

Preparation of a Free-Running Text Corpus for Maltese Concatenative Speech Synthesis

**Mark Borg, Keith Bugeja, Colin Vella,
Gordon Mangion, Carmel Gafà**

Research & Development Dept.,
Crimsonwing p.l.c.

9th April 2011

The Maltese Text to Speech Synthesiser

- Crimsonwing (Malta) p.l.c. awarded tender to develop the Maltese Text to Speech Synthesiser by the Foundation for Information Technology Accessibility (FITA)
- Project co-financed (85%) by the EU's ERDF (European Regional Development Fund), and national funds (15%)
- Operational Programme I – Cohesion Policy 2007-2013 *Investing in Competitiveness for a Better Quality of Life*



CRIMSONWING
Better people : better business



FITa



INVESTING IN YOUR FUTURE
EUROPEAN REGIONAL DEVELOPMENT FUND
COHESION FUND
MALTA 2007-2013

The Maltese Text to Speech Synthesiser

- Features:
 - 3 different voices: male, female, child
 - High quality: Studio recorded (44 KHz 16bit sound quality)
 - Neutral discourse
 - Windows SAPI compliant (Speech Application Programming Interface)
 - Inter-operability with any application that is SAPI compliant (e.g. Window-Eyes, etc.)
 - Freely available for download
 - Available in 2012

Text to Speech (TTS) Synthesis

- 1st generation (1960's to mid-1980's):
 - Formant synthesis
 - Articulatory synthesis (based on vocal tract models)
 - Robotic sounding
- 2nd generation (mid-1980's to mid-1990's):
 - Concatenative synthesis
 - Single instance of each recorded unit
 - Heavy DSP (digital signal processing)
 - Can suffer from audible glitches at concatenation points
 - 1st work in Maltese TTS falls here (P. Micallef, PhD 1998)
- 3rd generation (mid-1990's onwards):
 - Concatenative Synthesis with Unit Selection
 - Multiple instances of each recorded unit
 - Choosing the best 'chain' of candidate units
 - Less DSP

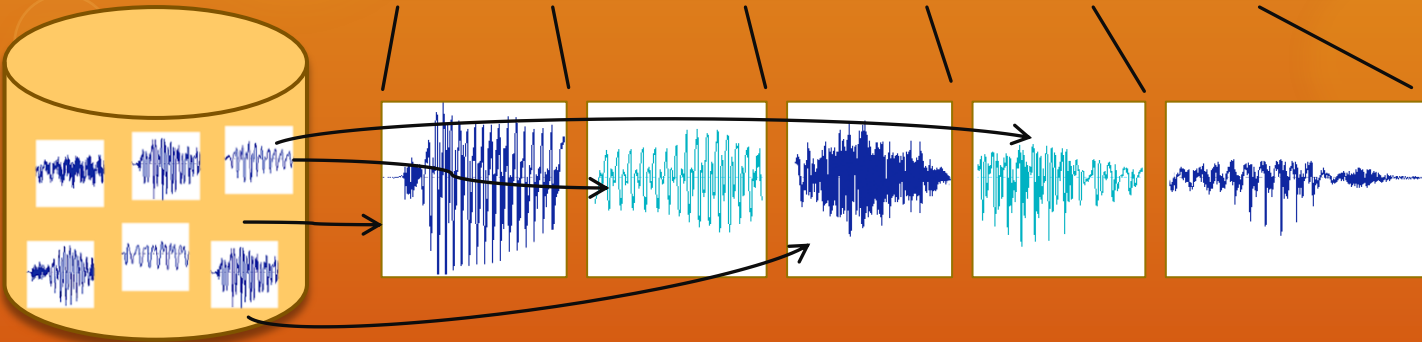
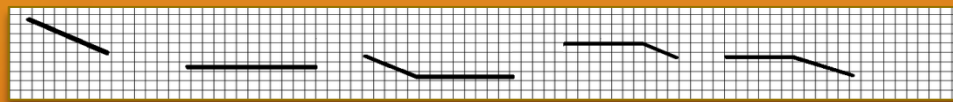
Concatenative Speech Synthesis

Dan mhux xogħol hafif, imma jrid isir.

- Given some utterance to be synthesised...

Concatenative Speech Synthesis

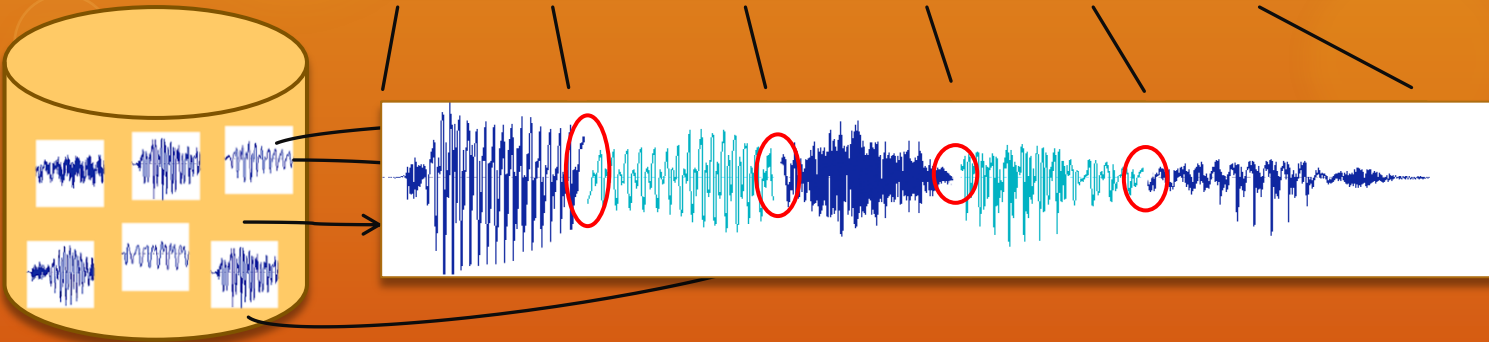
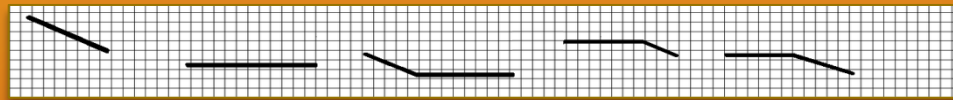
dən mʊf ʃɔ:l həfɪf, ɪmme jɹɪt ɪsɪr.



- Given some utterance to be synthesised...
- A phonemic transcription is generated
- The required prosodic model is generated
- Database with recorded speech, segmented into audio segments (units)
- The given utterance is divided into segments (units) and the best matching units from the database are selected
- The units are concatenated together

Concatenative Speech Synthesis

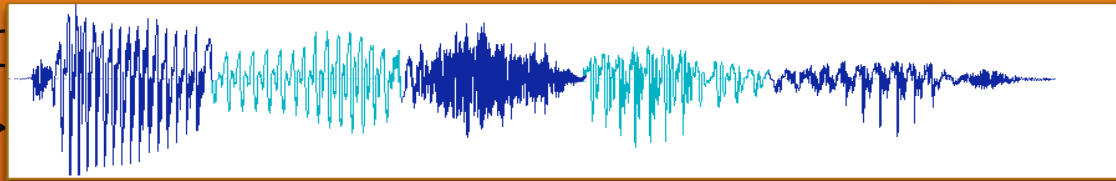
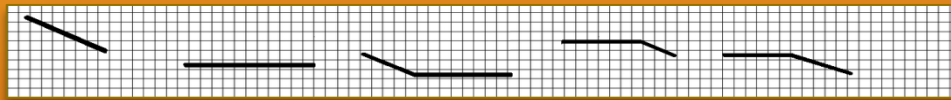
dən mʌf ʃɔ:l həfɪf, ɪmme jrit ɪsɪr.



- Given some utterance to be synthesised...
- A phonemic transcription is generated
- The required prosodic model is generated
- Database with recorded speech, segmented into audio segments (units)
- The given utterance is divided into segments (units) and the best matching units from the database are selected
- The units are concatenated together
- Some DSP is applied to smooth the joins between the units

Concatenative Speech Synthesis

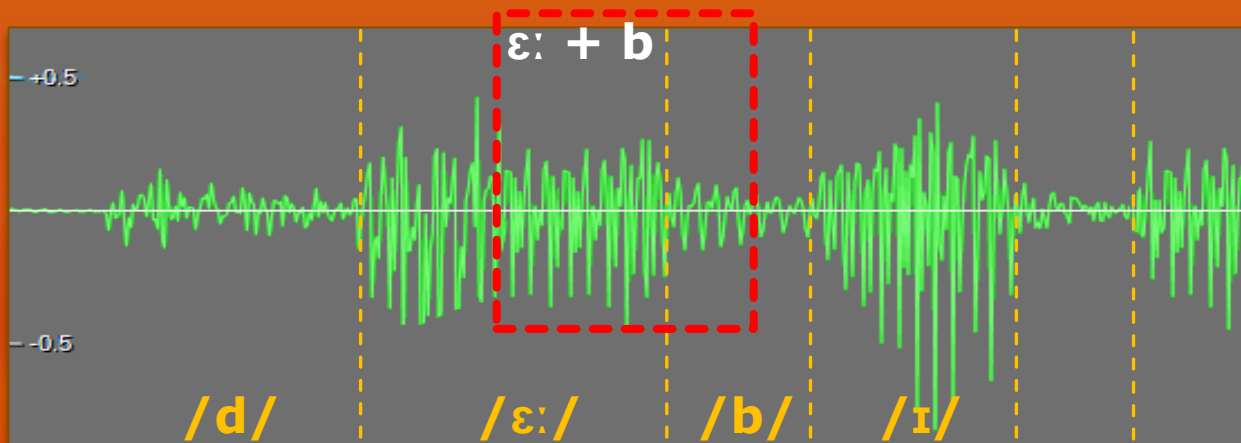
dən mʊf ʃɔ:l həfɪf, ɪmme ʃrɪt ɪsɪr.



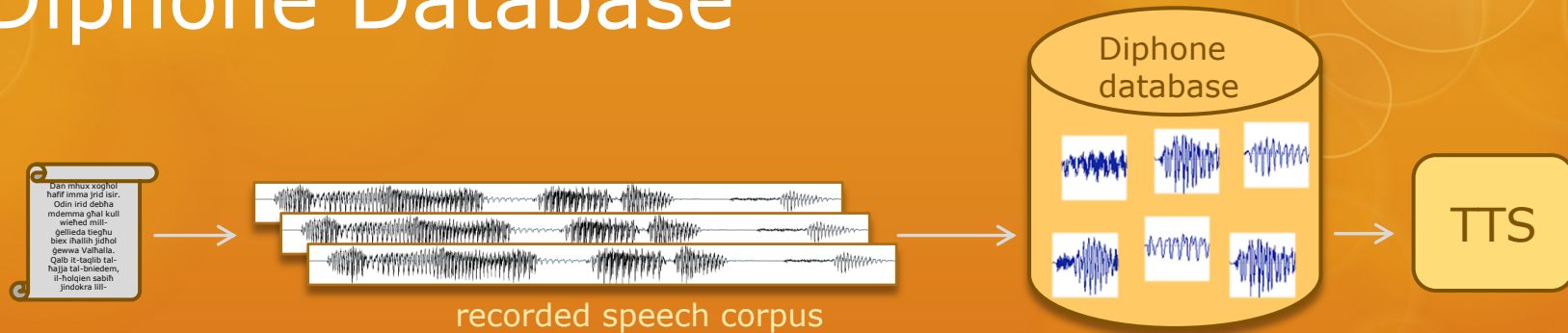
- Given some utterance to be synthesised...
- A phonemic transcription is generated
- The required prosodic model is generated
- Database with recorded speech, segmented into audio segments (units)
- The given utterance is divided into segments (units) and the best matching units from the database are selected
- The units are concatenated together
- Some DSP is applied to smooth the joins between the units

Concatenative Speech Synthesis

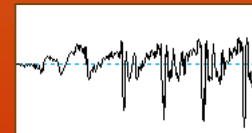
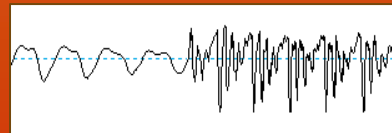
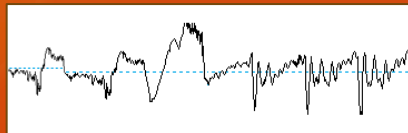
- What type of units to use for TTS?
- Half-phonemes, phonemes, diphones, triphones, syllables, etc.
- Closed vs. Open domains
- Co-articulation effects
- Diphones chosen for the Maltese TTS engine.
 - Compromise between number of units, co-articulation effects
 - Easier to do concatenation at the stationary parts of speech signals



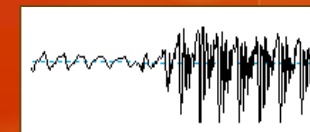
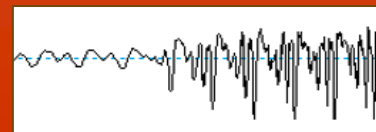
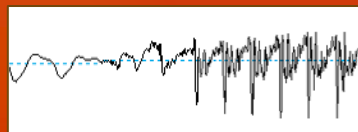
Diphone Database



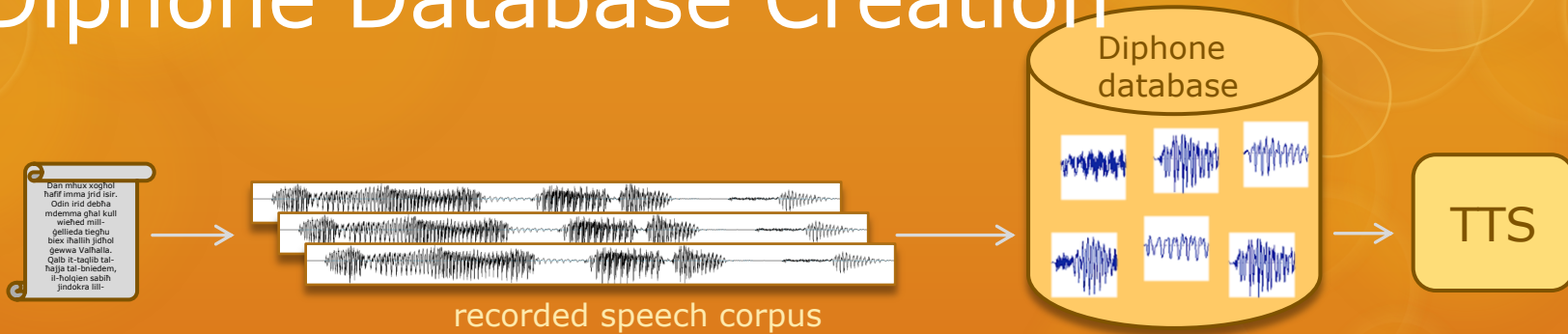
- Quality of synthesised speech is highly dependent on the corpus of recorded speech used to create the diphone database
- Large database required for sufficiently natural-sounding speech (spanning several to tens of hours)
- Large number of diphones needed for unit selection TTS



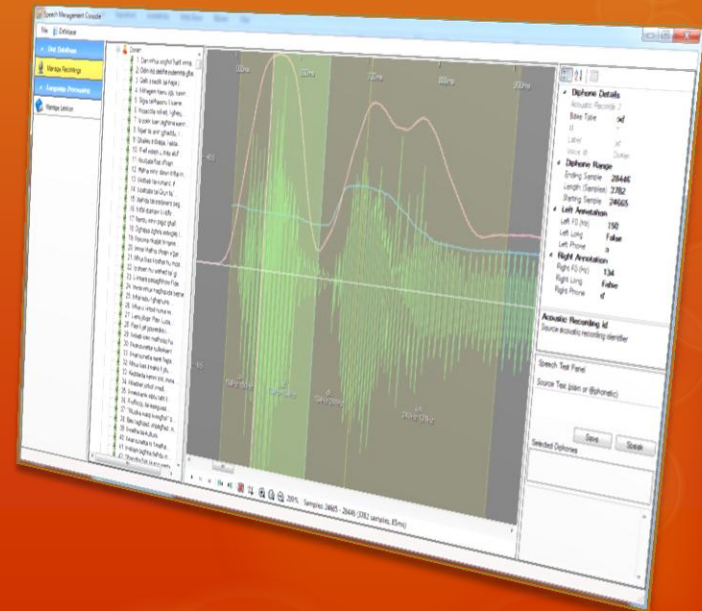
/b/ + /a/



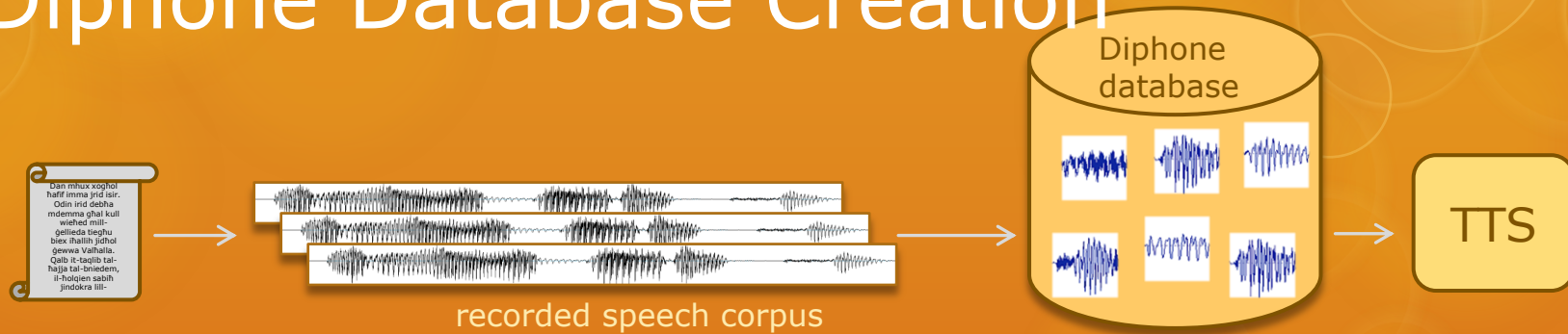
Diphone Database Creation



- Diphone cutting:
 - Manual process
 - Performance of automatic diphone segmentation methods is currently limited
 - Semi-automatic methods still require manual intervention
 - Labour and time intensive
 - Also recording constraints



Diphone Database Creation



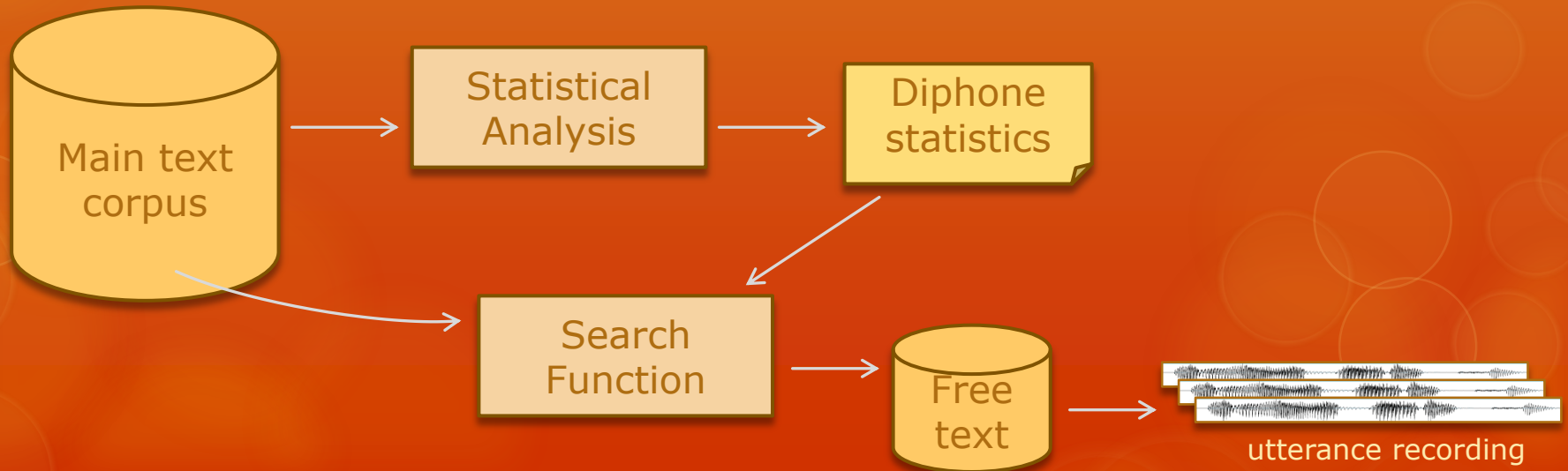
- Diphone Coverage
- How many of the potential diphones occur in Maltese?
- Which are the most frequent diphones?
 - Need statistics on diphone frequency and variation
- Therefore scope for this work

Free Text and Rainbow Text

- Often, the text used for the speech recordings is randomly sampled from a large corpus
 - Called “free text”
 - Consisting of a free-running text sample, made up of sentences of regular structure and reasonable length
 - Enables the speaker to read the text easily and with the expected prosodic patterns, so that naturalness is preserved
 - No optimisations applied towards the extraction of an optimal sample
- Or, a “rainbow text” is manually prepared by an expert
 - Consists of diphones embedded in carefully constructed sentences (many times, non-sensical)
 - E.g.: **/ɪ:/ + /w/** il-k***liewi*** fix-x***tiewi*** hu kliem ***siewi***
 - Covers at least one instance of each diphone
 - Unnatural

Automatic Generation of a Free-Running Text Corpus

- The aim of this work is:
 - To develop an automated search function (search process) that maximises diphone coverage when choosing the free text needed for utterance recording
 - The selected free text will be a small manageable portion of the full text corpus, and that is as representative of the main corpus as much as possible



Automatic Generation of a Free-Running Text Corpus

- *Diphone coverage* measure:
 - Not just getting one instance of all the diphones that can occur in Maltese
 - Get more instances of the most frequent diphones
 - Diphone position distribution
 - We attempt to capture prosodic variations on each diphone, by using the diphone position distributions in phrases and words
 - For the position in phrases, we use unit position
 - For the position in words, we use syllable number
 - By capturing phrase positions of diphones, we try to approximate variations due to intonation
 - By capturing syllable positions of diphones, we try to approximate stress in words

Main Text Corpus Preparation

- A number of text sources:
 - Online newspapers, websites, official documents, and Maltese books.
- Diverse nature of texts
 - Text cleaning and normalisation into a homogenous corpus

Text Source	Words	Normalised	Size as % of corpus
Maltese Books [†]	144,549	140,968	0.4
Il-Bibbja (The Bible)	633,373	633,305	1.9
Maltese Wikipedia	1,051,510	955,275	2.9
“Il-Ġens” newspaper	1,293,505	1,238,752	3.7
“In-Nazzjon” newspaper	1,228,972	1,191,008	3.6
“L-Orizzont” newspaper	10,081,676	9,783,125	29.5
Parliament Debates	20,094,864	19,166,440	57.9
Totals:	34,528,449	33,108,873	100

[†]Some of the books courtesy of Merlin Library Bookshop and Publishers Ltd.

Text Cleaning and Normalisation

- Character conversion to Unicode (UTF-8) standard
 - Legacy encodings, HTML codes for extended graphemes
- Filtering of semiotic elements like numbers, dates, emails, etc.
 - Semiotic class analyser and verbalisation not available at this stage

Examples:

38.79, MMXI, 7/4/2011, 11:45, 7.5m², 720x576, A320-200, H1N1,

Text Cleaning and Normalisation

- Character conversion to Unicode (UTF-8) standard
 - Legacy encodings, HTML codes for extended graphemes
- Filtering of semiotic elements like numbers, dates, emails, etc.
 - Semiotic class analyser and verbalisation not available at this stage
- Filtering of known abbreviations and acronyms
 - Via lookup into exception structure

Examples:

UHM, GWU, SCUBA, Dr., il-GDPs, ež, p.ež,

Text Cleaning and Normalisation

- Character conversion to Unicode (UTF-8) standard
 - Legacy encodings, HTML codes for extended graphemes
- Filtering of semiotic elements like numbers, dates, emails, etc.
 - Semiotic class analyser and verbalisation not available at this stage
- Filtering of known abbreviations and acronyms
 - Via lookup into exception structure
- Detection and filtering of unknown abbreviations and initials
 - Via regular expression matching

Examples:

L. N. Abela, U.N., Q.K., U.S., USA, i.e.,

Text Cleaning and Normalisation

- Character conversion to Unicode (UTF-8) standard
 - Legacy encodings, HTML codes for extended graphemes
- Filtering of semiotic elements like numbers, dates, emails, etc.
 - Semiotic class analyser and verbalisation not available at this stage
- Filtering of known abbreviations and acronyms
 - Via lookup into exception structure
- Detection and filtering of unknown abbreviations and initials
 - Via regular expression matching
- Detection and filtering of foreign text
- Filtering of elements like surnames which are written using Latin characters

Examples:

Chetcuti, Camenzuli, Muscat, B'Bugia,

Text Cleaning and Normalisation

- Character conversion to Unicode (UTF-8) standard
 - Legacy encodings, HTML codes for extended graphemes
- Filtering of semiotic elements like numbers, dates, emails, etc.
 - Semiotic class analyser and verbalisation not available at this stage
- Filtering of known abbreviations and acronyms
 - Via lookup into exception structure
- Detection and filtering of unknown abbreviations and initials
 - Via regular expression matching
- Detection and filtering of foreign text
- Filtering of elements like surnames which are written using Latin characters
- Segmentation of text into phrases and phrase classification (statement, question, exclamative)

After text cleaning and normalisation, the final text corpus size is of just over 33 million words.

Grapheme to Phoneme (G2P)

- Phonemic transcription of the given text
- Low degree of heterography in Maltese
 - Relationship between the orthography (graphemes) and the sounds (phonemes) is relatively straight forward
- A set of context-sensitive rewrite rules is generally sufficient for phonemic transcription of Maltese text

Phonemes

Graphemes	Phonemes (short)	Phonemes (long)
a	ɐ	ɛ:
e	ɛ	ɛ:
i	ɪ	i:
o	ɔ	ɔ:
u	ʊ	u:
ie		i:

Graphemes	Phonemes
à	à
è	è
ì	ì
ò	ò
ù	ù

Graphemes	Phonemes	Graphemes	Phonemes
b	b	p	p
ç	tʃ	q	ʔ
d	d	r	r
f	f	s	s
ğ	dʒ	t	t
g	g	v	v
h	h	w	w
j	j	x	ʃ
k	k	x	ʒ
l	l	z	ts
m	m	z	dz
n	n	ż	z

Graphemes	Phonemes
<i>silence</i>	#

G2P rules

- A set of context-sensitive rewrite rules: $F = \langle f_i \rangle$
 - $f_i: xGy \rightarrow xPy$ $G \in \{\text{graphemes}\}, P \in \{\text{phonemes}\}$
 - $f_i: xGy \rightarrow xPy \mid \text{condition}$
- Most rules define a straightforward mapping between graphemes and phonemes
- Other rules incorporate vowel lengthening, devoicing, and voicing assimilation
- Approximately 110 rewrite rules for Maltese G2P

G2P rules

$f_i: xGy \rightarrow xPy \mid \text{condition}$

#	Left context x	Grapheme(s)	Right context y	Phoneme(s)	Conditions?	Example
f_1		ghu		ɔ ʊ		tieghu
f_5		ej		ɛ ɪ		fejn
f_{84}		r		r		ras
f_{41}		b	ć,f,h,k,p,q,s,t,x,z,_	p		libsa
f_{42}		b		b		borma
f_{55}	_	gh				ghar
f_{56}		gh	_	h		fieragh
f_{11}		agħa		e:		mbagħhad
f_{15}	<i>consonant</i>	a	_	e:	<i>1-syllable</i>	ra
f_{46}	<i>vowel</i>	d	s,ds	ts		ghadsa
f_{99}	<i>vowel</i>	zz	<i>vowel</i>	dz	<i>word-list</i>	gazzetta
f_{100}		z		ts		zalza
f_{89}		t	b,d,ġ,g,v,ż	d		tbajja
f_{90}	<i>vowel</i>	t	x	tʃ		ratx
f_{91}	<i>vowel</i>	t	s	ts		ghatsa
f_{92}		t		t		torta

G2P Conversion

Dan mhux xogħol ħafif imma jrid isir. Odin irid debħa mdemma għal kull wieħed mill-ġellieda tiegħu biex iħallih jidħol ġewwa Valħalla. Qalb it-taqlib tal-ħajja tal-bniedem, il-ħolqien sabiħ jindokra lill-imgarrba. Il-bhejjem kienu jiġu hawn biex jixorbu minn nixxieġha, u ċ-ċrief li jgerrmu l-qoxra taz-zkuk, kien jitfgħu leħħa ħafifa fl-għabex. Sigra tal-fraxxnu...



den muʃ ʃɔ:l heʃif imme jrit isir # ɔdin irit deʃhe mdeɪmma e:l kull wi:heʃ
mi:lɔʒe:li:de ti:ɔʊ bi:ʃ i:he:li:h jithɔl ɔʒe:ʊwe velhe:le # ʔelp itteʔli:p telhe:je
telbni:deɪm # ilhɔʔi:n sebi:h jindokra lillimɔʒerrbe # ilbe:jeɪm ki:nʊ jidʒʊ
e:ʊn bi:ʃ jixɔrbʊ minn niʃʃi:e: # ʊ ʃʃri:f li jgerrmu lʔɔʃre teʃteʃku:k # ki:n jifɔʊ
leħhe heʃife fle:beʃ # siɔʒre telʃreʃʃnʊ li ki:net weʔe:t # de:ret phel ɔʒgent...

- Phonemic transcription of corpus yielded 153.5 million diphones
- G2P rules give very good results for Maltese
- A few exceptions:
 - E.g. Word "sur" can be pronounced as:
 - /s ʊ r/ ("Mr."), or as
 - /s u: r/ ("fortified wall")

Statistical Analysis of Corpus

- Statistical analysis of corpus to get diphone coverage statistics:
 - Diphone frequency counts
 - Distribution of positions of diphones in word syllables
 - Distribution of positions of diphones in phrases



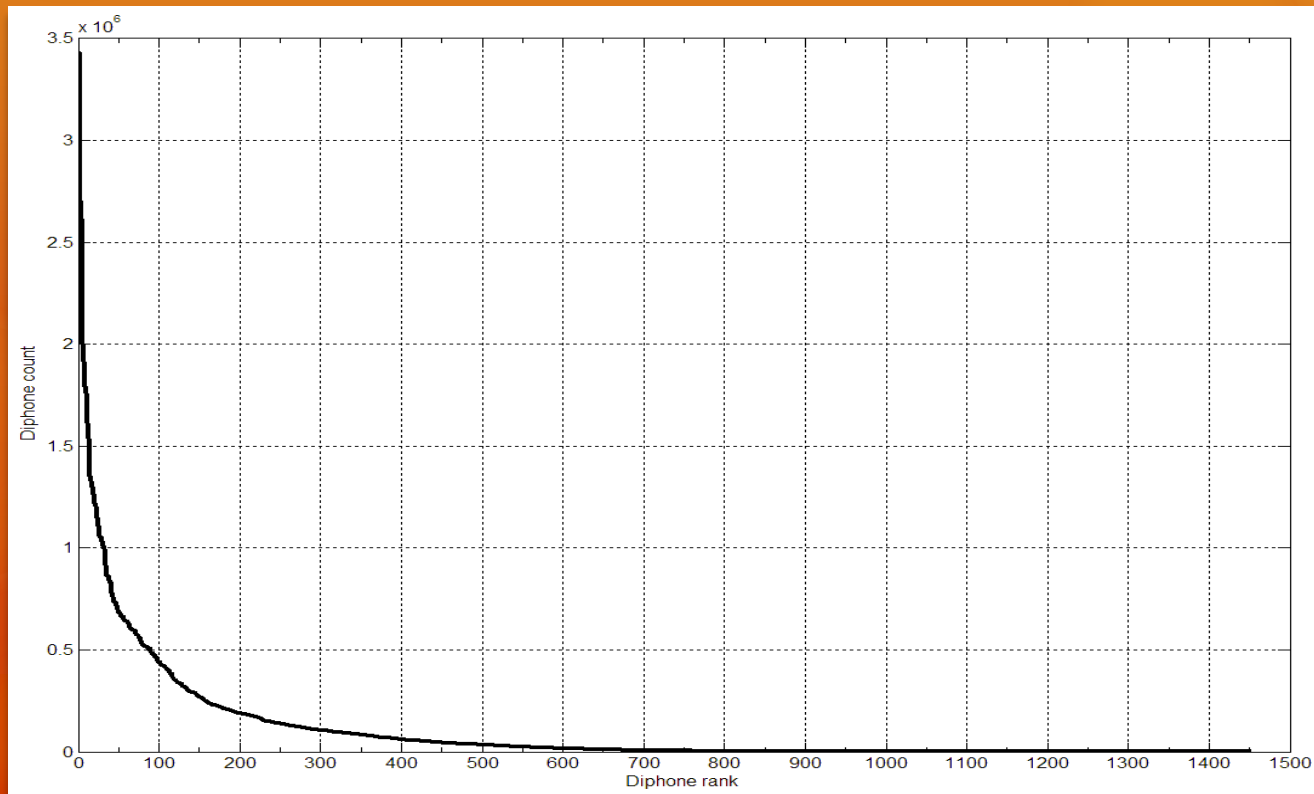
Statistical Results

○ Phoneme Frequency Counts:

Vowels	ɪ	18292597	Fricatives	s	6347603	Affricates	ts	1700828
	e	15448552		f	3034957		ɟʒ	1141271
	ɛ	7778560		h	2162600		tʃ	951329
	ʊ	7618576		ʃ	1658811		dʒ	8632
	ɔ	5096767		z	1048078			
	e:	3195630		v	989300	Nasals	n	9752059
	ɪ:	2226554		ʒ	5502		m	6645891
	ɛ:	756956						
	ɔ:	173623	Plosives	t	12253833	Liquids	l	12560881
	i:	171468		k	4470418		r	7656107
	à	95733		d	4148424			
	u:	74202		p	3242782	Glides	j	4629206
	ò	7051		b	2512670		w	1076580
	ù	3897		ʔ	1702567			
	è	3403		g	821119	Silence	#	5311123
	ì	304						

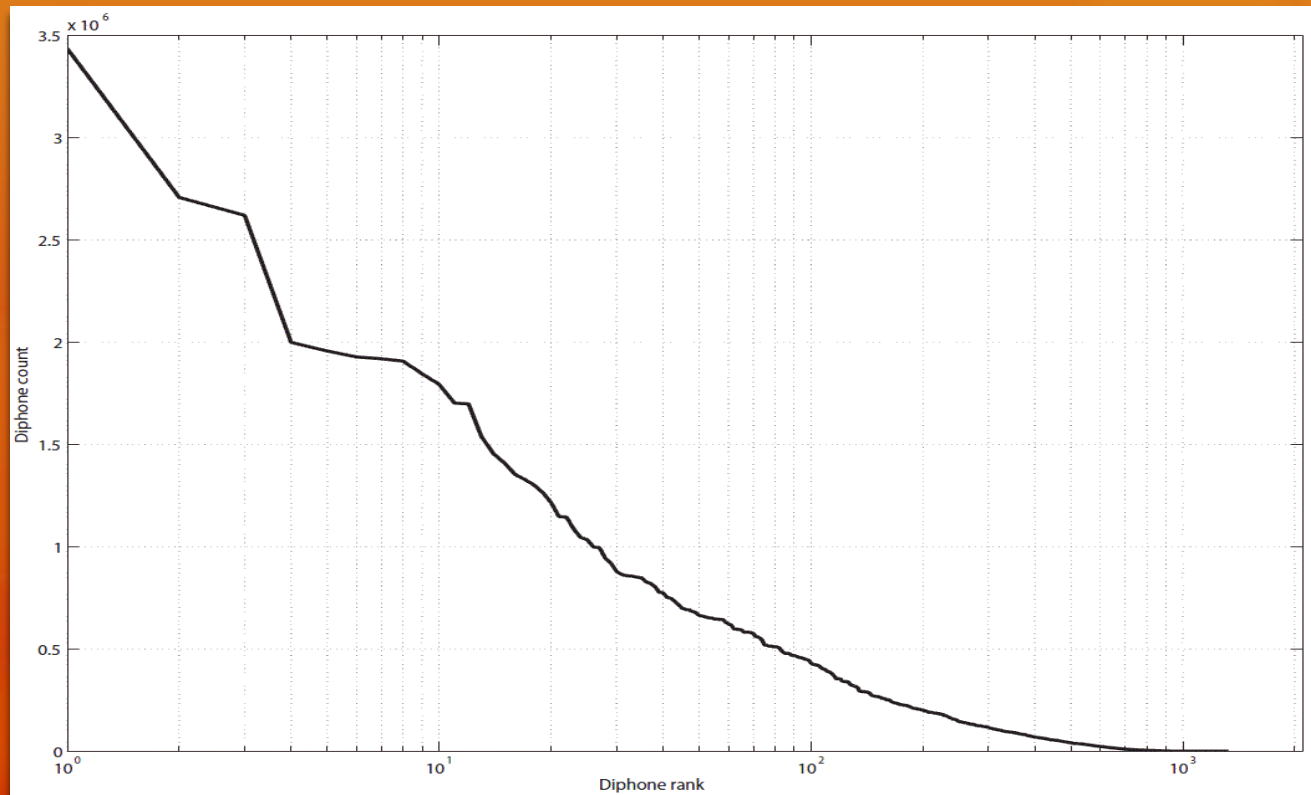
Diphone Statistics

- The potential number of distinct diphones is 1681
 - (1681=41x41 possible phoneme combinations)
- A total of 1450 distinct diphones were found in the corpus
- Diphones ranked by their frequency counts:



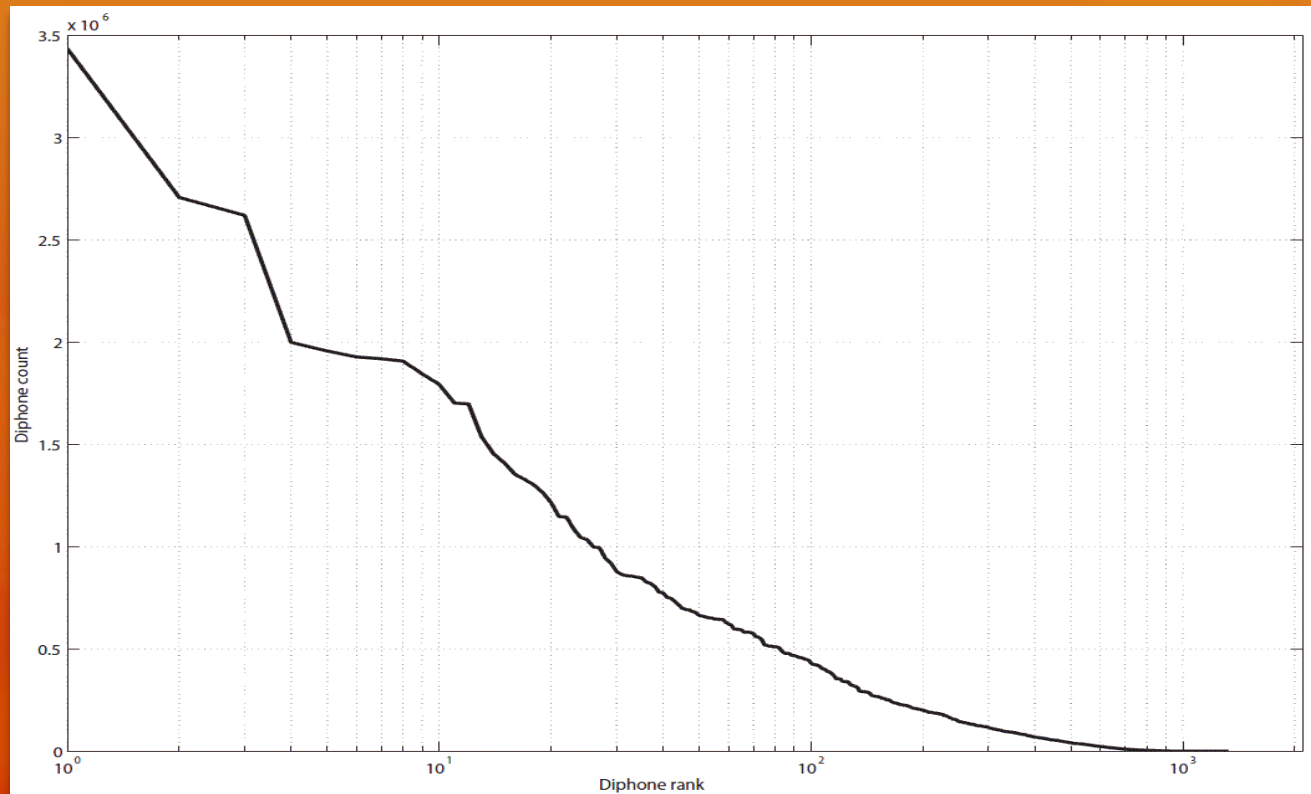
Diphone Statistics

- The potential number of distinct diphones is 1681
 - (1681=41x41 possible phoneme combinations)
- A total of 1450 distinct diphones were found in the corpus
- Diphones ranked by their frequency counts:



Diphone Statistics

- Manual check on last few hundred diphones
 - Last 101 diphones are a result of transcription errors, unfiltered foreign words, etc.
- Final number of distinct Maltese diphones stands at 1349 (80% of the potential phoneme combinations)



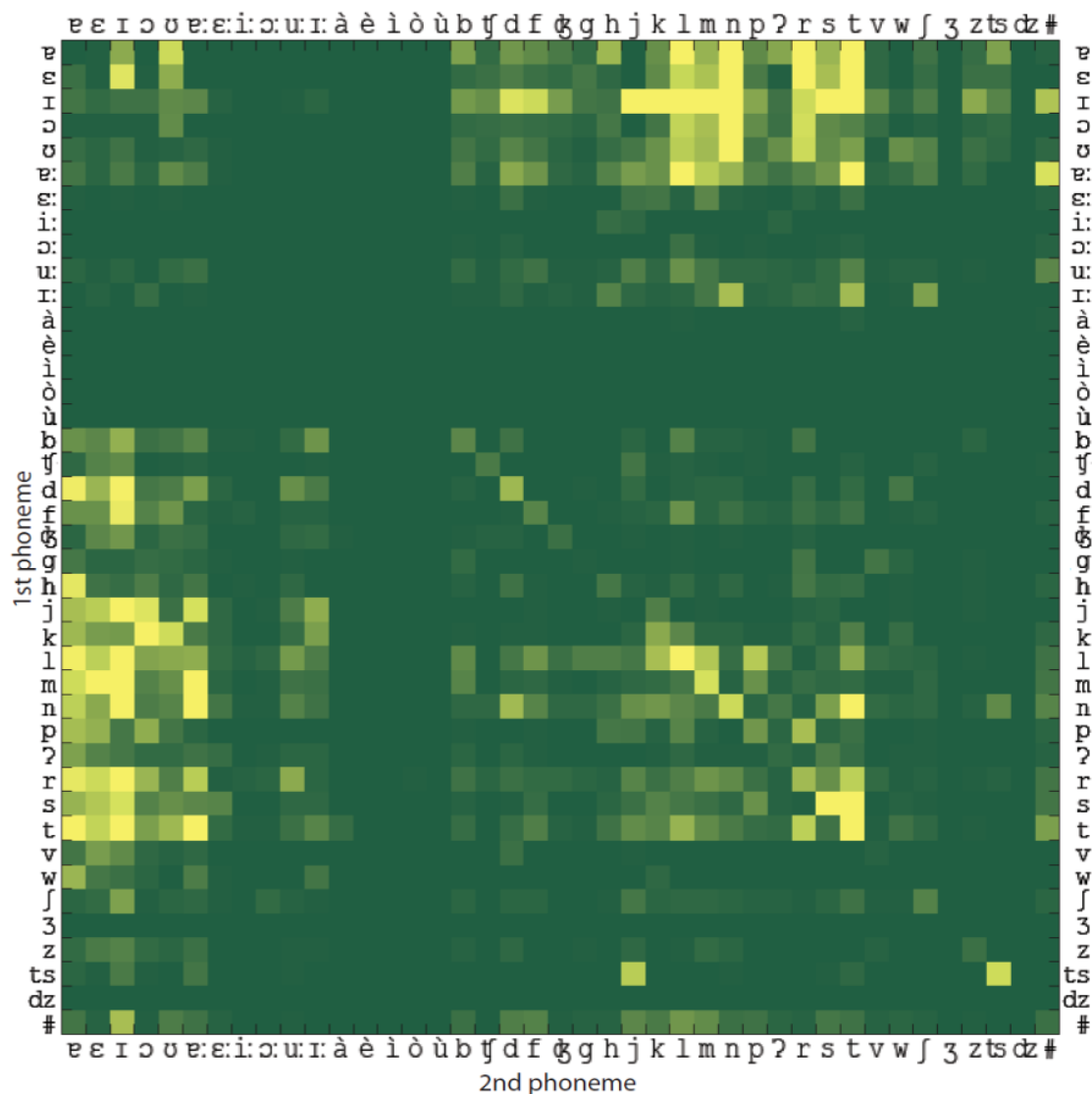
Diphone Statistics

- Diphones ranked by frequency count
- The statistics show:
 - The top 71 diphones from the 1349 distinct diphones account for 50% of all diphones in the corpus
 - The top 322 diphones account for 90% of all diphones in the corpus

diphone	count	%	Cumulative %
l+r	3435162	2.24	2.24
t+e	2791679	1.82	4.06
r+l	2707466	1.76	5.82
r+n	2619469	1.71	7.53
e+l	2538302	1.65	9.18
n+r	1998740	1.30	10.49
r+s	1918389	1.25	11.74
t+r	1907160	1.24	12.98
r+t	1793403	1.17	14.15
e+r	1762632	1.15	15.30
l+l	1698126	1.11	16.40
e+t	1614533	1.05	17.46
m+r	1537180	1.00	18.46
s+t	1454783	0.95	19.41
t+t	1353185	0.88	20.29
n+t	1328681	0.87	21.15
r+e	1319982	0.86	22.01
r+j	1301130	0.85	22.86
r+r	1264935	0.82	23.69
n+e	1242290	0.81	24.50
o+n	1218616	0.79	25.29
e+n	1206091	0.79	26.08
e+n	1149137	0.75	26.82
e+r	1146904	0.75	27.57
d+e	1106694	0.72	28.29
t+e:	1092714	0.71	29.01
m+e	1059509	0.69	29.70
j+e	1055206	0.69	30.38

Table 6: Most frequent diphones in Maltese

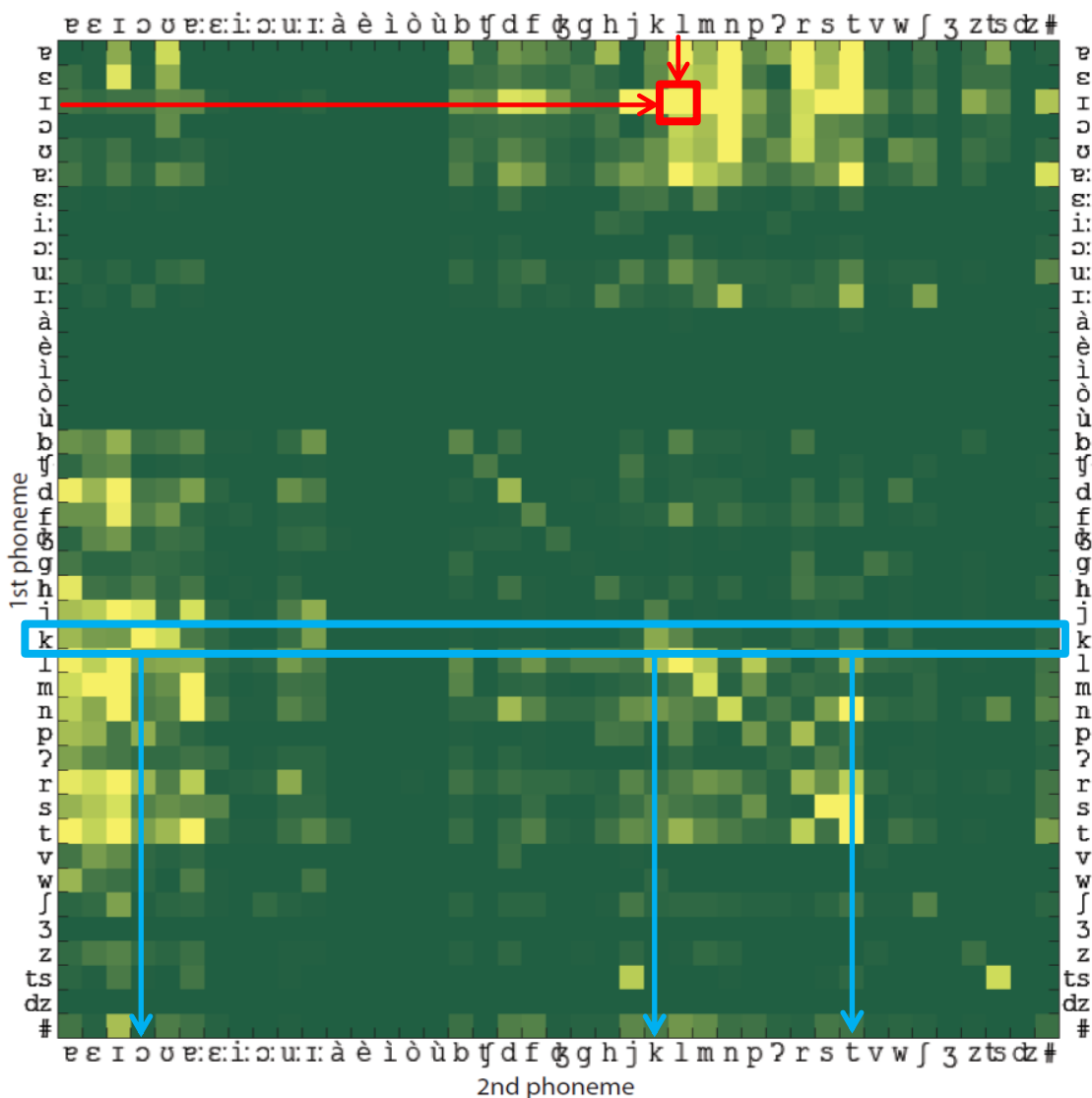
Diphone Statistics



diphone	count	%	Cumulative %
l+r	3435162	2.24	2.24
t+v	2791679	1.82	4.06
r+l	2707466	1.76	5.82
r+n	2619469	1.71	7.53
e+l	2538302	1.65	9.18
n+r	1998740	1.30	10.49
r+s	1918389	1.25	11.74
t+r	1907160	1.24	12.98
r+t	1793403	1.17	14.15
e+r	1762632	1.15	15.30
l+l	1698126	1.11	16.40
e+t	1614533	1.05	17.46
m+r	1537180	1.00	18.46
s+t	1454783	0.95	19.41
t+t	1353185	0.88	20.29
n+t	1328681	0.87	21.15
r+v	1319982	0.86	22.01
r+j	1301130	0.85	22.86
r+r	1264935	0.82	23.69
n+e	1242290	0.81	24.50
o+n	1218616	0.79	25.29
e+n	1206091	0.79	26.08
ε+n	1149137	0.75	26.82
ε+r	1146904	0.75	27.57
d+e	1106694	0.72	28.29
t+e:	1092714	0.71	29.01
m+e	1059509	0.69	29.70
j+v	1055206	0.69	30.38

6: Most frequent diphones in Maltese

Diphone Statistics

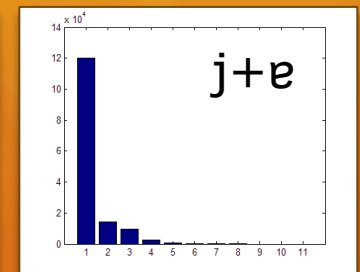
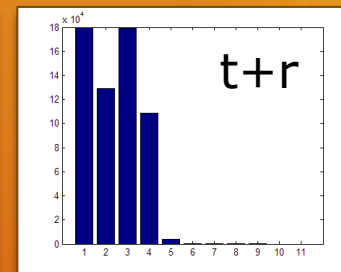
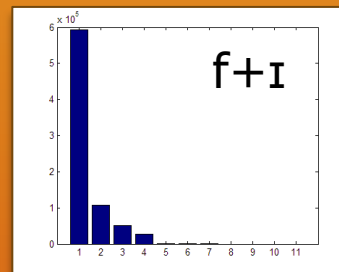
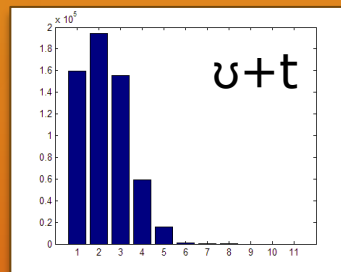
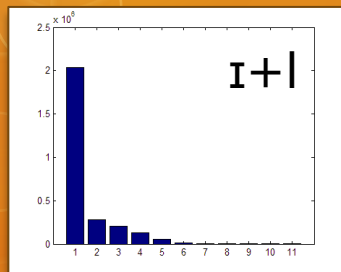


diphone	count	%	Cumulative %
l+r	3435162	2.24	2.24
t+v	2791679	1.82	4.06
i+l	2707466	1.76	5.82
r+n	2619469	1.71	7.53
e+l	2538302	1.65	9.18
n+r	1998740	1.30	10.49
r+s	1918389	1.25	11.74
t+i	1907160	1.24	12.98
r+t	1793403	1.17	14.15
e+r	1762632	1.15	15.30
l+l	1698126	1.11	16.40
e+t	1614533	1.05	17.46
m+i	1537180	1.00	18.46
s+t	1454783	0.95	19.41
t+t	1353185	0.88	20.29
n+t	1328681	0.87	21.15
r+v	1319982	0.86	22.01
r+j	1301130	0.85	22.86
r+r	1264935	0.82	23.69
n+e	1242290	0.81	24.50
o+n	1218616	0.79	25.29
e+n	1206091	0.79	26.08
ε+n	1149137	0.75	26.82
ε+r	1146904	0.75	27.57
d+v	1106694	0.72	28.29
t+e:	1092714	0.71	29.01
m+e	1059509	0.69	29.70
j+v	1055206	0.69	30.38

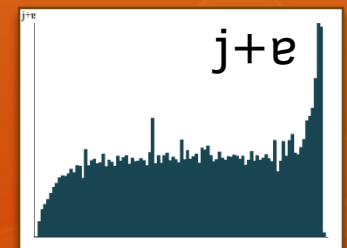
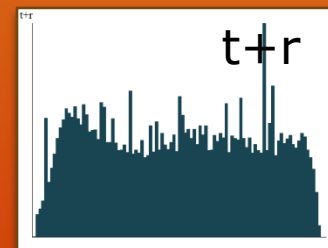
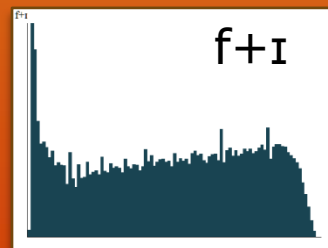
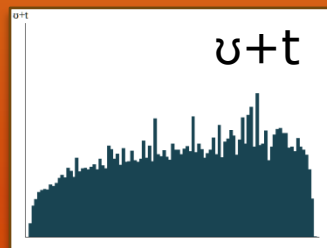
6: Most frequent diphones in Maltese

Diphone Statistics

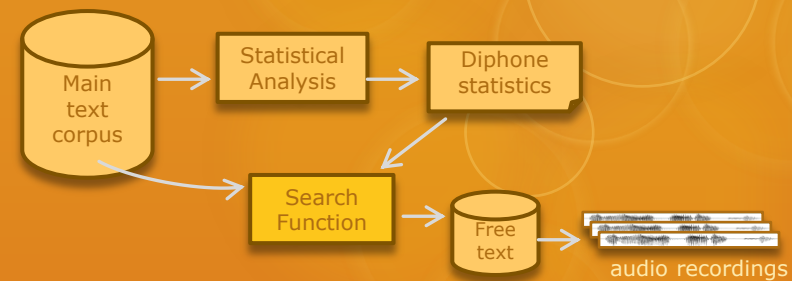
- Distribution of positions of diphones in word syllables:



- Distribution of positions of diphones in phrases:

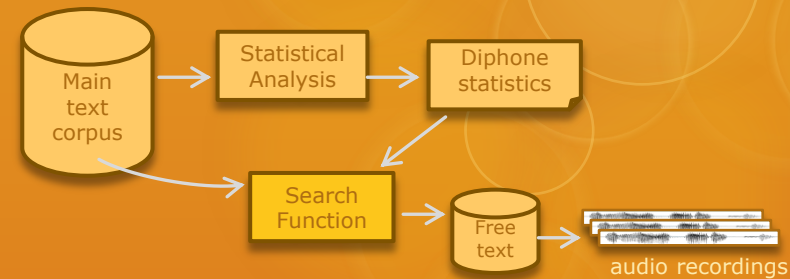


Free Text Selection



- After analysing the main corpus, the gathered statistics are used to compile a descriptor space containing the identifying features of this global text
 - Feature vectors:
⟨diphone, position score, frequency score⟩

Free Text Selection



- Automated selection process:

- Iterative process

- ○ The corpus is divided into text blocks of equal word count (rounded up to the nearest sentence)
- The text blocks are analysed, their feature vectors compiled, and their scores are ranked
- The text block with highest score corresponds to that text block which has features most similar to the features of the full text corpus
- The top entry is composited into a selection which contains all the top entries from previous iterations
- Text block size gets shorter during subsequent iterations
- Process finishes when the desired free text size is reached

Ranking Score

- Ranking score Δ is based on 2 important diphone features: position score ϕ_s , and frequency score ψ_s

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

where:

ϕ_w, ψ_w are weights for the position and frequency components respectively

Ranking Score

- Diphone frequency score ψ_s :
 - the ratio of the distinct diphone occurrences between the text block being analysed, and the global corpus
- Diphone position score ϕ_s :

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

where:

g = diphone count in global corpus

D_l = set of diphones occurring in the text block

$\phi_s(d)$ = diphone position score for individual diphone d

Ranking Score

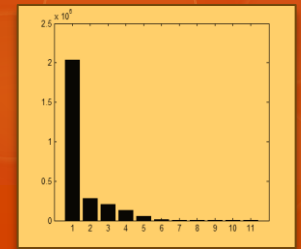
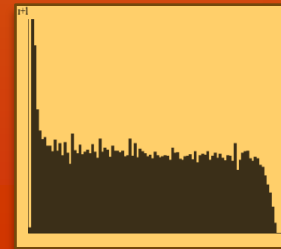
- Position score $\phi_s(d)$ for diphone d is given by:

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

where:

λ_s = phrase position score (and weight λ_w)

μ_s = syllable position score (and weight μ_w)



Ranking Score

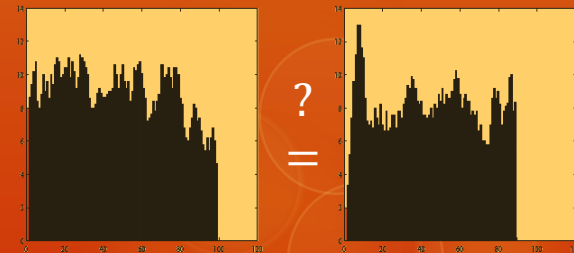
- Both phrase position score λ_s , and syllable position score μ_s are similarity scores computed using a scale invariant method on the position histograms of phrases and syllables respectively:

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

where:

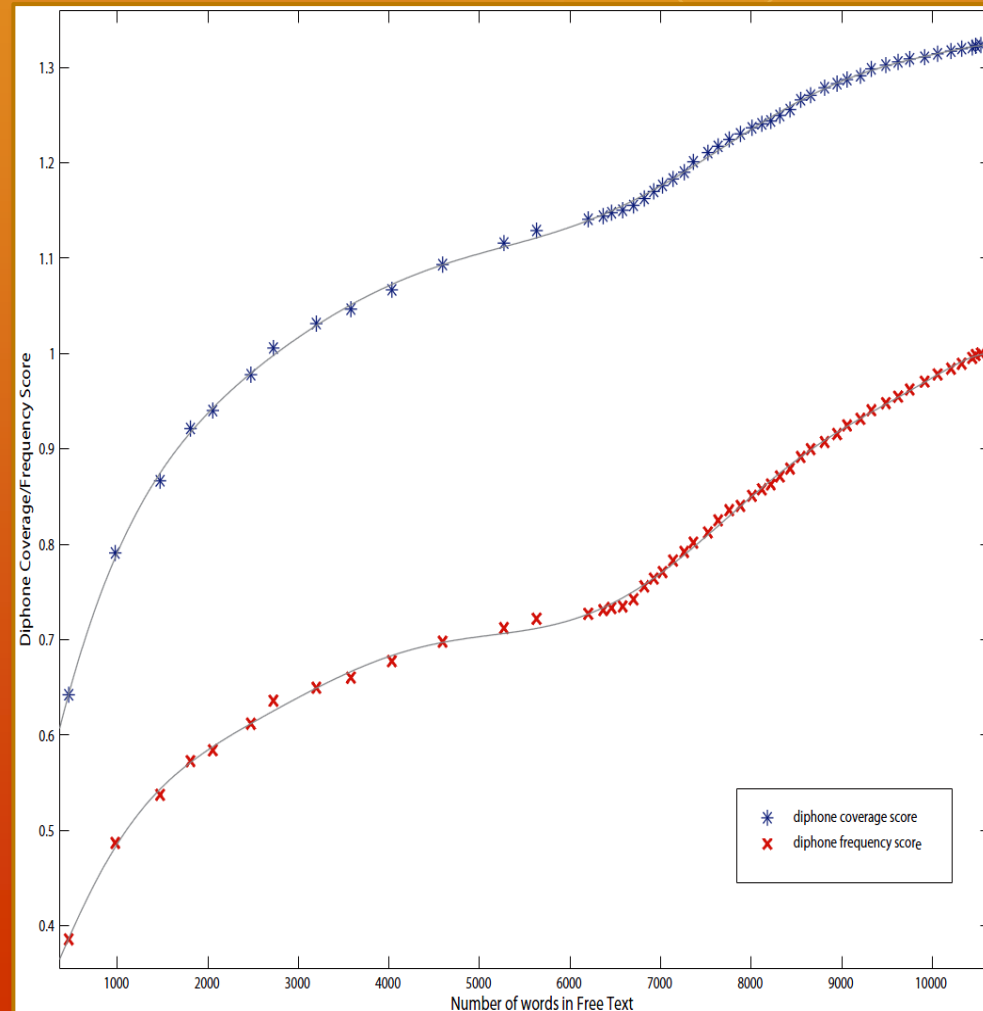
G_d^k , $k \in \{\lambda, \mu\}$ are the histograms for the global syllable and phrase positions for diphone d

L_d^k , $k \in \{\lambda, \mu\}$ are the histograms for the local (i.e. text block's) syllable and phrase positions for diphone d



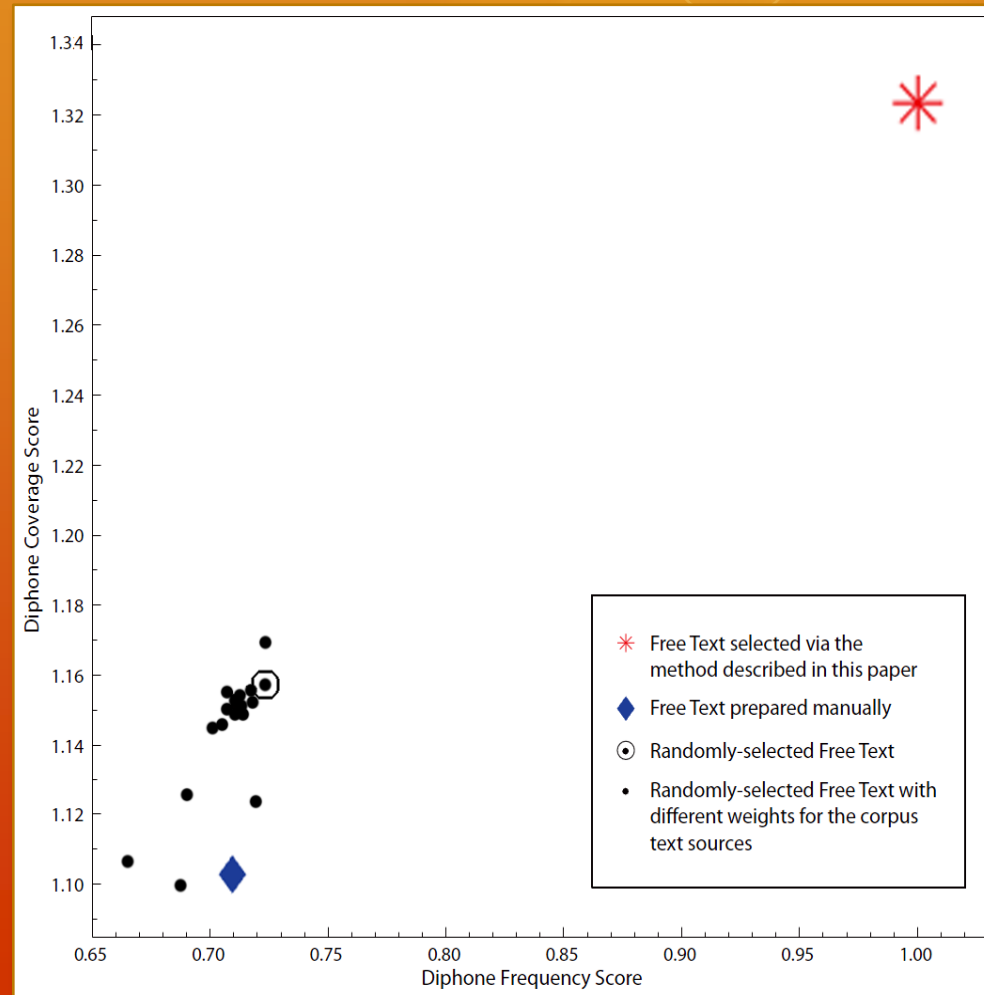
Free Text Selection Results

- Compiled a free text of approx. 10,000 words
- 50 iterations
- Initial text block size: 500 words
- At around the 6,500-word mark, the varying block size goes down to 1 sentence in size
- Final ranking score:
 - $\Delta(\phi_S, \phi_W, \psi_S, \psi_W) = 1.32$
(max $\Delta = 1.41$)
 - $\psi_S = 1.0$

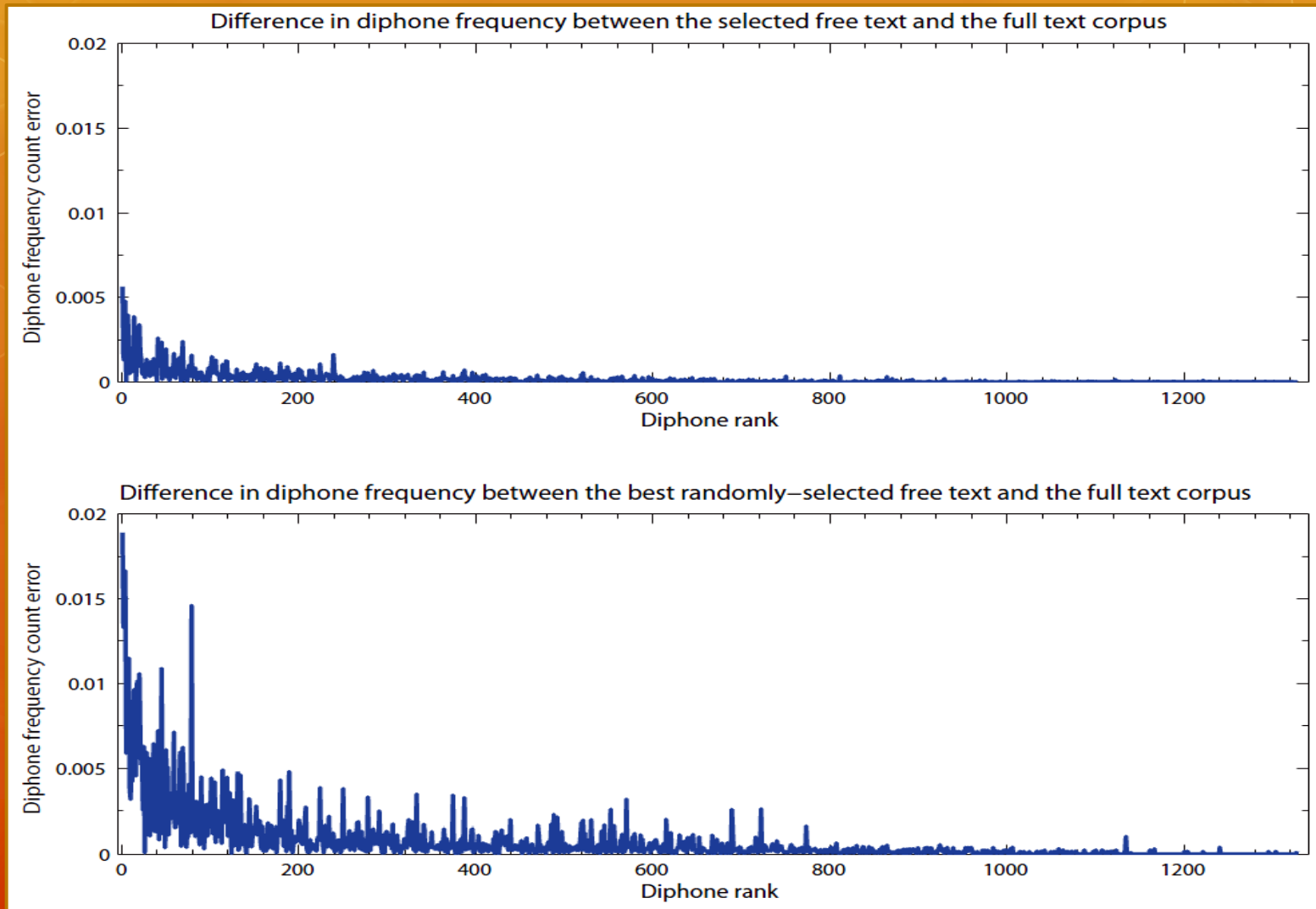


Evaluation of Free Text Selection

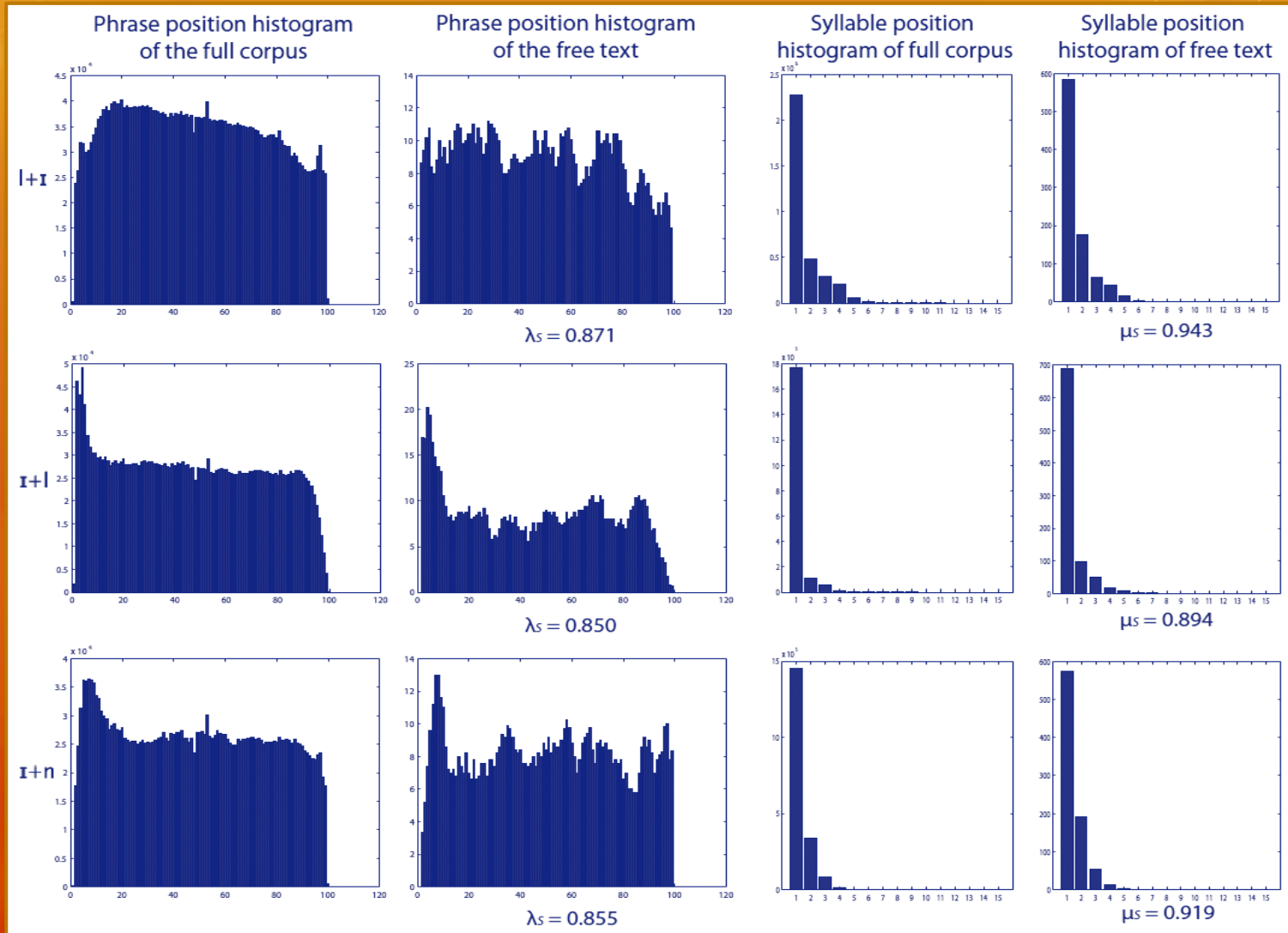
- Evaluation of our free text selection method against:
 1. Random free text selection method
 2. Weighted random free text selection method
 3. Manually-generated text (by an expert)
- Our selection method outperforms all the others
- Note that 10,000 words \ll 33 million words ($\approx 0.03\%$)



Evaluation of Free Text Selection



Free Text Selection Results




Conclusion

- We presented details on the generation of a free-running text corpus for Maltese concatenative speech synthesis
- We performed statistical analysis to obtain diphone statistics for Maltese text
- We developed a novel automated selection algorithm in the compilation of the free text corpus
- This algorithm defines the diphone coverage measure as a weighted combination of diphone frequencies and their syllable and phrasal positions
- As a result we have achieved a greater diphone coverage than other standard methods like random or manual free text generation
- Both the free text corpus and the statistics collected during this study will be directly applied to the ongoing development of the Maltese Text-to-Speech engine

Grazzi - Thanks



Synthesised Speech

Phrase	Synthesised
Dan mhux xogħol ħafif, imma jrid isir	
Ġmielhom mela	
Il-kanzunetta kullimkien	
Il-melħa tal-kultura	
Imma mhux magħquda bejniethom	
Kapitlu Sbatax	
L-Insara baqagħlhom f'idejhom xi bliet u fortizzi	
Mhux biss il-baħar hu mqalleb	
Tesawru tas-sentimenti	
U à propositu	