Resources and tools for pre-processing speech data in a lesser-known variety of English

Alexandra Vella[1], Sarah Grech[1], Ian Padovani[2], Maria-Christina Micallef[3]

[1]University of Malta, Malta; [2]Utrecht University, The Netherlands; [3]Uppsala University, Sweden
{alexandra.vella, sarah.grech}@um.edu.mt, i.padovani@students.uu.nl, maria-christina.micallef.8671@student.uu.se

ABSTRACT

Research on lesser-known language varieties can be hindered from the outset by the need for both data and tools for automating the required pre-processing work. For speech, whilst more ecologically valid data in the form of video and audio are sometimes available, these need to be accompanied by a machine-readable text, ideally segmented and labelled, both allowing for searchability. A significant initial commitment is needed even before the relevant phonetic and phonological research can begin. This paper demonstrates and evaluates the efficacy of already existing tools (YouTube captioning and WebMAUS forced alignment) in automating the pre-processing work required using Maltese English (MaltE), whilst also showcasing a sample analysis of the pronunciation of post-vocalic 'r' in the variety. As a low resource variety of English, MaltE presents a test case for showing how existing resources and tools can be utilised to work with language varieties which are digitally less well-supported.

Keywords: resources & tools, low resource, phonetics & phonology, Maltese English, variety of English

1. INTRODUCTION

A fair amount of manual effort is needed both to collect spoken data and to make such data ready for analysis, phonetic and/or phonological as well as otherwise [23]. Whilst this is the case for all languages, resources such as corpora and an increasingly wide range of automatic speech processing tools are available for so-called high resource languages such as English, Chinese and Spanish; Japanese and a few other European languages such as French can also boast an increasing wealth of resources and tools [12]. Low resource languages (see [6] for definition) trail behind, although the situation continues to improve. [19] report on their success in using WebMAUS to carry out forced alignment of data from a number of language documentation corpora whilst [14] survey research which shows that careful matching of a high to a low resource language can result in direct transfer of good performance to a new language using an existing model without any additional training. Similarly, [13] show that, whilst manual checking is still likely to be needed to correct errors, forced alignment software developed for American English performed well for six varieties of British English.

In this paper, we report on our application of two tools, YouTube's automatic captioning and the CLARIN-D web application WebMAUS, for pre-processing available data from Maltese English.

2. BACKGROUND

2.1. The Maltese English variety

Maltese English, MaltE, the variety of English of bilingual speakers of Maltese and English [22], can be said to be a lesser-known variety of English [17]. Whilst characterised by a great deal of both intra- and inter-speaker variability [4, 24], it is also readily identified by its speakers and others familiar with the variety [7, 18].

The accent of MaltE is particularly distinct in spite of the variability. Amongst the phonetic/phonological characteristics which have been noted [8, 24] are: (i) substitution of θ and ð; (ii) pronunciation of a velar plosive following ŋ; (iii) variability in the degree and nature of post-vocalic 'r' pronunciation; (iv) no clear/dark 'l' distinction; (v) aspiration and audible release of plosives; (vi) intervocalic voicing and final devoicing; (vii) vowel quality and durational differences (especially in the pronunciation of SSBE æ and ɜː, as well as ə); (viii) lack of vowel weakening and of syllabic consonants; (ix) differences in lexical, compound and phrase stress; (x) distinct rhythm and intonation patterns.

In spite of its distinctness, empirical evidence for many of the phonetic/phonological characteristics of MaltE is scarce, partly due to the lack of pre-processed language corpora which, as [19] point out, provide an extremely good starting point for much general and applied linguistic research, sociophonetics included, with usefulness increasing the more fine-grained the temporal alignment of the transcription is. Moreover, although MaltE is spoken in daily life by most Maltese speakers, increasingly so as Malta becomes more "cosmopolitan" [24], collecting ecologically valid data from this variety is

never straightforward. Looking towards available data which can be pre-processed easily to give researchers access to data accompanied by a text, as well as by word and phone level time-aligned segmentations, would definitely be a step in the right direction. However, anecdotes of difficulties that technologies such as speech recognition systems have with different accents are frequent (see [15], but also [5]). It is therefore not clear whether and to what extent available tools can deal with data from MaltE, although reports of the sort in [13] are encouraging.

2.2. YouTube automatic captioning

Little information on the backend to YouTube's automatic captioning tool is available on the YouTube Help page [25]. It includes the caveat that "automatic captions may misrepresent the spoken content due to mispronunciations, accents, dialects or background noise". One might expect the variability in MaltE to exacerbate such difficulties resulting in a decrease in accuracy. On the whole, however, whilst the presence of all the usual elements of spontaneous speech including various discourse markers, normal disfluencies, and, in dialogue data, overlaps, make the captioning process less straightforward, the interface is generally user-friendly. Unlike other tools such as Google's Speech-to-Text API, YouTube's automatic captioning tool is accessible to people without prior programming knowledge, thus making it available to a wider pool of researchers (as well as other users).

2.3. WebMAUS

The ***M**unich **AU**tomatic **S**egmentation Service* **MAUS**, developed by the Bavarian Archive of Speech Signals and the corresponding CLARIN-D WebMAUS [9, 10, 16, 20] service uses acoustic models combined with pronunciation rules and, in some cases, a language model, as the basis for providing word and phone time-aligned segmentations given a) a speech signal; and b) an accompanying orthographic transcript [21]. Once again, it differs from other equivalent tools such as the Montreal Forced Aligner [26] in not requiring programming knowledge.

The WebMAUS service is available for more than 25 languages, including a number of dialects. It is currently available for five varieties of English: Australian (AU), American (US), British (GB), Scottish (SC) and New Zealand (NZ). The documentation on the different parameter sets suggests differences in the acoustic and pronunciation modelling and in the nature and amount of manually segmented and labelled data used for training. The results of using untrained forced alignment are however promising, see [19] and also [5], and, although [14] report on the importance of the need to select a high resource language or language variety to apply to a low resource one with care, they also note good performance if the match is well made.

At the start of this work, we tested WebMAUS's English (GB), (SC) and (US). Though Maltese speakers of English have come to be increasingly exposed to American English, British English continues to be favoured, particularly in education [24]. Australian and New Zealand English, particularly their vowel systems, present differently enough not to be in contention. After testing, it was found that the rhoticity of Scottish English served to minimise segmentation errors, and, to some extent also labelling errors. WebMAUS forced alignment was therefore carried out using English (SC).

3. METHODOLOGY

3.1. The data

For the purposes of this paper, we focus on a small amount of data taken from CoSME, a Corpus of Spoken Maltese English which is being collected from a variety of available sources, e.g. online platform or YouTube material, recordings of radio or television programmes, public, including University-based events, etc. Specifically, the data we report on here are: (1) a graduand monologue; and (2) a TimesTalk (dialogue) interview involving a MaltE interviewer and a MaltE singer, Ira Losco. We report on the application of tools available for use with more mainstream varieties of English to these MaltE data. The procedure used involved creating a text using the YouTube automatic captioning tool, followed by WebMAUS forced alignment segmentation and labelling of the relevant sound files.

The first requirement in the pre-processing of spoken data, whether in video or audio format, involves creating a text. This is because most automatic segmentation and labelling tools (specifically in this case WebMAUS) require a sound file together with an accompanying text. Our procedure, broadly based on [3], involves the following:

Step 1 – Use YouTube captioning to create a first instance of a text.
Step 2 – Use established conventions to "clean up" the text.
Step 3 – Make any boundary adjustments and add other details as necessary.
Step 4 – Run WebMAUS on the cleaned up version of the text created using YouTube to generate a Praat TextGrid containing the automatically generated segmentation/labelling.

In (the relatively rare) cases when a video or audio recording comes with a ready transcript, step 1 can be

skipped: steps 2 and 3 are nevertheless still necessary in order to ensure that the accompanying transcript is as "loyal" to what is in the actual audio as possible, thus minimising WebMAUS errors.

In the case of the dialogue data, we have not, so far, separated the speech signal into chunks on the basis of speaker (this is however something we intend to do as we continue with this work).

Overlap is marked in post-editing of the captions text and then removed from the original sound file and accompanying transcript using a simple Python script. In our pipeline, periods marked as overlap of any length are replaced by brief silence. This allows WebMAUS to annotate multiple voices in a single channel without the difficulties arising from overlap.

3.2. Using YouTube captioning

In order to use YouTube to automatically generate closed captions, audio must be uploaded to a YouTube account in video format. In the case of audio-only data, a still image was added to the sound file and the result saved in video (e.g. MP4) format.

YouTube automatically detects the language and generates captions accordingly. These can be manually edited in the available interface, and mistakes that occur, due to variety differences or for other reasons, can be corrected. Moreover, for the purpose of preparing the text for further processing, the more accurate the representation of an utterance is, the better. [19] report that the quality of the forced alignment is better when elements of spontaneous speech in the signal are included in the transcript. The closed captions were therefore manually edited by our annotators to include such elements. A Python script containing simple rules was also used to clean up other elements in the text such as the timestamps that appear in the subtitles, before passing it through WebMAUS. This was done to eliminate the risk that these would be interpreted as actual words in the audio signal.

3.3. Using WebMAUS

The edited YouTube close-captioned texts were saved as .txt and uploaded together with original audio files to WebMAUS. Forced alignment was carried out using English (SC) and a segmented and labelled Praat TextGrid [1] was generated. The ORT-MAU and MAU tiers were duplicated and edited manually by our annotators, see Fig. 1.

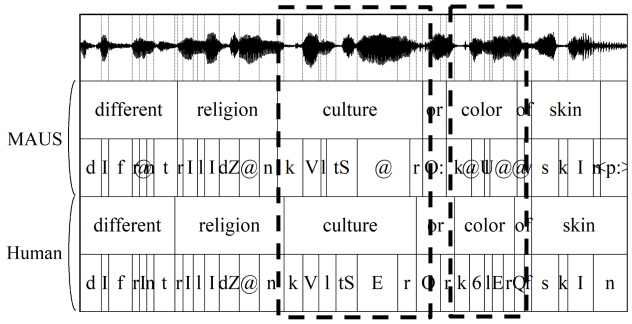

Figure 1: An example of an excerpt from a Praat TextGrid containing WebMAUS automatically generated word and phone level segmentation/labelling, together with the manual correction of the automatic output in the (bottom two) duplicated tiers.

4. EVALUATION OF TOOLS USED

4.1. Some general points

YouTube is quick at generating closed captions, typically taking from 30 minutes up to a few hours, depending on the file duration. The manual "cleaning up" required once the captions have been generated is also fairly quick, taking one of the annotators roughly an hour to cover 15 minutes of recording, although dialogue videos with high degrees of overlapping speech take more time. Still, creating a first instance text using this procedure takes far less time than full manual processing by human annotators.

[19] suggest that a "real-time factor up to 400" may be involved in manual segmentation and labelling of audio data at word and phone level. It took the annotator significantly less time to correct WebMAUS output.

4.2. YouTube captioning

Unlike the language documentation corpora data reported on in [19], recordings from available sources as outlined in 3.1 do not come with an accompanying text. In the case of MaltE, YouTube's closed captioning tool provides a surprisingly accurate starting solution.

A Word Error Rate (WER) analysis of our unedited (Orth) and edited (Clean) YouTube closed captioned texts is shown in Table 1.

Audio	Duration	Orth WER	Clean WER
Graduand	270	0.42%	1.55%
Interview	557	2.93%	6.99%

Table 1: WER scores for the data analysed. *Orth WER* compares to a plain orthographic transcription. *Clean WER* compares to a transcript including expanded acronyms, digits, added fillers etc.

Errors related to specific features of the variety do however occur. An example which arises from a feature of the pronunciation of MaltE, the substitution

of /ð/ with [d] (or [d̪], see feature i in 2.1 above), resulted in closed captioning of *though no* as *don't know*. Another example of a relatively common error type involves words such as *flags*, citation form /flægz/, being captioned as *flax* [flæks] a "mistake" likely due to final devoicing in Maltese English (see feature vii) or *flecks* [flɛks]: this is the result of substitution of /æ/ with [ɛ] (see feature viii).

As reported elsewhere, e.g. [14], proper nouns such as names and places often cause problems: the proper noun *Losco* was close-captioned as the (relatively) phonetically close approximation, *lost cause*. Errors related to detection of name and place entities in speech are not specific to the variety, see [14]. YouTube close-captioned *a new way* as *anyway* possibly for reasons related to speech rate. This is another issue not specific to MaltE, see [19].

4.3. WebMAUS

To evaluate the effectiveness of this step in our procedure, we ran WebMAUS on 270 seconds of data taken from the beginning, middle and end of a MaltE graduand monologue, (1) above, and an entire interview of 9 minutes and 17 seconds, involving two MaltE speakers, (2) above. Manual correction was carried out by our annotators. Boundaries at word beginnings and endings often needed adjusting (see Fig. 1) but segmentation was generally good. Discrepancies were more frequent at phone level, not just in the case of boundaries, but also with respect to phone labelling, a matter for further investigation.

Inter-annotator WebMAUS vs human annotator agreement was measured by means of a Cohen's Kappa score [11] based on comparison of the automatic and manual annotations. The results of the comparison for the graduand monologue and the interview data are shown in Table 2:

Sound File	Duration	Cohen's Kappa
Graduand	270	0.79
Interview	557	0.61

Table 2: Cohen's Kappa scores for the data analysed.

To utilise this metric, the audio and accompanying transcriptions were divided into non-overlapping frames 2 msecs in duration. Cohen's Kappa checks each frame to see if the segmentation and labelling of the phones in the two transcripts match or not, also taking into account the likelihood of them only matching by chance. As can be seen from Table 2, the agreement score for the two text-types is 0.79 and 0.61 respectively, indicating substantial agreement between WebMAUS and the human annotator. The score is higher for the graduand monologue data. This may be partly due to differences between scripted and unscripted speech. The former tends to display fewer normal disfluencies, and slower, more deliberate speech, making the latter more difficult to segment and label by comparison. Another possible reason for the difference is that different annotators were involved in annotating the data from the two sources, the latter receiving a more detail-oriented treatment.

4.4. WebMAUS vs human annotation of 'r'

Comparison of the WebMAUS vs human annotator data provides a wealth of information on the phonetic detail in the analyses. Phone substitutions made by the MaltE speakers clearly needed to be manually corrected. Characteristics of MaltE such as relatively heavy aspiration, including in word-final position, often resulted in the right-edge boundary being placed earlier by MAUS as compared to by the human annotator, the latter including the (often relatively long) period of aspiration as part of the relevant segment.

A characteristic we have started to examine in depth, but which we have minimal space to discuss here, is the MaltE realisation of 'r'. This, (see [2]) is shifting from non-rhotic to (more) rhotic. Examining the .csv file which served as the basis for the Cohen's Kappa analysis is interesting on several counts. It shows that the replacement rules employed in the English (SC) WebMAUS may not always generate truly Scottish English output. Fig. 1 (see the broken line box) shows an example in the labelling of *color* (YouTube spelling). WebMAUS does not include an 'r' following the vowel in spite of the rhoticity one would expect, also given the vowel following in *of skin*. Another example in Fig. 1 (see the heavy broken line box), *culture*, also includes a post-vocalic 'r'. Both cases involve an [ɛ] vowel similar in quality to Maltese /ɛ/ but shorter than what one would expect for SSBE, a sort of compensatory shortening. Further, in the interview data, 120 of the 215 instances of 'r' are post-vocalic. The human annotator labelled clear 'r' phones in all of these: WebMAUS gave only 15 'r' phones in this position. Two of the authors of this paper continue in the sociophonetic analysis of the MaltE pronunciation of 'r', including post-vocalically. Approximant, tap, trill and also affricated realisations have been noted, although the precise details of such realisations are still to be determined. Using data pre-processed as described above would provide a highly useful fast track entry point to the analysis.

5. CONCLUSION

In conclusion, we have shown that tools are available which allow researchers (and others), regardless of programming experience, to curate and pre-process

MaltE data, providing them with easy access to word and phone level time-aligned segmentations of ready data for analysis. As processing power, machine learning and natural language processing techniques continue to become more powerful, the time seems ripe to do all that is possible to increase the reach of already available tools to low resource languages and language varieties. User-friendly tools such as those we have described help bridge the gap allowing easy access to newcomers to these fields. We have shown that use of YouTube to create a first instance text to accompany a video/audio recording doubtlessly adds a welcome measure of automation at the early stages of transcription. Similarly, applying WebMAUS to MaltE gives good results, although manual correction remains necessary. Equally importantly, more in depth analysis of the accuracy of the output of both the closed captioning and forced alignment tools, and of the manual adjustments made by the human annotators, is likely to lead to insights of all sorts.

6. REFERENCES

[1] Boersma, P., Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 17 Dec 2022, http://www.praat.org/

[2] Bonnici, L. M. 2010. *Variation in Maltese English: The interplay of the local and the global in an emerging postcolonial variety*. PhD thesis, University of California.

[3] Cangemi, F., Fründt, J., Hanekamp, H., Grice, M. 2019. A semi-automatic workflow for orthographic transcription and syllabic segmentation. In: Piccardi, D., Ardolino, F., Calamai, S. (eds), *Gli Archivi Sonori al Crocevia tra Scienze Fonetiche, Informatica Umanistica e Patrimonio Digitale.* [*Audio Archives at the Digital Crossroads of Speech Sciences, Digital Humanities and Digital Heritage*.], Officinaventuno, Milano, 419-425.

[4] Caruana, S., Mori, L. 2021. Rethinking Maltese English as a continuum of sociolinguistic continua through evaluations of written and oral prompts. *English World-Wide. A Journal of Varieties of English* 42(3), 245-272.

[5] Choe, J., Chen, Y., Chan, M.P.Y., Li, A., Gao, X., Holliday, N. 2022. Language-specific effects on automatic speech recognition errors for World Englishes. *Proc. 29th International Conference on Computational Linguistics*, 7177-7186.

[6] Cieri, C., Maxwell, M., Strassel, S., Tracey, J. 2016. Selection criteria for low resource languages. *Proc. 10th International Conference on Language Resources and Evaluation (LREC '16)*, 4543-4549.

[7] Grech, S. 2015. *Variation in English: Perception and patterns in the identification of Maltese English*. PhD thesis, University of Malta.

[8] Grech, S., Vella, A. 2019. Rhythm in Maltese English. In: Paggio, P., Gatt, A. (eds), *The Languages of Malta*. Berlin: Language Science Press, 203-223.

[9] Kisler, T., Reichel, U. D., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech and Language* 45, 326-347.

[10] Kisler, T., Schiel, F., Sloetjes, H. 2012. Signal processing via web services: The use case WebMAUS. *Proc. Digital Humanities*, Hamburg, 30-34.

[11] Kolesnyk, A. S., Khairova, N. F. 2022. Justification for the use of Cohen's Kappa statistic in experimental studies of NLP and text mining. *Cybernetics and Systems Analysis* 59(2), 280-288.

[12] Laumann, F. 2022. Low-resource language: what does it mean? https://medium.com/neuralspace/low-resource-language-what-does-it-mean-d067ec85dea5

[13] MacKenzie, L., Turton, D. 2020. Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard* 6 s1.

[14] Magueresse, A., Carles, V., & Heetderks, E. 2020. Low-resource languages: A review of past work and future challenges. arXiv preprint arXiv:2006.07264

[15] Reid, M., MacGregor, G. 2020. There can be no true Scottish speech recognition systém. https://jabde.com/2020/11/14/no-true-scottish-sls/

[16] Schiel, F. 1999. Automatic phonetic transcription of non-prompted speech. *Proc. 14th ICPhS* , San Francisco, 607-610.

[17] Schreier, D., Trudgill, P., Schneider, E. W., Williams, J. P. 2010. *The Lesser-Known Varieties of English: An introduction*. Cambridge: Cambridge University Press.

[18] Stilon, E. 2018. Perceptions of Maltese English: An experimental study. MA dissertation, University of Malta.

[19] Strunk, J., Schiel, F., Seifart, F. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. *Proc. 9th International Conference on Language Resources and Evaluation (LREC '14)*, 3940-3947.

[20] The **M**unich **AU**tomatic **S**egmentation System **MAUS**.https://www.bas.uni-muenchen.de/Bas/BasMAUS.html

[21] Tour de CLARIN: WebLicht and WebMAUS. https://www.clarin.eu/blog/tour-de-clarin-weblicht-and-webmaus

[22] Vella, A. 2012. Languages and language varieties in Malta. *International Journal of Bilingual Education and Bilingualism* 16, 532-552.

[23] Vella A., Grech, S. 2021. What can a corpus tell us about phonetic and phonological variation? In: O'Keeffe, A., McCarthy, M. J. (eds), *The Routledge Handbook of Corpus Linguistics*. Routledge, 281-295.

[24] Vella, A., Grech, S. Submitted. English in Malta.

[25] YouTube Help. https://support.google.com/youtube/answer/6373554?hl=en

[26] McAuliffe, Michael, Michaela Socolof, Elias Stengel-Eskin, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger (2017). Montreal Forced Aligner [Computer program]. Version 1.0.