant that many users resort to English for their electronic and online communication.

On the other hand, this also implies that people whose language proficiency is Maltese-dominant might simply abstain from communicating through certain channels where their native language is not represented. Indeed, where speech is concerned, this is arguably also true for English speakers in Malta, since available speech interfaces, for

example on mobile devices, tend to be developed for the recognition or synthesis of varieties of English (such as British, American or Australian) with a larger number of speakers, while Maltese English, which has been argued to be a variety in its own right (Grech, 2015), is unsupported. As a matter of fact, digital support for the Maltese language has improved drastically since the publication of the METANET white paper (and this is likely true for a number of other languages covered by the white paper series). To take some examples, large annotated corpora for Maltese are now available, with accompanying tools for segmentation and labelling, including tokenisation and part-of-speech annotation (Gatt and Čéplö, 2013).[2]. Advances have been made in the development of electronic lexicons (Camilleri, 2013) and in automatic morphological analysis and labelling (Borg and Gatt, 2017; Ravishankar et al., 2017) as well as dependency parsing (Tiedemann and van der Plas, 2016; Zammit, 2018).
Advances in speech technology for Maltese have however been comparatively limited. While there have been successful attempts to build speech synthesis systems using concatenative techniques (Micallef, 1997; Borg et al., 2011), no tools currently exist for Automatic Speech Recognition (ASR). This is partly due to a substantial data bottleneck where resources for speech engineering are concerned.

1.2. Aims of the present paper

The present paper addresses this gap, presenting a new corpus for ASR, built in the context of ongoing work in the MASRI (Maltese Automatic Speech Recognition) project. The paper describes the MASRI-HEADSET Corpus (MHC), the first corpus in Maltese suitable for training ASR systems.
The rest of this paper is structured as follows. Section 2. describes the design strategy and the recording process leading up to the corpus. In Section 3., we demonstrate its suitability for creating acoustic models for ASR by running some experiments in CMU-Sphinx[3] and Kaldi[4]. We also discuss the construction of pronunciation models of Maltese. Section 4. concludes the paper with a brief overview of ongoing work on expanding Maltese speech resources, and on deploying recent techniques for ASR that reduce the dependency on large data sources.

2. The MASRI-HEADSET Corpus (MHC)

MASRI-HEADSET is the first Maltese corpus specifically designed with ASR systems in mind. While it is comparatively small, it constitutes the first step in the creation of larger-scale resources for speech recognition in Maltese.
In this section, we describe the design of MHC, including the selection of participants whose voices were recorded for the corpus, as well as the process whereby textual prompts

Figure 1: Regions in the Maltese islands (NSO, 2019)

were selected. Finally, we discuss the characteristics of the final release version of the corpus.

2.1. Corpus Design and Collection

Participants Expressions of interest in participation were solicited via online advertisements. Initially, 61 people expressed interest. In an effort to ensure adequate representation of accent and speaker variation, participants were selected to achieve a balanced sample in terms of gender, age and geographical location, the latter being defined in terms of the six regions of the Maltese islands defined by the National Statistics Office for data collection purposes (NSO, 2019), as shown in Figure 1. Out of a total of 61 people who expressed interest, 25 individuals were recruited. The sample is approximately gender-balanced, with 13 female and 12 male speakers (age range = 18 to 31; mean age = 23.9, SD = 3.54). Table 1 provides a summary of the participants and the average number of hours and utterances split by gender. Participation was remunerated at the rate of €15 for approximately one hour of recording. Written consent was obtained and the procedure was screened by the procedures of the University of Malta Research Ethics Committee.[5]

	Female	**Male**
Number of Speakers	13	12
Average amount per speaker	19m:57s	18m:53s
Average utterances per speaker	158.0	150.8

Table 1: Overview of the MASRI-HEADSET Corpus

Language dominance: Given the bilingual situation in Malta, where apart from Maltese, English is also an official language, and where some individuals are English-dominant, prospective participants were asked to respond to an online language background questionnaire. The 25 participants selected were all Maltese-dominant speakers, with some also speaking dialects of Maltese, in addition

[2]Many of these resources are available through the Maltese Language Resource Server: `https://mlrs.research.um.edu.mt`

[3]An open source speech recognition toolkit. See `https://cmusphinx.github.io/`

[4]An open source speech recognition toolkit. See `https://kaldi-asr.org/index.html`

[5]`https://www.um.edu.mt/urec`

to standard Maltese. Speakers with dominance in English rather than Maltese were not eligible for participation in this particular corpus collection.

Sentence selection: Approximately 200 samples of sentential prompts were selected from the Korpus Malti v3.0[6] (Gatt and Čéplö, 2013). Samples were built using the following procedure, which was intended to ensure that each sample would contain as broad a range of phonological and phonotactic variation as possible:

1. The full corpus was transcribed automatically using a rule-based grapheme-to-phoneme (G2P) mapping procedure;
2. A trigram model over phone sequences was constructed based on the full corpus;
3. The corpus text was broken up into blocks of approximately 75 orthographic words each, with each block paired with its orthographic transcription. Blocks were randomly shuffled. Samples were then constructed by greedily adding blocks to a sample, as per the following algorithm:
4. For $i = 1$ to n samples, with size m:
   (a) initialise sample s_i to $\emptyset$
   (b) While $|s_i| < m$, sample the next best block and include it in s_i

where the 'next best block' is defined in terms of coverage: given the current distribution of trigrams in blocks already included in the sample, the best next block is the one that results in the greatest amount of variation in the sample once it is included. Once a sample was constructed, prompts containing proper names and/or words in English were pruned.
The result of this procedure was a set of samples, each consisting of blocks of text (our prompts). Each of the 25 participants was required to read each block of text in the sample assigned to them.

Recording methods: Speech recordings were made in a quiet room at the Faculty of ICT at the University of Malta. The recordings were done in a dual fashion: close (using a headset) and far-field (via a microphone at a distance of approximately 3 metres from the participant). Only the headset data is included in the release under discussion.
Each speaker read prompts from one of the samples, shown on a computer screen in random order using SpeechRecorder (Draxler and Jänsch, 2004), which was also used to capture the recordings. The headset recordings were captured using a Sennheiser PC-8 (48KHz) mic/headset and the recording sessions typically lasted between 45 minutes and one hour for each of the participants.

2.2. Characteristics of the Release Version

After recording, the data underwent a process of pruning and normalisation, resulting in the following characteristics:

- The MHC has an exact duration of 8 hours and 6 minutes. It has 3864 audio files.
- Every audio file contains only the voice of one single speaker with no background noise.
- Utterances with stuttering and/or mispronunciations were pruned.
- Data in MHC is classified by speaker, with all recordings of one speaker stored in a separate directory.
- Data is also classified according to the gender (male / female) of the speakers.
- Audio files in the MHC are distributed in a 16khz @ 16bit mono format.
- The MHC corpus contains 3864 utterances (one per audio file), with a total of 11,503 unique words or tokens.
- Every audio file has a unique ID that is compatible with ASR engines such as Kaldi (Povey et al., 2011) and CMU-Sphinx (Lamere et al., 2003).
- All textual transcriptions in MHC are lowercased. All punctuation is removed, with the exception of hyphens (-) and apostrophes ('), which are part of Maltese orthography.

Table 2 provides a detailed overview of each participant in terms of the total time recorded and the number of utterances read.

3. Experiments

In this section, we describe a number of experiments in constructing ASR systems using a typical architecture consisting of language model, pronunciation model and acoustic model. These are intended to provide baseline results, demonstrating that the MHC is suitable for training ASR systems using off-the-shelf toolkits, namely, the Pocketsphinx[7] (Huggins-Daines et al., 2006) and Kaldi (Povey et al., 2011) engines. The choice of these toolkits was based on the following rationale.
MHC is a relatively small dataset compared, for example, to standard datasets for English (Panayotov et al., 2015; Roter, 2019). State of the art systems, such as DeepSpeech (Amodei et al., 2016), Wav2Letter (Collobert et al., 2016) or Espresso (Wang et al., 2019) are known to require datasets which are orders of magnitude larger than MHC. At the same time, Sphinx and Kaldi remain widely used for quick prototyping, and perform a series of transformations on input data that make them less data hungry than more sophisticated systems. Since our aim is to produce an initial set of baseline results, these toolkits were appropriate.

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

[7] Pocketsphinx is a real-time version of the CMU-Sphinx (Lamere et al., 2003).

Speaker ID	Total time	No. of Utterances
F_01	12m:57s	112
F_02	13m:27s	127
F_03	24m:19s	182
F_04	20m:48s	166
F_05	19m:35s	174
F_06	18m:11s	125
F_07	22m:56s	161
F_08	19m:29s	171
F_09	21m:15s	170
F_10	23m:16s	169
F_11	17m:46s	144
F_12	21m:18s	178
F_13	24m:10s	175
M_01	15m:20s	118
M_02	22m:02s	164
M_03	15m:18s	117
M_04	17m:19s	125
M_05	22m:26s	159
M_06	17m:57s	179
M_07	21m:26s	177
M_08	19m:17s	170
M_09	20m:05s	166
M_10	18m:02s	134
M_11	20m:08s	164
M_12	17m:18s	137

Table 2: Details for MASRI-HEADSET Corpus

3.1. Training and test data

For the purposes of these experiments, the MHC was randomly divided into training (3614 files, totalling 7h:35m:21s) and test (250 files, totalling 30m:57s) sets. Note that the corpus, as distributed, does not reproduce the split, but is distributed as one whole set. Nevertheless, one can reproduce the experiments of this paper with the help of our ”LREC2020 Experiment Files” available in our project website[8].

While the selected test set is small, it was sampled to ensure representativeness, by including 10 audio files from each of the 25 speakers represented in the corpus. The aim is thus to enable an evaluation of a relatively simple ASR model on test data with limited speaker variation (in the sense that data from all speakers is also represented in the training set).

In constructing a language model and pronunciation dictionary for the experiments reported below, none of the utterances or lexical items in the test set were included.

3.2. The Language Model

The language model was created using part of the Korpus Malti v3.0 (Gatt and Čéplö, 2013), a corpus of written or transcribed Maltese divided into different genres, including: culture, news, academic, religion, sports, etc. The corpus is annotated with part of speech information and has a size of approximately 250 million tokens. A substantial proportion of the tokens are also lemmatised.

For the purposes of these experiments, sentences containing Maltese-English code-switching or extensive borrowing from English were excluded – Engish words were identified via the CMU Pronounciation Dictionary (Weide, 1998). We also removed sentences with digits and proper names (to distinguish between proper names and other tokens, we used CIEMPIESS-PNPD (Mena, 2019)). After the selection process, we ended up with more than 28,000 sentences. In addition, we included the transcriptions from the MHC training data (3,614 in total) so as to reduce perplexity.

Using the above data, a 3-gram language model was produced using the SRI Language Modelling Toolkit.[9]

For the experiments with Sphinx, the model could be directly included in ARPA format; for Kaldi, the format was transformed using the off-the-shelf Kaldi executable `arpa2fst`.

3.3. The Pronunciation Model

Traditional ASR systems, such as Sphinx, Kaldi or HTK (Young and Young, 1993), require a pronunciation model in the form of a pronunciation dictionary. The format of these kind of dictionaries is very simple: a list of words alphabetically sorted, each of them followed by a sequence of phonemes separated by spaces. An example is shown in figure 2 for the lemma *abbanduna* ‘to abandon’ and its derivations.

```
abbanduna    ɐ b ɐ n d ʊ n ɐ
abbandunajt  ɐ b ɐ n d ʊ n ɐ ɪ t
abbandunat   ɐ b ɐ n d ʊ n ɐ t
abbandunata  ɐ b ɐ n d ʊ n ɐ t ɐ
abbandunati  ɐ b ɐ n d ʊ n ɐ t ɪ
abbandunaw   ɐ b ɐ n d ʊ n ɐ ʊ
```

Figure 2: Format of the Pronounciation Dictionary

Words in the Korpus Malti v3.0 corpus were converted into their phonetic transcription using a grapheme-to-phoneme tool (G2P), which is a new Python 3 implementation of an earlier model by Borg et al. (2011), originally built in the context of a text-to-speech system for Maltese (see Section 1.). The phoneme set used by the G2P tool is given in the Appendix, Table 5.

The resulting pronunciation dictionary contains more than 4,000 words. Table 3 shows the phoneme distribution in the MHC corpus.

Table 3 reveals that there is no point in considering G2P rules for the low-pitch vowels due to their small counts. While such vowels are known to help to improve naturalness in speech synthesis systems, their low frequency suggests they might have limited value for ASR. We leave this as a question for future work.

[8] `https://www.um.edu.mt/projects/masri/downloads.html`

[9] `http://www.speech.sri.com/projects/srilm/`

No.	Phoneme	Counts	Percentage
1	ɐ	24340	12.1790%
2	ɪ	22622	11.3193%
3	l	14133	7.0717%
4	t	14087	7.0487%
5	n	11227	5.6176%
6	ʊ	11155	5.5816%
7	ɛ	10920	5.4640%
8	m	9286	4.6464%
9	r	8597	4.3017%
10	k	7380	3.6927%
11	s	6679	3.3420%
12	ɔ	5420	2.7120%
13	h	5383	2.6935%
14	j	5285	2.6444%
15	d	5281	2.6424%
16	ɪː	4946	2.4748%
17	b	4575	2.2892%
18	f	4573	2.2882%
19	ʔ	3881	1.9419%
20	ɐː	3800	1.9014%
21	ʃ	3309	1.6557%
22	p	2877	1.4396%
23	w	1953	0.9772%
24	ʤ	1820	0.9107%
25	z	1304	0.6525%
26	ʧ	1151	0.5759%
27	ɛː	870	0.4353%
28	g	769	0.3848%
29	iː	574	0.2872%
30	v	480	0.2402%
31	ʦ	425	0.2127%
32	ɔː	359	0.1796%
33	ʤ	136	0.0681%
34	ʒ	129	0.0645%
35	à	92	0.0460%
36	è	21	0.0105%
37	ì	7	0.0035%
38	ù	6	0.0030%
39	ò	1	0.0005%

Table 3: Phoneme Distribution of the MASRI-HEADSET

3.4. Sphinx Setup

For this experiment, Sphinx was configured using standard settings: continuous Hidden Markov Models (HMM) with 3 states per HMM and 1,050 tied states. The number of tied states was determined empirically by running the training stage several times and selecting the model with the lowest Word Error Rate (WER).

As features, we computed MFCC vectors through Sphinx for all of the 13 coefficients, plus the first and second derivatives.

In addition to the above setup, we also performed additional experiments applying Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT).

3.5. Kaldi Setup

We carried out two experiments in Kaldi: one with traditional HMMs and the other with neural networks using the `nnet4d3` recipe.[10] The following processes were used during HMM training:

- Linear Discriminant Analysis (LDA)
- Maximum Likelihood Linear Transform (MLLT)
- Speaker Adaptive Training (SAT)

3.6. Results

Table 4 shows the results for all the experiments performed, providing information about the Word Error Rate (WER) and the Sentence Error Rate (SER). The overall perplexity of the language model in all experiments was 182.3744, with 87 out-of-vocabulary (OOV) words.

Experiment	WER	SER
Sphinx HMMs	19.40%	64.00%
Sphinx HMMs+LDA/MLLT	18.40%	59.60%
Kaldi HMMs	12.54%	39.60%
Kaldi NNs	**10.56%**	**37.60%**

Table 4: Results of the experiments performed

As the table shows, the best results were obtained by Kaldi with a neural network setup. It is interesting to note that the difference between Sphinx and Kaldi is at almost 10%. This is consistent with similar experiments done on Spanish (Mena et al., 2017).

4. Conclusions and future work

This paper described the design of MASRI-HEADSET (MHC), a new corpus for Maltese speech aimed to support research on automatic speech recognition (ASR). This is the first corpus of its kind for the Maltese language, for which speech resources are still somewhat limited.
Preliminary experiments with the corpus, while somewhat small-scale, suggest that high-quality acoustic models can be created in spite of the small size of the MHC, due to the low noise in the recordings. Our best results (WER of 10.56%), obtained with a neural network setup, point to the importance of performing future experiments with modern speech recognizers based on neural networks such as DeepSpeech (Amodei et al., 2016), Wav2Letter (Collobert et al., 2016) or Espresso (Wang et al., 2019).
The corpus will be freely available from our website[11] for research purposes, with additional licensing options for commercial use.
While MHC is an important first step towards achieving adequate support for Maltese speech technology, it is limited in size. Our current work is addressing this issue, as well as

[10] `nnet4d3` is an architecture composed of two p-norm layers with 1000 neurons in each layer. The number of pdfs that correspond to the outputs of the neural network is 1504 (Lim and Kim, 2019).

[11] `https://www.um.edu.mt/projects/masri/index.html`

other challenges in developing viable ASR technology for spoken Maltese, from multiple angles.

First, we are studying recent techniques for enhancing ASR in weakly supervised or unsupervised settings, with a view to maximising the use of the available data, either via pre-training (Schneider et al., 2019), or via augmentation techniques such as spectrogram perturbation (Park et al., 2019) and noise superposition (Hannun et al., 2014).

Second, a crowdsourcing effort is currently underway, based on the Common Voice initiative. Common Voice[12] (Roter, 2019) is an open-source, web-based platform created by Mozilla in order to crowd-source speech data for various languages. Language communities are encouraged to localise the Common Voice website to their language, as well as provide a minimum of 5,000 validated sentences to be used as prompts to be read out and validated by users. This process is currently underway for Maltese, with the website having been completely localised and sentence prompts already sampled from the Korpus Malti v3.0. The prompts are currently being validated for use. Once this process is complete, users will be able to contribute to the construction of a larger, community-driven dataset for Maltese by donating their voices to the project, as well as validating other users' voice clips. As part of this effort, we are also developing unsupervised techniques designed to automatically validate user-contributed data, using heuristics to automatically rank text-speech pairs according to the likelihood that a recorded voice clip matches a textual prompt.

Finally, many of the important challenges for open-domain ASR for Maltese are connected to the complexities arising from the bilingual (Maltese/English) situation described in Section 1., where a majority first language with a relatively small number of speakers (in this case Maltese) exists alongside a far better resourced and more widely-spoken language (that is, English). A well-known issue in such intensive language contact situations is the widespread incidence of code-switching and lexical borrowing (Matras, 2009). Any viable ASR system for Maltese will also need to handle Maltese-English code-switching. We aim to address this both from the data-collection perspective, through selection of data from naturalistic sources, as well as crowd-sourcing and further data-collection efforts, and from the modelling perspective as we explore transfer learning and learning from mixed language resources.

Acknowledgements

Special thanks to Ayrton Brincat for helping us with the final revision of the release transcriptions of the corpus. We also thank KPMG Microsoft Business Solutions (formerly CrimsonWing) for providing their grapheme-to-phoneme conversion rules used as the basis for the G2P tool created in this project. The MASRI Project is funded by the University of Malta Research Fund Awards. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp used for MASRI research.

5. References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182.

Azzopardi-Alexander, M. and Borg, A. (2013). *Maltese*. Routledge.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Borg, C. and Gatt, A. (2017). Morphological analysis for the Maltese language: The challenges of a hybrid system. In *Proceedings of the 3rd Arabic Natural Language Processing Workshop (WANLP'17)*, Valencia, Spain. Association for Computational Linguistics.

Borg, M., Bugeja, K., Vella, C., Mangion, G., and Gafà, C. (2011). Preparation of a free-running text corpus for Maltese concatenative speech synthesis. In *Proceedings of the 3rd International Conference on Maltese Linguistics (GhiLM'11)*, Valletta. GhiLM.

Brincat, J. (2011). *Maltese and other languages: A linguistic history of Malta*. Midsea Books, Malta.

Camilleri, J. J. (2013). A Computational Grammar and Lexicon for Maltese. Master's thesis, Chalmers University of Technology, Gothenburg, Sweden.

Čéplö, S. (2018). *Constituent order in Maltese: A quantitative analysis*. Ph.D. thesis, Charles University, Prague, Czech Republic.

Collobert, R., Puhrsch, C., and Synnaeve, G. (2016). Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*.

Comrie, B. (2009). Maltese and the world atlas of language structures. *Introducing Maltese Linguistics*, pages 3–12.

Draxler, C. and Jänsch, K. (2004). Speech recorder (version 2.2. 1). https://www.bas.uni-muenchen.de/Bas/software/speechrecorder/.

European Parliament. (2018). Language equality in the digital age (European Parliament Resolution 2018/2028(INI)). http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf.

Fabri, R. (1993). *Kongruenz und die Grammatik des Maltesischen*. Niemeyer, Tuebingen.

Gatt, A. and Čéplö, S. (2013). Digital corpora and other electronic resources for Maltese. *Corpus Linguistics 2013*, page 96.

Gatt, A. and Fabri, R. (2018). Borrowed affixes and morphological productivity: A case study of two Maltese nominalisations. In P Paggio et al., editors, *Languages of Malta: Recent Studies*, pages 143–170. Language Science Press, Berlin.

Grech, S. (2015). *Variation in English: Perception and patterns in the identification of Maltese English*. Ph.D. thesis, University of Malta, Malta.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., and Ng, A. Y. (2014). Deep Speech: Scaling up end-to-end speech recognition. *arXiv*, 1412.5567.

[12] https://voice.mozilla.org/en

Hoberman, R. D. (2007). Maltese morphology. *Morphologies of Asia and Africa*, 1:257–281.

Huggins-Daines, D., Kumar, M., Chan, A., Black, A. W., Ravishankar, M., and Rudnicky, A. I. (2006). Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.

Krauwer, S. (2003). The basic language resource kit (blark) as the first milestone for the language resources roadmap. *Proceedings of SPECOM 2003*, pages 8–15.

Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., Warmuth, M., and Wolf, P. (2003). The CMU SPHINX-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, volume 1, pages 2–5.

Lim, M. and Kim, J.-H. (2019). Integration of tensorflow based acoustic model with kaldi wfst decoder. *arXiv preprint arXiv:1906.11018*.

Matras, Y. (2009). *Language Contact*. Cambridge University Press, Cambridge, UK.

Mena, C. D. H., Ruiz, I. V. M., and Camacho, J. A. H. (2017). Automatic speech recognizers for mexican spanish and its open resources. *Journal of Applied Research and Technology*, 15(3).

Mena, C. (2019). The ciempiess proper-names pronouncing dictionary. In *Corpus presented at OpenCor 2019 Conference, Guanajuato City, Mexico.[Available online at https://opencor.gitlab.io/corpora-list/]*.

Micallef, P. (1997). *A text to speech synthesis system for Maltese*. Ph.D. thesis, University of Surrey.

Mifsud, M. (1995). *Loan verbs in Maltese: A descriptive and comparative study*. Brill, Leiden.

NSO. (2019). *Regional Statistics (Malta), 2019 Edition*. National Statistics Office, Malta.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *arXiv*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Ravishankar, V., Tyers, F., and Gatt, A. (2017). A Morphological analyser for Maltese. *Procedia Computer Science (Special Issue on Arabic Computational Linguistics)*, 117:175–182.

Rosner, M. and Joachimsen, J. (2012). *The Maltese Language in the Digital Age*. Springer, Berlin and Heidelberg.

Roter, G. (2019). Sharing our Common Voices – Mozilla releases the largest to-date public domain transcribed voice dataset. https://blog.mozilla.org/blog/2019/02/28/sharing-our-common-voices-mozilla-releases-the-largest-to-date-public-domain-transcribed-voice-dataset/, February.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised Pre-training for Speech Recognition. In *Proceedings of Interspeech 2019*, pages 1–9, Graz, Austria.

Tiedemann, J. and van der Plas, L. (2016). Bootstrapping a Dependency Parser for Maltese - A Real-World Test Case. In Martijn Weiling, et al., editors, *From Semantics to Dialectometry : Festschrift in honor of John Nerbonne*. College Publications, London.

Vella, A. (1994). *Prosodic structure and intonation in Maltese and its influence on Maltese English*. Ph.D. thesis, University of Edinburgh, Scotland.

Wang, Y., Chen, T., Xu, H., Ding, S., Lv, H., Shao, Y., Peng, N., Xie, L., Watanabe, S., and Khudanpur, S. (2019). Espresso: A fast end-to-end neural speech recognition toolkit. *arXiv preprint arXiv:1909.08723*.

Weide, R. L. (1998). The CMU pronouncing dictionary. http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/.

Young, S. J. and Young, S. (1993). *The HTK hidden Markov model toolkit: Design and philosophy*. University of Cambridge, Department of Engineering Cambridge, England.

Zammit, A. (2018). A dependency parser for the Maltese language using deep neural networks. Master's thesis, Department of Artificial Intelligence, University of Malta, Msida, Malta.

Appendix A: The Maltese Phonology

Consonants	Bilabial	Labiodental	Alveolar	Postalveolar	Palatal	Velar	Glottal
Plosive (voiceless)	p		t			k	ʔ
Plosive (voiced)	b		d			g	
Affricate (voiceless)			ʦ		ʧ		
Affricate (voiced)			ʣ		ʤ		
Fricative (voiceless)		f	s	ʃ			h
Fricative (voiced)		v	z	ʒ			
Nasal (voiced)	m		n				
Lateral Approximant (voiced)			l				
Trill (voiced)			r				
Approximant (voiced)					j	w	
Vowels			Front		Central		Back
Close			i				
				ɪ		ʊ	
Close-mid							
Open-mid				ɛ			ɔ
					ɐ		
Open							
Long Vowels			Front		Central		Back
Close			iː				
				ɪː		ʊː	
Close-mid							
Open-mid				ɛː			ɔː
					ɐː		
Open							
Low-Pitch Accented Vowels			Front		Central		Back
Close			ì				
						ù	
Close-mid							
Open-mid				è			ò
					à		
Open							

Table 5: The Phonological System of the Maltese Language