MARK BORG, KEITH BUGEJA, COLIN VELLA, GORDON MANGION & CARMEL GAFÀ (MALTA)

Preparation of a free-running text corpus for Maltese concatenative speech synthesis*

Taqsira

Sistemi ta' sinteżi tat-taħdit jeħtieġu korpus ta' diskors b'karatteristiċi fonetiċi u prożodiċi li huma rappreżentattivi tal-ilsien Malti. Dan l-istudju jippreżenta metodu ġdid li, b'mod awtomatiku, joħloq korpus ta' testi bil-Malti li fuqu jinbena repożitorju diġitali ta' taħdit. Korpus ta' madwar 33 miljun kelma nġabar minn siti elettroniċi, gazzetti, kotba u dokumenti uffiċjali; wara li dan tnaddaf u nqaleb għal rappreżentazzjoni fonetika (minn grafemi), ġie analizzat bl-għan li tinħareġ statistika li tgħin biex ikunu jistgħu jinqabdu l-ħsejjes kollha li hemm bżonn għall-prosodija ta' ilħna sintetiċi. B'hekk, ġie magħżul korpus iżgħar li, għaldaqstant, xorta jiġbor fih il-karatteristiċi u l-ħsejjes kollha tal-korpus il-kbir. Il-metodu li ntuża biex bih jinħoloq dan il-korpus huwa deskritt bir-reqqa f'din il-pubblikazzjoni. Il-kwalità tal-korpus hi mill-aqwa meta mqabbla ma' korpora maħluqa b'metodi oħra, inkluża l-għażla manwali.

1. Introduction

Text-to-speech systems based on concatenative speech synthesis employ the use of databases of recorded utterances which are strung together to produce speech output. The corpus of recorded speech is segmented into units of concatenation such as individual phones or diphones, and is often read from a training text compiled to provide a high degree of coverage of these basic units. The quality of output speech is highly dependent on the unit coverage of the speech database (Kominek & Black 2003) and in order to provide sufficiently natural speech output, large databases of recorded utterances are often required, spanning tens of hours (Kawai & Tsuzaki 2002). In automatic unit selection methods, a speech database is queried at runtime to find the best units to synthesize desired speech.

* This work was supported by the Foundation for Information Technology Accessibility (FITA) and Operational Programme I – Cohesion Policy 2007–2013, part-financed by the European Regional Development Fund (ERDF), at a co-financing rate of 85% EU funds and 15% National Funds. However, this paper does not necessarily represent the opinion of these entities, and they are not responsible for any use which may be made of its contents.

Often the training text is randomly sampled from a large corpus and no optimizations are applied towards the extraction of an optimal sample (Santen & Buchsbaum 1997). Nevertheless, when building a database for an open domain application, recording every possible speech event from a random selection of sentences is practically impossible (Bozkurt et al. 2003).

In our study, we consider the diphone as the basic unit of concatenation for speech synthesis for a number of reasons. A diphone is a unit which starts from the stable region (middle) of one phone and extends to that of the next phone, thus also allowing acoustic information on the transition between phones to be captured. The stable regions around the diphone boundaries simplify concatenation of such units at the speech signal level (Laws 2003). Moreover, the diphone as a unit allows for reasonable coverage of the language's phonetic content while retaining inexpensive database construction. We avoid longer unit sizes such as triphones because full coverage is harder to achieve due to a combinatorial explosion in the number of units. Moreover, we do not consider half-phonemes because although coverage is simplified, a larger unit size is required for high quality synthetic speech (Bozkurt et al. 2003).

In this paper we present a novel search function used to maximize diphone coverage when choosing a training source text for utterance recording. We discuss preparation of the corpus in section 2, followed by the statistical analysis of its phonemic and prosodic content in section 3. We then describe our method for free text selection in section 4, a method that we subsequently evaluate in section 5. Finally, we present our conclusions and suggestions for further work.

2. Preparation of the corpus

The text corpus used in this study was acquired from newspapers, websites, official documents and books written in Maltese. Notwithstanding, the diverse nature of these texts required us to normalize them into a homogeneous corpus that could be easily analyzed. This process spanned two broad stages: text cleaning and grapheme-to-phoneme conversion.

Standard Maltese operates with a system of 24 consonantal phonemes (if [dz] is given full phonemic status) and 11 vocalic sounds. Furthermore, there are 7 diphthongal segments, each composed of one of the eleven vocalic sounds together with an [ɪ] or [ʊ] (Borg & Azzopardi-Alexander 1997). Please refer to Tables 1 and 2.

/p/	/b/	/m/	/t/	/d/	/n/	/k/	/g/	/ʔ/ (q)
/f/	/v/	/s/	/z/ (ż)	/ʃ/,/ʒ/ (x)				
/ʧ/ (ċ)	/ʤ/ (ġ)	/ts/, /dz/ (z)						
/h/ (ħ)								
/j/	/w/	/l/	/r/					

Table 1: Consonantal phonemes with orthographic correspondences in brackets in non-obvious cases

Monophthongs			Diphthongs	
Orthographic	Phonetic Realization		Orthographic	Phonetic Realization
	Short	Long		
a	ɐ	ɐː	aw, għu	ɐʊ
e	ɛ	ɛː	aj, għi	ɐɪ
i	ɪ	iː	ew	ɛʊ
o	ɔ	ɔː	ej, għi	ɛɪ
u	ʊ	uː	iw	ɪʊ
ie		ɪː	oj	ɔɪ
			ow, għu	ɔʊ

Table 2: The eleven vocalic sounds and seven diphthongal segments of Maltese with indication of orthographic correspondences

Maltese is written in the Latin alphabet; nevertheless, due to the use of a number of characters, namely *ċ*, *ġ*, *ħ*, and *ż*, involving the use of diacritics, it cannot be fully represented using an ASCII character map. This was the source of some confusion when third parties independently developed fonts without agreeing on any standard (Dalli 2000). As a consequence, some texts required conversion from these legacy encodings to the Unicode (UTF-8) standard.

Moreover, while the source texts comprising the corpus are for the most part verbal ones, they nevertheless also contain other elements, such as numbers, dates, email addresses and abbreviations. In order to be properly handled, such elements require the use of a semiotic class analyzer to generate the associated verbalizations. Source texts may also contain words whose phonetic form cannot be realized correctly at the grapheme-to-phoneme stage; these include surnames and foreign words which do not follow pronunciation rules for Standard Maltese. A decision was taken to filter out these exceptional cases for the purposes of the analysis reported here.

2.1. Text cleaning

The leading motivation behind the text cleaning stage is that of compiling a homogeneous text corpus from source texts to be used in the grapheme-to-phoneme stage. The sources used in the composition of our corpus came in a variety of formats and encodings.

Predominantly, texts like the Parliamentary Debates were embedded in Microsoft Word documents and encoded in UTF-8, while online newspapers were embedded in HTML and encoded in ASCII, with extended graphemes represented by HTML codes.

Thus, in the text cleaning stage, the text sources are converted to UTF-8 text files, and in the process a number of filters are applied, which:

1. remove known acronyms and abbreviations through lookup in an exceptions file;
2. detect unknown abbreviations and initials;
3. detect foreign and alphanumeric words.

As a result of applying the text cleaning filters just mentioned, the size of the corpus was reduced by approximately 4.1%, ending with a final corpus size of just over 33 million words (see Table 3).

	Text Source	Number of words	Number of normalized words
1	Il-Bibbja (The Bible)	633,373	633,305
2	Maltese Wikipedia	1,051,510	955,275
3	Newspapers	12,604,153	12,212,885
4	Parliament Debates	20,094,864	19,166,440
5	Maltese Books	144,549	140,968
	Total:	34,528,449	33,108,873

Table 3: Text sources

In the final phase of text cleaning, the text is segmented into phrases, using a heuristic approach based on punctuation marks. For the purpose of this study, we did not make a distinction between different types of phrase breaks (e.g. as in phrases separated by commas versus those separated by end-of-sentence markers). Moreover, each phrase is classified into one of three categories, depending on whether it is a statement, a question or an exclamation.

2.2. Grapheme to phoneme

Many different strategies and algorithms have been adopted over the years for the process of grapheme-to-phoneme (G2P) conversion, ranging from rule-based approaches and finite state transducers, to data driven machine-learning algorithms based on neural networks, HMMs, etc. (Divay & Vitale 1997). Compared to languages such as English, Maltese is a fairly homographic language and thus tends to exhibit a one-to-one correspondence between most of the orthographic symbols (the graphemes) and the sounds they represent (the phonemes). For this reason, using a set of context-sensitive rewrite rules is generally sufficient for the phonemic transcription of Maltese text.

The set of G2P rules adopted here is based on previous work by Micallef (1998) and Farrugia (2005). The set of rules used is listed in Table 7. Most of the rules define a

straightforward mapping between a letter and its corresponding phoneme.[1] In the case of the historical consonant represented by the digraph *għ*, while this is normally silent, it can change the pronunciation of neighboring letters (e.g. *għuda* → /ɔʊdɐ/ [2], English 'wood'), lengthen adjacent vowels (e.g. *għadu* → /ɐːdʊ/, English 'enemy') (Hume et al. 2009), or can be voiced as /h/ in certain situations such as when in word-final position (e.g. *qlugħ* → /ʔlʊːh/, English noun 'sails') or when occurring together with the letter *h* (e.g. *magħhom* → /mɐhhɔm/, English 'with them'). Similar behavior is also exhibited by the normally-silent consonant *h*.

A number of G2P rules encode the effect of consonant devoicing that occurs in word-final position or when in a certain consonant cluster in word-medial positions (e.g. *bieb* → /bɪːp/, English 'door'). For a certain limited number of words, the consonants *x* and *z* are mapped to /ʒ/ and /dz/ respectively, rather than the normal /ʃ/, /ts/ (e.g. *xbejba* → /ʒbɛɪbɐ/, English 'maiden'; *mezzi* → /mɛdzɪ/, English 'methods'); the G2P rules handling these cases are activated based either on context or on a pre-defined word list.

Diphthongs in Maltese can have at least two possible phonetic realizations; both are considered correct and are in nearly equal use. For example, *tiegħi* (English 'mine') can be realized as /tɪːɛɪ/ or /tɪːɐɪ/; the one adopted for the G2P process was selected based on the authors' consensus.

Previous research in the area of Maltese speech synthesis (Micallef 1998), appears to show that when grave accents occur on long vowels in stressed open syllables (e.g. the word-final vowel *è* in *kafè*, English 'coffee'), these can be approximated by normal long vowels only to a certain degree, because of slight differences in certain acoustic features. As a practical measure, these accented vowels were treated independently, on a par with other elements of the phonemic inventory. Hence /kɐfè/, instead of /kɐfɛː/. A problem that can occur during the phonetic transcription of such cases is caused by the fact that these accented vowels can be written in three ways: (i) explicitly using accented vowels (*kafè*); (ii) with an apostrophe following the stressed vowel (*kafe'*); (iii) or left unmarked (*kafe*). Complicating things further, an apostrophe after a word-final vowel is also used to indicate the presence of the silent *għ* at word-final position, e.g. *laqa'*, English 'he received'. Discriminating and handling ambiguous cases of this sort requires a combination of specific G2P rules and lexicon-based information.

The set of G2P rules are implemented in terms of regular expressions and applied to the input text starting from the most specialized rules, then followed by the generic ones. A silence phoneme (represented by /#/) is used to mark the phrase breaks detected by the phrase segmentation algorithm described in the previous section. The input text

---

[1] In this document, letters or words in their orthographic form are written in italics, while their phonemic equivalent are represented in regular font style and enclosed within slashes (/.../).

[2] Since the transcription examples given in this document are used in the context of a discussion on G2P rules, all transcriptions are given in / / irrespective of whether the level of transcription is a more phonetic rather than a phonemic one.

is then scanned left to right, and replaced with its phonemic transcription. An example run of the G2P process is illustrated in Figure 1.

Input text	G2P Rules	Phonemic transcription	Rule no.
Żewġ dgħajjes bla qlugħ ▲	/ ż / → z	z	(104)
Żewġ dgħajjes bla qlugħ ▲	/ e w / → ɛ ʊ	z ɛ ʊ	(6)
Żewġ dgħajjes bla qlugħ ▲	/ ġ / ċ,f,ħ,k,p,q,s,t,x,_ → ʧ	z ɛ ʊ ʧ	(56)
Żewġ dgħajjes bla qlugħ ▲	/ d / → d	z ɛ ʊ ʧ d	(47)
Żewġ dgħajjes bla qlugħ ▲	consonant / għ / vowel,j →	z ɛ ʊ ʧ d	(52)
Żewġ dgħajjes bla qlugħ ▲	/ a j / → ɐ ɪ	z ɛ ʊ ʧ d ɐ ɪ	(3)
Żewġ dgħajjes bla qlugħ ▲	/ j / → j	z ɛ ʊ ʧ d ɐ ɪ j	(68)
Żewġ dgħajjes bla qlugħ ▲	/ e / → ɛ	z ɛ ʊ ʧ d ɐ ɪ j ɛ	(36)
Żewġ dgħajjes bla qlugħ ▲	/ s / → s	z ɛ ʊ ʧ d ɐ ɪ j ɛ s	(87)
Żewġ dgħajjes bla qlugħ ▲	/ b / → b	z ɛ ʊ ʧ d ɐ ɪ j ɛ s b	(41)
Żewġ dgħajjes bla qlugħ ▲	/ l / → l	z ɛ ʊ ʧ d ɐ ɪ j ɛ s bl	(71)
Żewġ dgħajjes bla qlugħ ▲	consonant / a / _ → ɐ: (single syllable word)	z ɛ ʊ ʧ d ɐ ɪ j ɛ s blɐ:	(15)
Żewġ dgħajjes bla qlugħ ▲	/ q / → ʔ	z ɛ ʊ ʧ d ɐ ɪ j ɛ s blɐ: ʔ	(82)
Żewġ dgħajjes bla qlugħ ▲	/ l / → l	z ɛ ʊ ʧ d ɐ ɪ j ɛ s blɐ: ʔl	(71)
Żewġ dgħajjes bla qlugħ ▲	/ u / → ʊ	z ɛ ʊ ʧ d ɐ ɪ j ɛ s blɐ: ʔlʊ	(39)
Żewġ dgħajjes bla qlugħ ▲	/ għ / _ → h	z ɛ ʊ ʧ d ɐ ɪ j ɛ s blɐ: ʔlʊh	(55)

Figure 1: Example run of the G2P rules on the phrase "Żewġ dgħajjes bla qlugħ"

Processing of the phrase proceeds in left-to-right order. Each row in this figure shows the current position of the reading head, indicated by an arrow (▲), the G2P rule that is activated at this position, and the phonemic output obtained so far. The G2P rules are formatted as: *left-context*/*grapheme(s)*/*right-context* → *phoneme(s)*. The left and right contexts may be empty in the absence of a context which applies for the given rule. The underscore (_) symbol denotes a word boundary, while a group of graphemes separated by commas indicate that any one of the graphemes in question can occur as context. Finally, the rule numbers shown in parentheses refer to the rules as defined in Table 7.

While a rule-based approach for Maltese G2P working on the orthographic level gives quite good results, it is not sufficient to cover all possible pronunciations. The Maltese language has a small number of heterophonic homographs (words with different spoken sounds but with the same written form); these can only be differentiated via semantic interpretation (Farrugia 2005). For example, *sur* can be pronounced as /sʊr/ (English 'Mr.') or as /suːr/ (English 'fortified wall'). It is envisioned that the final Maltese TTS system will have a lexicon containing a list of exception words with their phonemic transcription. The G2P module will make use of this lexicon and apply the G2P rules described here for unknown (out-of-vocabulary) words. Due to the nature of the Maltese orthography, and based on the results of the G2P module obtained so far, it is expected that the size of this lexicon will be quite small.[3]

Vowels	ɪ	18,292,597	Fricatives	s	6,347,603	Affricates	ts	1,700,828
	ɐ	15,448,552		f	3,034,957		ʤ	1,141,271
	ɛ	7,778,560		h	2,162,600		ʧ	951,329
	ʊ	7,618,576		ʃ	1,658,811		dz	8,632
	ɔ	5,096,767		z	1,048,078			
	ɐː	3,195,630		v	989,300	Nasals	n	9,752,059
	ɪː	2,226,554		ʒ	5,502		m	6,645,891
	ɛː	756,956						
	ɔː	173,623	Plosives	t	12,253,833	Liquids	l	12,560,881
	iː	171,468		k	4,470,418		r	7,656,107
	à	95,733		d	4,148,424			
	uː	74,202		p	3,242,782	Glides	j	4,629,206
	ò	7,051		b	2,512,670		w	1,076,580
	ù	3,897		ʔ	1,702,567			
	è	3,403		g	821,119	Silence	#	5,311,123
	ì	304						

Table 4: Phoneme frequency count

3. Statistical analysis of the corpus

Statistical analysis of the phonetic transcription of the text corpus is performed for two main reasons: (1) to obtain statistics, such as frequency counts, of the diphone units that will help in the design and fine-tuning of the Maltese text-to-speech system, and (2) to arrive at a free-text sample that is as representative as possible of the main corpus. The

[3] When work on the Maltese TTS system was finished, and after the original version of this document was written, we investigated the size of this lexicon with respect to phonemic transcription exceptions, i.e., words for which an incorrect phonemic transcription is generated by the G2P rules of this paper. Results from this investigation validated our expectation that the number of exceptions will be quite small – more in Appendix 3.

latter consists of selecting phonetically-rich text blocks, made up of sentences of regular structure and reasonable length, that should enable the speaker to read them easily and with the expected prosodic patterns, so that naturalness is preserved. This is in contrast to text that is constructed manually with the intent of covering a wide range of sounds. However such constructed text tends to be nonsensical, more difficult to read and often assumes a uniform diphone frequency distribution.

Table 4 above gives the phoneme frequency counts of the text corpus.

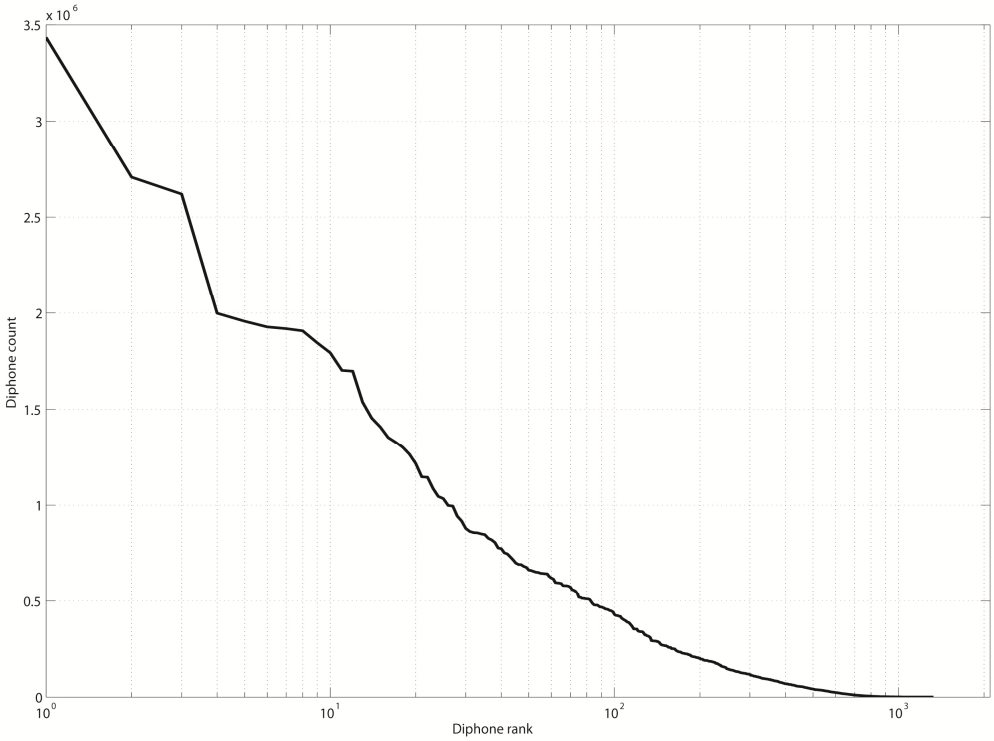

Figure 2: Diphone absolute frequency by rank

3.1. Diphone analysis

Phonetic transcription of the full text corpus yielded approximately 153.5 million diphones, which were then analyzed to find the number of distinct diphones and their frequency counts. Out of a total of 1681 (=41x41) possible phoneme combinations, 1450 distinct diphones were found in the text corpus. Figure 2 above shows these diphones ordered by their frequency count in a descending order and plotted on a semi-log scale.

Diphone	Count	%	Cumulative %
l+ɪ	3,435,162	2.24	2.24
t+ɐ	2,791,679	1.82	4.06
ɪ+l	2,707,466	1.76	5.82
ɪ+n	2,619,469	1.71	7.53
ɐ+l	2,538,302	1.65	9.18
n+ɪ	1,998,740	1.30	10.49
ɪ+s	1,918,389	1.25	11.74
t+ɪ	1,907,160	1.24	12.98
ɪ+t	1,793,403	1.17	14.15
ɐ+r	1,762,632	1.15	15.30
l+l	1,698,126	1.11	16.40
ɐ+t	1,614,533	1.05	17.46
m+ɪ	1,537,180	1.00	18.46
s+t	1,454,783	0.95	19.41
t+t	1,353,185	0.88	20.29
n+t	1,328,681	0.87	21.15
r+ɐ	1,319,982	0.86	22.01
ɪ+j	1,301,130	0.85	22.86
r+ɪ	1,264,935	0.82	23.69
n+ɐ	1,242,290	0.81	24.50
ɔ+n	1,218,616	0.79	25.29
ɐ+n	1,206,091	0.79	26.08
ɛ+n	1,149,137	0.75	26.82
ɛ+r	1,146,904	0.75	27.57
d+ɐ	1,106,694	0.72	28.29
t+ɐː	1,092,714	0.71	29.01
m+ɐ	1,059,509	0.69	29.70
j+ɐ	1,055,206	0.69	30.38
l+ɐ	1,052,547	0.69	31.07
j+ɪ	1,034,211	0.67	31.74
ɛ+t	1,009,907	0.66	32.40
ʊ+n	996,849	0.65	33.05
s+s	995,425	0.65	33.70
ɪ+m	918,871	0.60	34.30
m+ɛ	865,573	0.56	34.86
k+ɔ	864,716	0.56	35.42

Table 5: Most frequent diphones in Maltese

It can be seen that the curve of this figure exhibits a gradual decrease to 0. The last few hundred diphones were validated manually to check whether they occur naturally in the Maltese language or not. It was found that 101 of these diphones are caused by transcription errors or foreign words, leaving a final total of 1349 distinct Maltese diphones.

The data suggests a Zipfian distribution, exhibiting rapid drops in frequency at the top ranks, which is a common occurrence in natural language processing (Manning 1999).

Table 5 lists the 35 most frequent diphones, which together account for approximately one third of all the diphones in the corpus. The statistics obtained also show that the first 71 diphones from the 1349 distinct diphones account for 50% of all diphones in the corpus, and that the first 322 diphones account for 90% of all diphones. Figure 3 below shows the diphone frequency counts as a transition matrix, the lighter the color the higher the count.

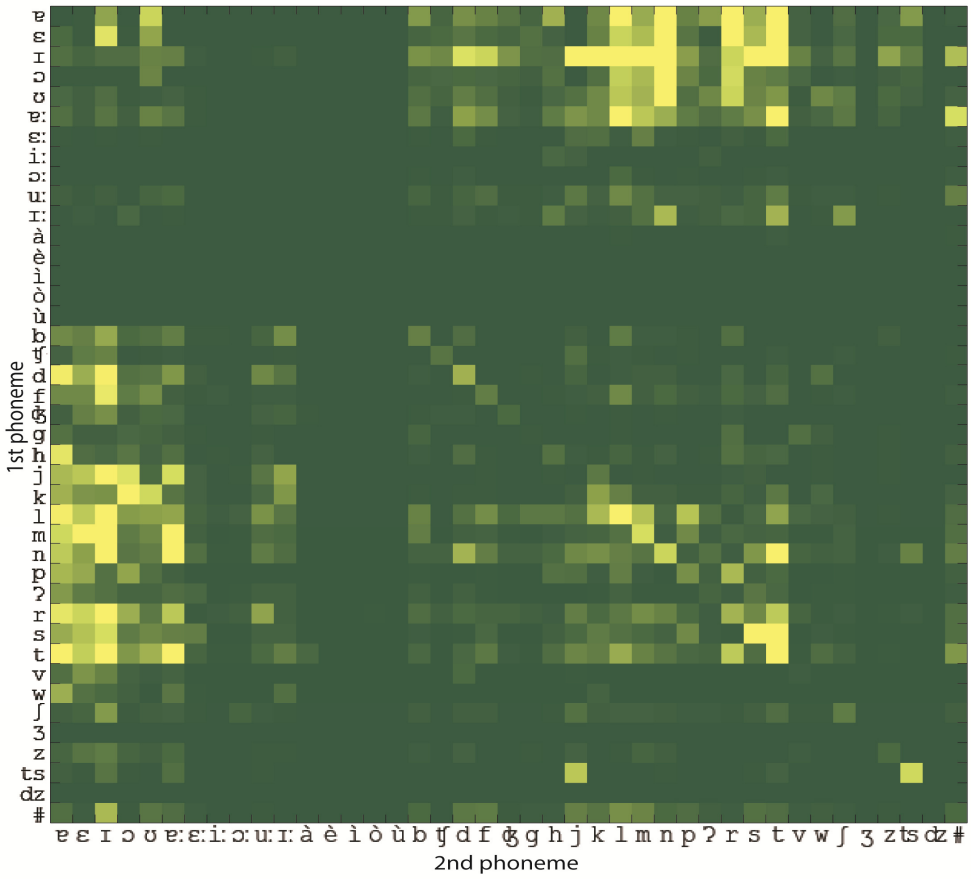

Figure 3: Diphone transition matrix

Analysis of the various source texts in the corpus (refer to Table 6), reveals that the Newspaper texts are the most phonetically rich, i.e., 97% of the 1349 diphones occurred at least once in the Newspaper texts (even though these texts make up just 37% of the total corpus). This is followed by the Maltese Wikipedia web pages (91.8%), followed by the Parliament Debates (89.7%), Il-Bibbja (81.6%), and the Maltese books (75.6%).

```
INPUTS:   Original_corpus, Optimal_sample_size, Initial_text_block_size
OUTPUT:   Optimal_sample

SET Optimal_sample TO empty set
SET Text_block_size TO Initial_text_block_size

WHILE size of Optimal_sample < Optimal_sample_size

        Partition Original_corpus into text blocks of Text_block_size words,
                rounded to nearest sentence

        FOR EACH Text_block not in Optimal_sample
                Generate New_sample by concatenating Optimal_sample with Text_block
                Compute feature vectors of scores for New_sample
        END FOR

        Rank all New_sample instances by scores

        SET Optimal_sample TO highest ranking New_sample
        SET Text_block_size TO Text_block_size / 2

END WHILE
RETURN Optimal_sample
```

Figure 4: Free text selection algorithm

4. Free text selection

The aim of this free text selection method is to distil an optimal sample from the normalized corpus in terms of its phonemic and prosodic features. The selection of free text is carried out incrementally. After the main corpus has been analyzed, the statistics gathered are used to compile a synthesized descriptor, a space containing the identifying features of this global text. We describe this space via a number of vectors of the form:

$$\langle diphone, position\ score, frequency\ score \rangle$$

The selection process, which is iterative in nature, divides the corpus into text blocks of equal word count, rounded to the nearest sentence, which get shorter during subsequent iterations. These blocks are analyzed and their feature vectors compiled and ranked. The top entry is composited into a selection which contains all the top entries from

previous iterations. The ranking mechanism generates a score for the current selection taking into account each individual text block, with highest scores being proportional to the similarity of features between the global text and the selection. The number of unique diphones occurring in a ranked text block determines the number of feature vectors associated with it. We base the ranking score on two important diphone features, position ϕ and frequency ψ, computed using a general 4-D weighted distance function:

$$\Delta(\phi_s, \phi_w, \psi_s, \psi_w) = \sqrt{(\phi_s \cdot \phi_w)^2 + (\psi_s \cdot \psi_w)^2}$$

The weights for the frequency and position components, ϕ_w and ψ_w are fixed throughout the process. In the diphone position score ϕ_s we attempt to capture prosodic variations on each diphone, by trying to match the diphone position distribution in phrases and words: in phrases by unit position, in words by syllable number. By capturing phrase positions of diphones, we try to approximate variations due to intonation, while by capturing syllable positions we try to approximate stress in words. The diphone frequency component modulates the position score, factoring the diphone occurrences into the final score. The final score represents the diphone coverage of the given text block with respect to the global text block. While the diphone frequency score is computed as the ratio of diphone occurrences between the text block being ranked and the global text, the diphone position score is given by the weighted sum of each of the respective individual diphone position scores for all diphones present in the text chunk, and is defined as:

$$\phi_s = \frac{1}{g} \sum_{d \in D_l} \phi_s(d)$$

where g is the diphone count in the global text, D_l is the set of diphones occurring in the ranked text block and $\phi_s(d)$ is the diphone position score for each individual diphone d, computed as follows:

$$\phi_s(d) = \Delta\big(\lambda_s(d),\, \lambda_w(d),\, \mu_s(d),\, \mu_w(d)\big)$$

where the tuple (λ_s, λ_w) represents the diphone phrase position score and weight, and (μ_s, μ_w) represents the diphone syllable score and weight. Both the phrase position score λ_s and the syllable position score μ_s are similarity scores computed using a scale invariant method on the position histograms of phrases and syllables respectively. Let D_g be the set of all distinct diphones occurring in the global text. Let G_d^{λ} and G_d^{μ} be the histograms for the global syllable and phrase positions for diphone d where $d \in D_g$. Let L_d^{λ} and L_d^{μ} be the histograms for the local syllable and phrase positions for diphone d where $d \in D_g$. We define the similarity function $S_k(d)$ for diphone d as follows:

$$S_k(d) = \frac{\sum_{i=1}^{|G_d^k|} \min\left(L_d^k(i), G_d^k(i)\right)}{\max\left(\sum_{i=1}^{|L_d^k|} L_d^k(i), \sum_{i=1}^{|G_d^k|} G_d^k(i)\right)}$$

where $k\epsilon\{\lambda,\mu\}$. The selection process compiles a free text of approximately 10000 words from the main corpus using the techniques and metrics specified. The free text selection algorithm is summarized in Figure 4.

5. Results

In this section, we discuss the performance of our free-text selection method and compare it to other approaches for collating a speech corpus.

Figure 5 shows how the diphone coverage score $\Delta(\phi_s, \phi_w, \psi_s, \psi_w)$ and diphone frequency score ψ of the chosen free text changed with each of the 50 iterations required to achieve a 10000 word free text. The initial text block size is of 500 words. At around the 6500-word mark, the varying text block size (rounded to the nearest sentence) goes down to just 1 sentence in size, and the diphone frequency score curve exhibits a marked increase, which is also reflected in the diphone coverage score. The final value of ψ is 1.0, meaning that all the 1349 diphones occur at least once in the free text.

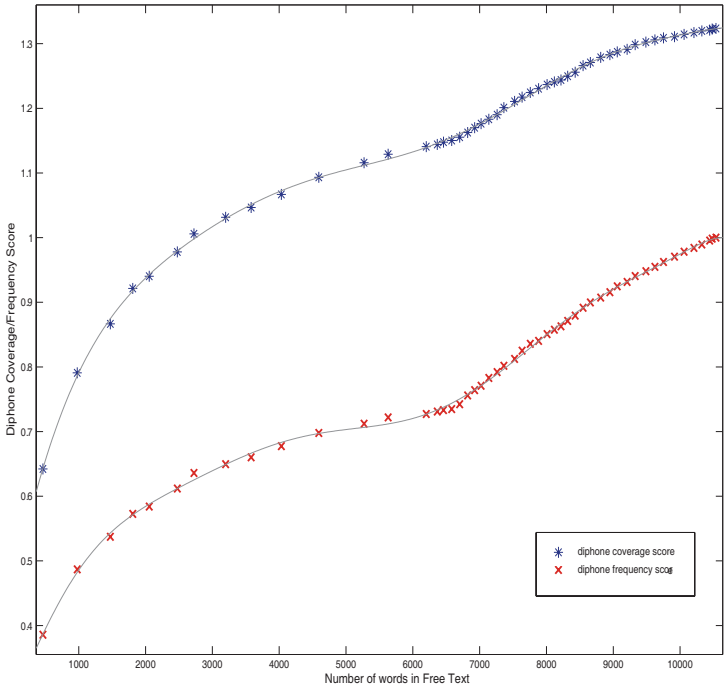

Figure 5: Diphone coverage and frequency scores of the selected Free Text at each iteration

In order to evaluate the effectiveness of the free text selection method described in this paper, a comparison was made with other selection methods, mainly: (1) a random free

text selection method, (2) a weighted random selection method, and (3) against a manually-generated text. The weighted random selection method performs importance sampling of the text sources which have been previously weighted. Therefore, a text source with a higher weight is a more probable candidate for selection than one with lower weight. Once a text source has been chosen, a text block is randomly selected. For each different tuple of weights, a run of 100 free text candidates were generated and the best ranked candidate was selected for comparison. The manually generated text was prepared by a linguistic expert and consists of diphones embedded in carefully constructed sentences (somewhat similar to the rainbow passage text for English); unlike free text, the sentences of the manual text may be nonsensical.

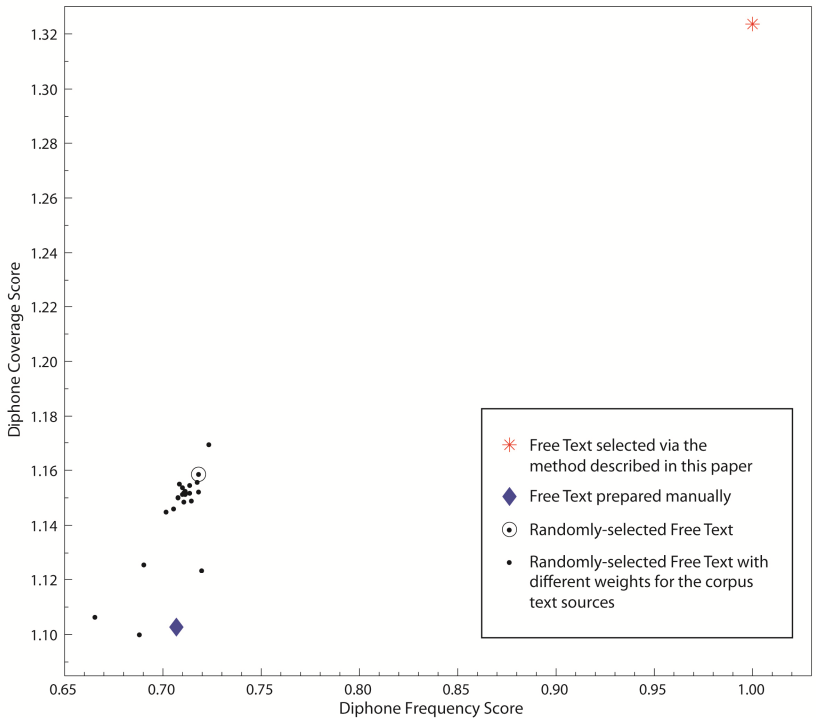

Figure 6: Diphone coverage scores of free text obtained through various selection methods

Figure 6 compares the scores obtained with the different free text selection methods. It can be seen that the selection method described in this paper outperforms all the others, both in terms of diphone frequency score $\psi = 1.0$ and the diphone coverage score $\Delta(.) = 1.324$. The best-scoring random-based selection achieves scores of $\psi = 0.723$ and $\Delta(.) = 1.169$, and uses the text source weights $\omega = \langle 1, 2, 3, 1, 2 \rangle$. Surprisingly enough, the manual text does not fare well ($\Delta(.) = 1.103$, $\psi = 0.707$); but in defense it must be said that the diphone statistics mentioned earlier were not available to the linguistic expert, hence the low diphone frequency score.

Figure 7 (top part) gives the difference between the diphone frequency counts of the free text selected by our method and the global frequency counts of the main corpus. Compared to the other free text candidates (bottom part of figure), this difference is small.

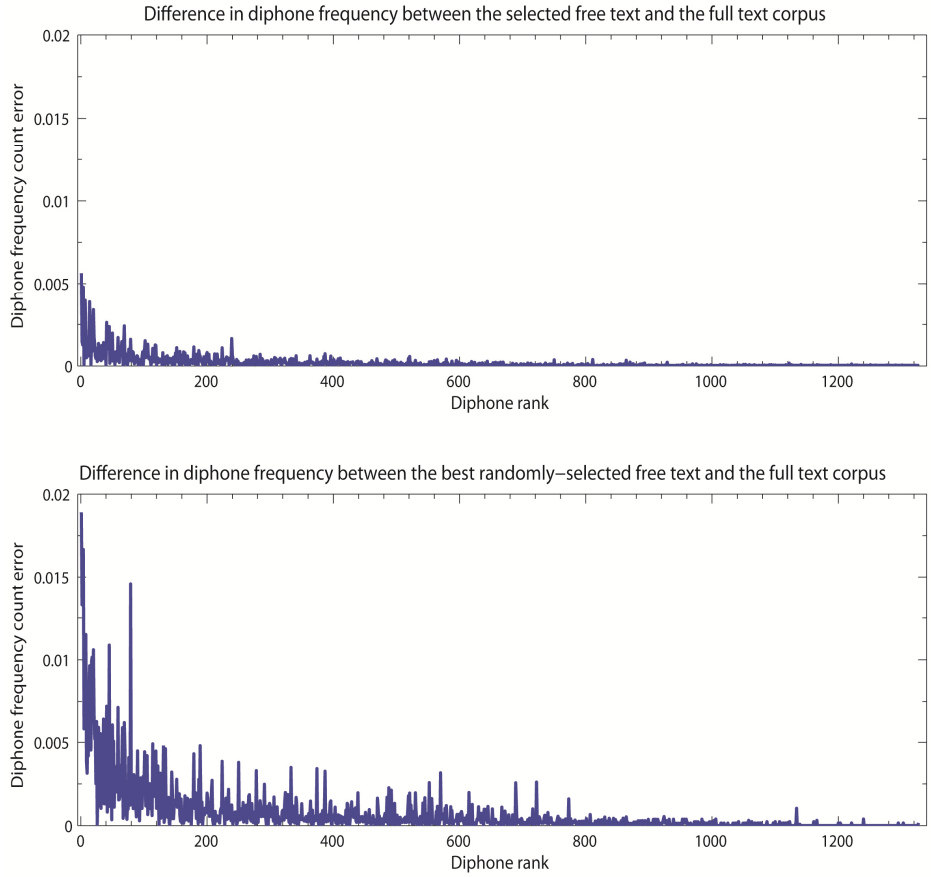

Figure 7: Error difference in diphone frequency between that of the selected free text and the full corpus

Finally, Figure 8 shows the phrase position and syllable position histograms of 3 frequent Maltese diphones (l + , + l, and + n); the histograms of the chosen free-text and the main corpus for these diphones are quite similar (e.g. $\lambda_s(\mathrm{l}+) = 0.871, \mu_s(\mathrm{l}+) = 0.943$).

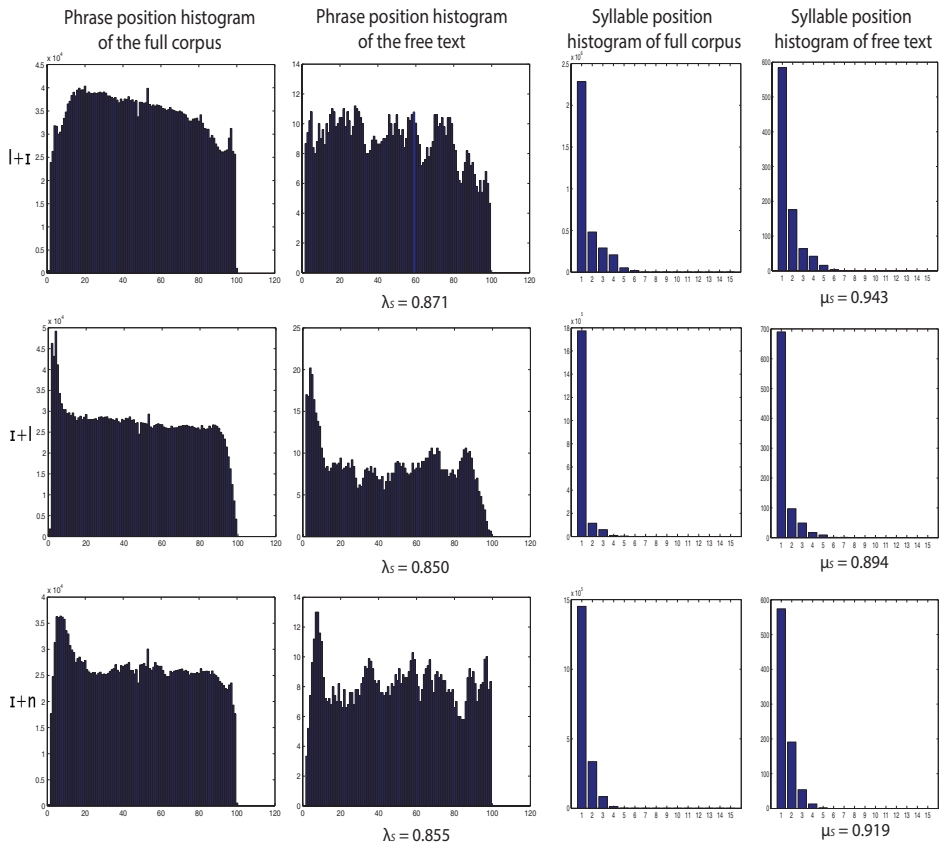

Figure 8: Phrase and syllable position histograms of the 3 diphones l+, +l, and +n, for the full text corpus and the free text

6. Conclusion

The work presented in this study details the generation of a free-running text corpus for Maltese concatenative speech synthesis. One of the major contributions of this work is the use of a novel free-text selection algorithm in the compilation of this corpus. This algorithm defines the diphone coverage measure as a weighted combination of diphone frequencies and their respective syllable and phrasal positions. As a result, we have achieved greater diphone coverage than other standard methods like weighted or manual selection. Both the free text corpus and the statistics collected during this study will be directly applied to the development of the Maltese TTS. It is worth noting that an additional advantage of our method is that it can also be applied to other languages given the availability of adequate G2P modules for the respective languages.

7. Future work

Work is in progress towards the improvement of corpus normalization by the addition of semiotic classification and the respective verbalization of these classes. The normalized text corpus can also lead to studies on word statistics which may prove to be key factors in future phases of development of Maltese TTS. In our study we have focused on the generation of a corpus using a specific distance metric that was based on phonemic and prosodic features. However, we believe that through the application of other distance metrics, it is possible to extract optimal corpora with a focus on other feature sets.

Appendix 1

Table 6 gives more detailed information about the text sources making up the corpus used as basis for the work of this paper.

Text source	No. of words	No. of normalized words	% of corpus	Original file format & encoding	Article range/download date, URL
Il-Bibbja (The Bible)	633,373	633,305	1.9	Microsoft Word documents, Unicode encoded	
Maltese Wikipedia	1,051,510	955,275	2.9	HTML text, Unicode encoded	Downloaded 8th – 15th April 2010 http://mt.wikipedia.org/wiki
"Il-Ġens" newspaper	1,293,505	1,238,752	3.7	HTML text, ASCII encoded	News articles dated 17th Oct 2009 to 19th Aug 2010 http://www.il-gensillum.com
"In-Nazzjon" newspaper	1,228,972	1,191,008	3.6	HTML text, ASCII encoded	News articles dated 15th June 2008 to 19th Aug 2010 http://www.maltarightnow.com
"L-Orizzont" newspaper	10,081,676	9,783,125	29.5	HTML text, ASCII encoded	News articles dated 29th Dec 2007 to 19th Aug 2010 http://www.l-orizzont.com
Parliament Debates	20,094,864	19,166,440	57.9	Microsoft Word documents, Unicode + legacy encodings	April 1992 – June 2010 (Debates of the 7th to the 11th legislatures) http://www.parlament.org.mt
Maltese Books	144,549	140,968	0.4	RTF documents, legacy encoding	Some of the books courtesy of Merlin Library Bookshop and Publishers Ltd.
Totals:	34,528,449	33,108,873	100%		

Table 6: Information on the text sources forming the corpus

Appendix 2: Maltese G2P Rules

Table 7 below lists the Maltese grapheme-to-phoneme rules, in order of priority, with specific rules given first, followed by generic ones. Each G2P rule is defined by a left and right context (where applicable), the grapheme character(s) and phoneme replacement(s), together with any condition that might apply to this rule. A word boundary is indicated by the _ character; and the meta-characters C and V stand for the consonants (*b*, *ċ*, *d*, *f*, *ġ*, *g*, *għ*, *h*, *ħ*, *j*, *k*, *l*, *m*, *n*, *p*, *q*, *r*, *s*, *t*, *v*, *x*, *ż*, *z*) and vowels (*a*, *e*, *i*, *ie*, *o*, *u*) respectively.

Since the consonant *għ* is a digraph (written down using a pair of characters), the G2P processing module pays special attention not to confuse the consonant *g* with the first character of the digraph *għ*, when checking a rule's left and right contexts. For example, rule 48 is not applicable to the grapheme *f* in the word *lifgħa*, but is applied to *f* in *tifga* (*lifgħa* → /lɪfɐː/, English 'leopard snake', and tifga → /tɪvgɐ/, English 'to choke').

As a practical measure, and in order to reduce the size of the diphone inventory in the final TTS system, length marks are not used for geminate consonants in our set of G2P rules. For example, the phonemic transcription of the word *giddieb* (English 'liar') generated by our G2P module is /gɪddɪːp/, and not /gɪdːɪːp/.

#	Left context	Grapheme(s)	Right context	Phoneme(s)	Rule condition	Example
1		għu		ɔ ʊ		tiegħu
2		għi		ɛ ɪ		tiegħi
3		aj		ɐ ɪ		mingħajr
4		aw		ɐ ʊ		jemigraw
5		ej		ɛ ɪ		fejn
6		ew		ɛ ʊ		żewġ
7		iw		ɪ ʊ		liwja
8		oj		ɔ ɪ		bojod
9		ow		ɔ ʊ		mowbajl
10	ie	għ	e	j		qiegħed
11		agħa		ɐː		mbagħad
12		egħe		ɛː		inxtegħel
13		ogħo		ɔː		bogħod
14	C	e	hiC,għiC	ɛː		fehimni
15	C	a	_	ɐː	single-syllable words	ra
16	C	e	_	ɛː	single-syllable words	re
17	C	o	_	ɔː	single-syllable words	ġo
18	C	u	_	uː	single-syllable words	kju
19		aha		ɐː		naraha
20		aho		ɔː		tahom
21		ehe		ɛː		deher
22	għ	a		ɐː		għar
23		a	għ	ɐː		fieragħ
24	għ	e		ɛː		għemil
25		e	għ	ɛː		qegħda

#	Left context	Grapheme(s)	Right context	Phoneme(s)	Rule condition	Example
26	għ	o		ɔː		għomja
27		o	għ	ɔː		logħba
28	h	a		ɐː		kollha
29		a	h	ɐː		tah
30	h	e		ɛː		hena
31		e	h	ɛː		xehda
32		ie	għC	ɛː		ibiegħdu
33		ie		ɪː		bieb
34		i	ħ,għ,h,q	iː		smigħ
35		a		ɐ		dar
36		e		ɛ		kelb
37		i		ɪ		bir
38		o		ɔ		qorq
39		u		ʊ		tuffieħ
40		b	ċ,f,ħ,k,p,q,s,t,x,z,_	p		libsa
41		b		b		borma
42		ċ	b,d,ġ,g,v,ż	ʤ		arċduka
43		ċ		ʧ		kċina
44	V	d	x,dx,tx	ʧ		riedx, roddx, ridtx
45	V	d	s,ds	ts		għadsa, imqaddsa
46		d	ċ,f,ħ,k,p,q,s,t,x,_	t		mard
47		d		d		dort
48		f	b,d,ġ,g,v,ż	v		fdal
49		f		f		fidda
50	V	għ	hV	h		tagħhom
51	V	għ	V		silent	lagħab
52	C	għ	V,j		silent	dgħajfa, dgħjufija
53	V,j	għ	C		silent	nilagħbu
54	_	għ			silent	għar
55		għ	_	h		fieragħ
56		ġ	ċ,f,ħ,k,p,q,s,t,x,_	ʧ		ħriġt
57		ġ		ʤ		ġebla
58		g	ċ,f,ħ,k,p,q,s,t,x,_	k		spag
59		g		g		gremxul
60	_	h			silent	hena
61		h	_	h		fih
62	C	h	V		silent	seraqhom
63	i,ie	h	V	j	different vowels as right & left context	fihom
64	u	h	a,i,ie,o,u	w		nafuha
65	V	h	V		silent	sehem
66		h			silent	
67		ħ		h		ħanut
68		j		j		jasal
69		k	b,d,ġ,g,v,ż	g		kbirna
70		k		k		kelb
71		l		l		lima
72	_i	m	d	m		imdejjaq
73	i	m	d	n		mimdud
74		m		m		mejda
75		n	b,p	m		denb, qanpiena

#	Left context	Grapheme(s)	Right context	Phoneme(s)	Rule condition	Example
76	i	n	l	l		inlumu
77	i	n	m	m		inmekkek
78	i	n	r	r		inrabbi
79		n		n		naqas
80		p	b,d,ġ,g,v,ż	b		
81		p		p		pipa
82		q		ʔ		qattus
83		r		r		ras
84	V	s	s_x	ʃ		miss xejn
85	V,s	s	x_	s		kinisx, rassx
86		s	b,d,ġ,g,v,ż	z		masġar
87		s		s		sena
88		t	b,d,ġ,g,v,ż	d		tbajja
89	V	t	x	ʧ		ratx
90	V	t	s	ts		għatsa
91		t		t		torta
92		v	ċ,f,ħ,k,p,q,s,t,x,z,_	f		kattiv
93		v		v		vapur
94		w		w		werqa
95		x	b,d,ġ,g	ʒ		xbajt
96	V	x	V	ʒ	applicability of rule determined by a word list	televixin
97		x		ʃ		kaxxa
98	V	zz	V	dz	applicability of rule determined by a word list	gazzetta
99		z		ts		zalza
100	V	ż	ż_x	ʃ		għożż xejn
101	V	ż	x	s		nehmiżx
102	ż	ż	x_	z		għożżx
103		ż	ċ,f,ħ,k,p,q,s,t,x,_	s		żfin
104		ż		z		żrar
105		à		à		università
106		è		è		kafè
107		ì		ì		Indrì
108		ò		ò		però
109		ù		ù		tabù

Table 7: Maltese grapheme-to-phoneme rules

Appendix 3: G2P Exception Lexicon

When work on the Maltese TTS system was complete, we performed an investigation on the Maltese words for which the set of G2P rules described in this paper generate an incorrect phonemic transcription, i.e. the G2P exception lexicon.

The Maltese TTS system makes use of a lexicon, which apart from abbreviations, acronyms, and foreign words, contains a total of 32,365 Maltese words. Of these 32,365 words, only 1,304 (4.0%) have a phonemic transcription different from those generated by the G2P rules. Surnames (e.g. *Asciak)*, names and toponyms (e.g. *Cospicua* (/kɔspiːkwɐ/) constitute approximately half of these G2P exceptions – 710 (2.2%) words in total.

Amongst the remaining 594 words (1.8%), one finds 54 heterophonic homographs (words with the same written form, but which have different spoken sounds and different semantic meanings). Some examples include: *ħakem* (/hɐːkɛm/ and /hɐkɛm/)*, baħħar*, *bajjad*, *kaħħal*, *saħħar*, *xandar*, *tarmak*, *sur*, *tajjar, qarsa*, *qala*, *hemm*, and *ġara*.

Other G2P exceptions include words like: *ċagħka* (/ʧɐːʔɐ/ instead of /ʧɐːkɐ/), *kewkba* (/kɛʊbɐ/ instead of /kɛʊgbɐ/), and *ġkieket* (/ʤgɪːgɛt/ instead of /ʧkɪːkɛt/). Some exceptions reflect the way certain words are commonly pronounced, like: *granmastru* (/grɐmmɐstrʊ/ instead of /grɐnmɐstrʊ/), *daqsxejn* (/dɐʔʃɛɪn/ instead of /dɐʔsʃɛɪn/), *għandna* (/ɐːnnɐ/ instead of /ɐːndnɐ/), *kooperattiva* (/kɔːpɛrɐttiːvɐ/ instead of /kɔɔpɛrɐttiːvɐ/), and *ġelledija* (/ʤɛllɛdɪjɐ/ instead of /ʤɛllɛdɪjɐ/).

The words *hieni* and *hienja* are exceptions to G2P rule 60 (/hɪːnɪ/ and /hɪːnjɐ/ respectively, and unlike *hena* which is transcribed correctly by rule 60 as /ɛːnɐ/). Similarly, the word *raheb* (/rɐːhɛp/) is an exception to G2P rule 65 (unlike *rahan*, transcribed correctly as /rɐːn/).

A number of G2P exceptions occur when words contain a consonant cluster of 3 or more consonants and conflicting voicing and devoicing G2P rules are activated. Examples of such exceptions include: *nobżqu* (/nɔpsʔʊ/ instead of /nɔbsʔʊ/), *fosdqa* (/fɔstʔɐ/ instead of /fɔztʔɐ/), *mrattba* (/mrɐdbɐ/ instead of /mrɐtdbɐ/), and *nixbħek* (/niʃphɛk/ instead of /nɪʒphɛk/). The incorrect phonemic transcription for the latter word *nixbħek* occurs when *x* is voiced to /ʒ/ because of the following *b* (via G2P rule 95), but at the same time, *b* is devoiced to /p/ because of the following *ħ* (via G2P rule 40). The G2P rules could be modified to cater for these cases – for example, G2P rule 95 could be modified so that if *b* occurs as the right context of *x*, an extra condition is added that specifies that *b* must not be followed by any of the following consonants: *p*, *t*, *k*, *f*, *ċ*, *s*, *x*, *q*, or *ħ*. Care must be taken to avoid the problem of combinatorial explosion with such modifications.

The small number of words (1.8%) in the G2P exception lexicon (which can be easily brought down to under the 1% mark if the G2P module is modified to handle correctly consonant clusters as described above), proves that a rule-based approach for the phonemic transcription of Maltese is a valid approach.

References

Borg, Albert & Azzopardi-Alexander, Marie (1997): *Maltese*. London/New York: Routledge.

Bozkurt, Baris; Ozturk, Ozlem & Dutoit, Thierry (2003): Text design for TTS speech corpus building using a modified greedy selection, in: *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003 – INTERSPEECH 2003)*. Geneva, Switzerland: ISCA, 277–280.

Dalli, Angelo (2000): Data representation formats for Maltese. Technical Report, University of Malta.

Divay, Michel & Vitale, Anthony J. (1997): Algorithms for grapheme-phoneme translation for English and French: applications for database searches and speech synthesis, in: *Computational Linguistics* 23, 495–523.

Farrugia, Paulseph-John (2005): Text-to-speech technologies for mobile telephony services. Master Thesis, University of Malta.

Hume, Elizabeth; Venditti, Jennifer; Vella, Alexandra & Gett, Samantha (2009): Vowel duration and Maltese 'għ', in: Comrie, Bernard; Fabri, Ray; Hume, Elizabeth; Mifsud, Manwel; Stolz, Thomas & Vanhove, Martine (eds.), *Introducing Maltese linguistics*. Amsterdam, Philadelphia: John Benjamins, 15–46.

Kawai, Hisashi & Tsuzaki, Minoru (2002): A study on time-dependent voice quality variation in a large-scale single speaker speech corpus used for speech synthesis, in: *Proceedings of the 2002 IEEE Workshop on Speech Synthesis*. California: IEEE, 15–18.

Kominek, John & Black, Alan W. (2003): CMU ARCTIC databases for speech synthesis. Technical Report, Carnegie Mellon University.

Laws, Mark R. (2003): Speech data analysis for diphone construction of a Maori online text-to-speech synthesizer, in: Hamza, M. H. (eds.), *Proceedings of the IASTED International Conference on Signal and Image Processing (SIP 2003)*. Honolulu, US: ACTA Press, 103–108.

Manning, Christopher D. & Schütze, Hinrich (1999): *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Micallef, Paul (1998): A text to speech system for Maltese. Ph.D. Thesis, University of Surrey.

Santen, Jan P. H. van & Buchsbaum, Adam L. (1997): Methods for optimal text selection, in: Kokkinakis, G. & Fakotakis, N. & Dermatas, E. (eds.), *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH 1997)*. Rhodes, Greece: ISCA, 2–5.