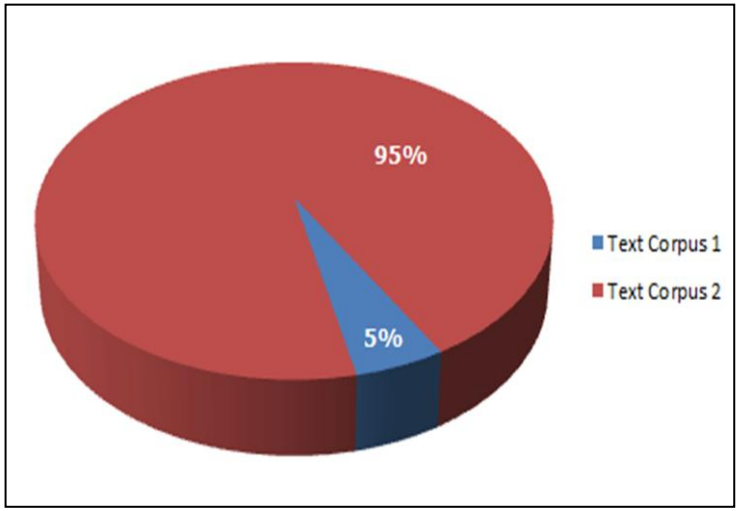
MALTESE DIPHONE STATISTICAL ANALYSIS (2/June/2010)

The Text Corpora

Text Corpus 1 – Maltese Wikipedia web pages	
URL:	http://mt.wikipedia.org/wiki/****
File Format:	HTML text, converted to text files.
Encoding used:	UTF8 Encoding
Total pages:	3,551
Total # of words:	1,052,556

Text Corpus 2 – Parliament Debates	
URL:	http://www.parliament.gov.mt/file.aspx?f=****
File Format:	MS Word document, converted to text files.
Encoding used:	UTF8 Encoding
Total files downloaded:	1,360
Total # of words:	20,095,349

TOTAL	
Total # of words:	**21,147,905**
Total # of sentences detected:	**914,683**



Description of the Analysis Process

Below is a brief description of the processing and analysis steps performed:

1. Pre-processing of text corpus:
   - 1.1. Converting text to Unicode UTF-8 encoding.
   - 1.2. Character normalisation
     - 1.2.1. Replacing HTML codes
       - 1.2.1.1. E.g., & → &, (→ (, Á → Á
     - 1.2.2. Normalisation of quote and double-quote characters
       - 1.2.2.1. E.g., ‘, ’, ` → ', “, ”, " → ".
     - 1.2.3. Replacement of non-letter characters (e.g., #, ~, \t, etc.) with space.
   - 1.3. Filtering out of emails and URLs
   - 1.4. Filtering out numbers and replacing them by a fixed token containing no decimal point.
   - 1.5. Detecting abbreviations and acronyms, and replacing them by a fixed token
     - 1.5.1. Comparing against a pre-defined list of known abbreviations and acronyms defined in file “_config\abbreviation list.txt”
       - 1.5.1.1. E.g., Mr., Dr., eż., p.eż., Nru., Ltd., HSBC, MCAST, etc.
       - 1.5.1.2. Handling also cases where the abbreviations and acronyms are pre-posed with the definite article and/or prepositions:
         - 1.5.1.2.1. E.g., il-Mons., l-Onor., is-CMTU, bħall-HSBC, mal-HSBC, għall-GDPs, etc.
     - 1.5.2. Detecting abbreviations consisting of a sequence of one letter followed by a fullstop.
       - 1.5.2.1. E.g., U.S., i.e., Q.K., or initials like M.
   - 1.6. Filtering out ellipsis (single-character: …, or made up of consecutive fullstops: . . .)
   - 1.7. Filtering out exception words and phrases defined in file “_config\exception list.txt”, and that have not been caught in the previous processing steps.
     - 1.7.1.E.g., IĊ-CHAIRMAN, L-Acting Speaker:, roman numerals like: XVII, MDXX, and Maltese surnames.

2. Splitting of text corpus into sentences
   - 2.1. Sentence boundary detection
     - 2.1.1.Via the detection of the sentence termination characters: ., !, ?.
       - 2.1.1.1. Handling exception cases involving sentences containing abbreviations ending with a fullstop and that have not been caught in earlier processing.
       - 2.1.1.2. Handling exception cases where the detected sentence consists of just one word and this word starts with a small-case letter; most probably this sentence should be joined with the previous one.
   - 2.2. Sentences are saved to a sentence corpus file, called “<OUTPUT_DIR>\sentence_corpus”.

3. Word Analysis
   3.1. Splitting of a sentence into a sequence of words
      3.1.1. Handling special cases where words are hyphenated due to paragraph formatting on a sentence break (\n\r), or due to an incorrect whitespace appearing between the definite article and a word:
         3.1.1.1. E.g., għarraf- \n\r niehom → għarrafniehom, tal- kelb → tal-kelb.
   3.2. Pre-processing of words:
      3.2.1. Filtering out quotes and double quotes wrapping words:
         3.2.1.1. E.g., ' <word> ' → <word>
   3.3. Analysing structure of words to convert *surface forms* to *lexical forms*, i.e., to normalise words consisting of combinations of definite articles, prepositions and words to arrive at a unique word list[1].
      3.3.1. Matching the word pattern: <word1> - <word2> where <word1> is the definite article
         3.3.1.1. E.g., l-kelb → l- + kelb
         3.3.1.2. Handling cases of the definite article with *consonant assimilation* (*ittri xemxin:* ċ, d, n, r, s, t, x, z, ż):
            3.3.1.2.1. E.g., d-dar → d- + dar
         3.3.1.3. Removing any outer */i/-epenthesis* from the definite article:
            3.3.1.3.1. E.g., il-kelb → l- + kelb, iċ-ċentru → ċ- + ċentru,
      3.3.2. Matching the word pattern: <word1> ' <word2> where <word1> is one of the prepositions: fi, bi, ta', ma'.
         3.3.2.1. E.g., f'post → fi + post
         3.3.2.2. E.g., b'ommi → bi + ommi
         3.3.2.3. E.g., t'Anna → ta' + Anna
         3.3.2.4. E.g., m'ommu → ma' + ommu
      3.3.3. Matching the word pattern: <word1> - <word2> where <word1> is a combination of the definite article and the prepositions: ma', ta', sa, ġo.
         3.3.3.1. E.g., mal-kelb → ma' + l- + kelb
         3.3.3.2. E.g., tan-negozju → ta' + n- + negozju
         3.3.3.3. E.g., sad-disgħa → sa + d- + disgħa
         3.3.3.4. E.g., ġon-nar → ġo + n- + nar
      3.3.4. Matching the word pattern: <word1> - <word2> where <word1> is a combination of the definite article and the prepositions: bi, fi.
         3.3.4.1. E.g., fl-art → fi + l- + art
         3.3.4.2. E.g., fil-baħar → fi + l- + baħar
         3.3.4.3. E.g., bin-nuqqas → bi + n- + nuqqas

[1] Some of the rules used for this step are as per: M. Rosner , "Finite State Analysis of Prepositional Phrases in Maltese", Univ. of Malta, 2003.

3.3.5. Matching the word pattern: <word1> - <word2> where <word1> is a combination of the definite article and the prepositions: lil, għal, bħal, minn.

3.3.5.1. E.g., lil-Libja → lil + l- + Libja

3.3.5.2. E.g., lill-partit → lil + l- + partit

3.3.5.3. E.g., mix-xejn → minn + s- + xejn

3.3.6. Matching the word pattern: <word1> - <word2> where <word1> is a combination of the definite article and the prepositions: dan, din.

3.3.6.1. E.g., dil-mozzjoni → din + l- + mozzjoni

3.3.6.2. E.g., daċ-ċajt → dan + ċ- + ċajt

3.3.7. If <word2> in steps 3.3.1 to 3.3.6 contains a hyphen within itself, then split the word.

3.3.7.1. E.g., fl-eks-sistema → fi + l- + eks + sistema

3.3.8. Handle special cases where in the pattern <word1> - <word2>, <word1> or <word2> consists of numeric digits:

3.3.8.1. E.g., fl-1974 → fi + l- + 1974

3.3.8.2. E.g., 18-il → 18 + l-

3.3.8.3. E.g., mill-15-il → minn + l- + 15 + l-

3.4. Attempt to detect non-Maltese words and filter them out from the final word list.

3.4.1. Detect words containing non-Maltese Latin letters, e.g., y.

3.4.2. Detect words containing characters not within the following Unicode blocks:

3.4.2.1. Basic Latin

3.4.2.2. Latin 1 Supplement (contains accented characters such as á)

3.4.2.3. Latin Extended –A (contains the Maltese letters ċ, ħ, etc.)

3.5. Merge words that differ only by letter case

3.6. Words are saved to a word list file, called "<OUTPUT_DIR>\word_list".

3.6.1. Words flagged as foreign in step 3.4, are saved to the file called "<OUTPUT_DIR>\word_list_foreign".

3.6.2. Words containing hyphens and which do not match any of the patterns analysed in step 3.3, are saved to a file called "<OUTPUT_DIR>\word_list_unknown".

4. Diphone Analysis

4.1. Strip out any accents from words, except if an accent occurs on vowels at the end of words

4.1.1. E.g., però, komunità

4.1.2. Normalise the accents, by changing acute accents (e.g., á) to grave accents (à).

4.1.2.1. E.g., Awtoritá → Awtorità

4.2. Perform grapheme to phoneme conversion based on the LTS rules given further below (see section "LTS Rules")

4.3. Gather phone statistics

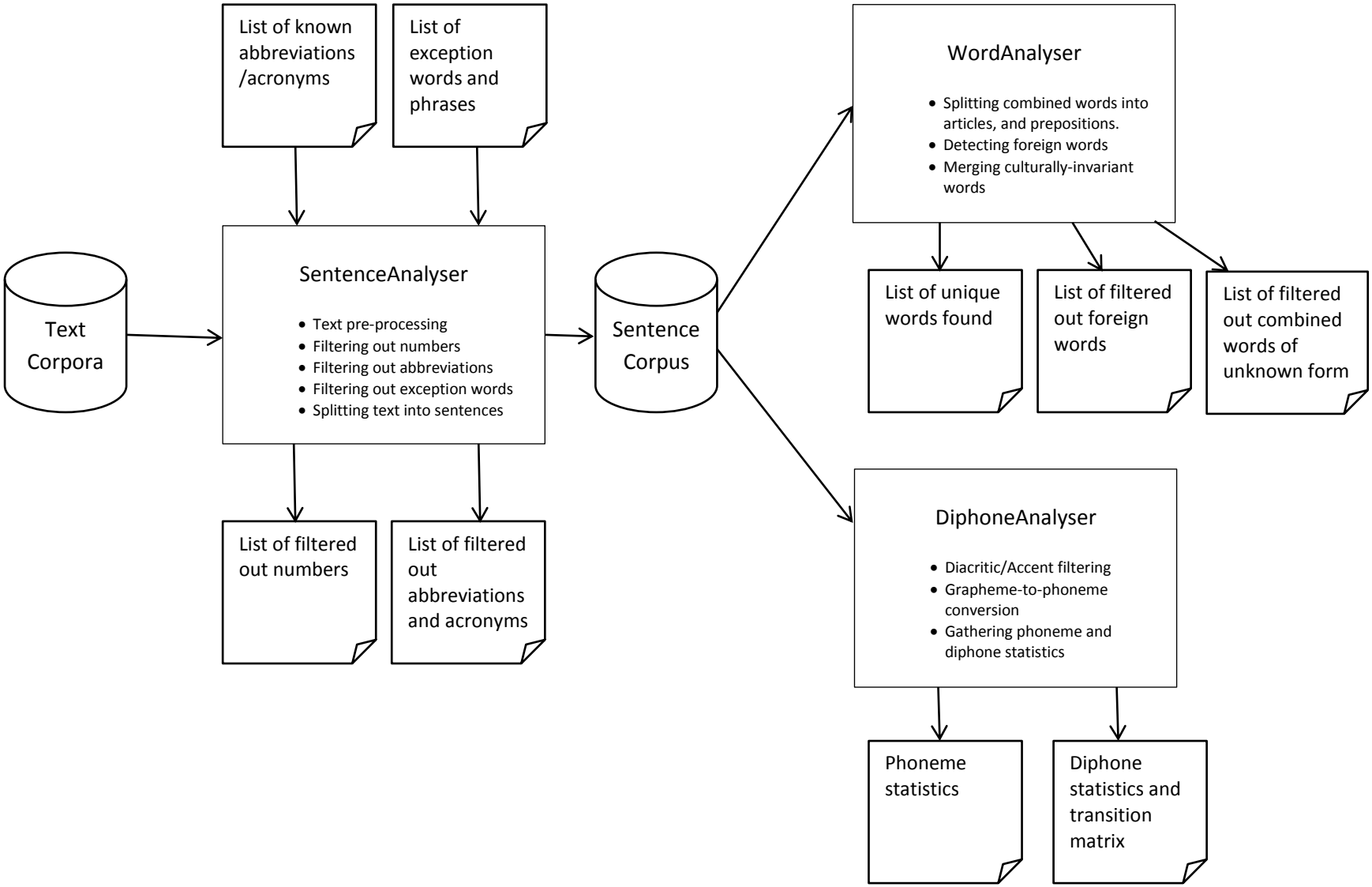
4.3.1. Phone statistics are saved to a file called "<OUTPUT_DIR>\phoneme_list".

4.4. Gather diphone statistics

4.4.1. Diphone statistics saved to a file called "<OUTPUT_DIR>\diphone_list".

4.4.2. Diphone statistics are also saved as a transition matrix in a file called "<OUTPUT_DIR>\diphone_transition_matrix".

4.4.3. A transition matrix of diphone statistics grouped by diphone class (e.g., vowel, affricate, etc.) is saved to a file called "<OUTPUT_DIR>\diphone_class_stats".



Phoneme List

The list of 43 phonemes used in this analysis (42 actual phonemes plus silence symbol) is given below:

vowels	
æ	a
ɛ	e
ɪ	i
ɔ	o
ʊ	u

half-length vowels	
ɪˑ	half-long i
ʊˑ	half-long u

lengthened vowels	
æː	long a
ɛː	long e
iː	long i
ɔː	long o
uː	long u
ɪː	ie

accented vowels	
à	accented i
è	accented e
ì	accented i
ò	accented o
ù	accented u

consonants	
b	b
ʧ	ċ
d	d
f	f
ʤ	ġ
g	g
h	ħ
j	j
k	k
l	l
m	m
n	n
p	p
ʔ	q
r	r
s	s
t	t
v	v
w	w
ʃ	x (1)
ʒ	x (2)
z	ż
ts	z (1)
dz	z (2)

silence symbol	
@	silence

Note:

- The silence symbol is used at the beginning and end of a sentence.
- Phonemes /ɪˑ/ and /ʊˑ/ represent half-length vowels, a length between /ɪ/ and /iː/, /ʊ/ and /uː/ respectively. So far the rules governing their use are not clear, and hence they have been left out for now in this analysis.

LTS Rules

The following are the grapheme-to-phoneme (or letter-to-sound, LTS) rules used in this analysis[2].

The rules are applied in the order given here (from specific rule to more generic rule order).

The matching pattern string is a regular expression string with 3 character look-back.

- Example 1: Given the pattern of rule 16 (i.e., "...ieh[bċdfġgħjklmnpqrstvwxżz]"), means that if the grapheme "ie" is followed by "h" and then followed by one of the characters in [bċdfġgħjklmnpqrstvwxżz] (i.e., a consonant), then the grapheme "ie" maps to the phoneme /ɛː/ (long ɛ).
- Example 2: Given the pattern of rule 30 (i.e., ".għa"), means that if the grapheme "a" is preceded by the 2 letters "għ", then the grapheme "a" maps to phoneme /ɐː/ (long a). A dot means 'any character', so in this case the third character before the grapheme "ie" can be anything, including a word boundary indicator.
- Example 3: Given the pattern of rule 32 (i.e., "..[bċdfġgħjklmnpqrstvwxżz]u#"), means that if the grapheme "u" is preceded by a consonant (one of the letters in the set [bċdfġgħjklmnpqrstvwxżz]), and if the grapheme "u" occurs at the end of a word (i.e., followed by the word boundary indicator "#"), then it is mapped to the phoneme /uː/ (long u).

Rule #	Grapheme(s)		Phoneme(s)	Matching pattern (regular expression string with 3 character look-back)	
1	għu	→	/ɔ/ /ʊ/	…għu	Diphthongs
2	ow	→	/ɔ/ /ʊ/	…ow	
3	oj	→	/ɔ/ /ɪ/	…oj	
4	iw	→	/ɪ/ /ʊ/	…iw	
5	għi	→	/ɛ/ /ɪ/	…għi	
6	ej	→	/ɛ/ /ɪ/	…ej	
7	ew	→	/ɛ/ /ʊ/	…ew	
8	aj	→	/ɐ/ /ɪ/	…aj	
9	għu	→	/ɐ/ /ʊ/	…għu	
10	aw	→	/ɐ/ /ʊ/	…aw	
11	aho	→	/ɔː/	...aho	Vowel lengthening
12	ogħo	→	/ɔː/	...ogħo	
13	o	→	/ɔː/	.għo	
14	o	→	/ɔː/	...ogħ	

[2] The LTS rules are taken from: Paulseph-John Farrugia, "Text to Speech Technologies for Mobile Telephony Services", MSc Thesis, University of Malta, 2005.
And have been amended to take into consideration the differences in the set of phonemes used here and those used in the above referenced source.

Rule #	Grapheme(s)		Phoneme(s)	Matching pattern (regular expression string with 3 character look-back)	
15	i	→	/iː/	...igħ | ...i[ħhq]	
16	ie	→	/ɛː/	...ieh[bċdfġgħjklmnpqrstvwxżz]	
17	ie	→	/ɛː/	...iegħ[bċdfġgħjklmnpqrstvwxżz]	
18	ehi	→	/ɛː/	...ehi	
19	ehe	→	/ɛː/	...ehe	
20	e	→	/ɛː/	..he	
21	e	→	/ɛː/	...eh	
22	egħi	→	/ɛː/	...egħi	
23	egħe	→	/ɛː/	...egħe	
24	e	→	/ɛː/	.għe	
25	e	→	/ɛː/	...egħ	
26	aha	→	/ɐː/	...aha	
27	a	→	/ɐː/	..ha	
28	a	→	/ɐː/	...ah	
29	agħa	→	/ɐː/	...agħa	
30	a	→	/ɐː/	.għa	
31	a	→	/ɐː/	...agħ	
32	u	→	/uː/	..[bċdfġgħjklmnpqrstvwxżz]u#	
33	o	→	/ɔː/	..[bċdfġgħjklmnpqrstvwxżz]o#	
34	e	→	/ɛː/	..[bċdfġgħjklmnpqrstvwxżz]e#	
35	a	→	/ɐː/	..[bċdfġgħjklmnpqrstvwxżz]a#	
36	à	→	/à/	...à#	Accented vowels
37	è	→	/è/	...è#	
38	ì	→	/ì/	...ì#	
39	ò	→	/ò/	...ò#	
40	ù	→	/ù/	...ù#	
41	ie	→	/ɪː/	...ie	Default vowels
42	u	→	/ʊ/	...u	
43	o	→	/ɔ/	...o	
44	i	→	/ɪ/	...i	
45	e	→	/ɛ/	...e	
46	a	→	/ɐ/	...a	
47	j	→	/j/	…j	Glides / Semi-vowels
48	w	→	/w/	…w	
49	r	→	/r/	…r	Liquids
50	l	→	/l/	…l	

Rule #	Grapheme(s)		Phoneme(s)	Matching pattern (regular expression string with 3 character look-back)	
51	n	→	/m/	…np | …nb	Nasals
52	n	→	/n/	…n	
53	m	→	/m/	…m	
54	v	→	/f/	...v#	Fricatives
55	f	→	/v/	...f[ptkqfsxħzċ]	
56	v	→	/f/	...v[ptkqfsxħzċ]	
57	v	→	/v/	...v	
58	f	→	/f/	...f	
59	ż	→	/s/	...ż#	
60	s	→	/z/	...s[bdgvżġmnlrwj]	
61	ż	→	/s/	...ż[ptkqfsxħzċ]	
62	ż	→	/z/	...ż	
63	s	→	/s/	...s	
64	x	→	/ʒ/	...x[bdgvżġ]	*...not sure about this!!!*
65	x	→	/ʃ/	...x	
66	għ	→	/h/	...għ	
67	h	→	/h/	...h#	
68	għ	→	/h/	...għ#	
69	ħ	→	/h/	...ħ	
70	z	→	/dz/	...z[bdgvżġ]	Plosives and Affricates
71	z	→	/ts/	...z	
72	ġ	→	/ʧ/	...ġ#	
73	ċ	→	/dʒ/	...ċ[bdgvżġ]	
74	ġ	→	/ʧ/	...ġ[ptkqfsxħzċ]	
75	ġ	→	/dʒ/	...ġ	
76	ċ	→	/ʧ/	...ċ	
77	q	→	/ʔ/	...q	
78	g	→	/k/	...g#	
79	g	→	/k/	...g[ptkqfsxħzċ]	
80	g	→	/g/	...g	
81	k	→	/k/	...k	
82	d	→	/t/	...d#	
83	d	→	/d/	...d	
84	t	→	/t/	...t	
85	p	→	/b/	...p[bdgvżċ]	
86	b	→	/p/	...b[ptkqfsxħzċ]	

Rule #	Grapheme(s)		Phoneme(s)	Matching pattern (regular expression string with 3 character look-back)	
87	b	→	/p/	...b#	
88	b	→	/b/	...b	
89	p	→	/p/	...p	
90	h	→		...h	Silent h
??	*??*	→	/ɪˑ/	*??*	*...rule unknown?*
??	*??*	→	/ʊˑ/	*??*	*...rule unknown?*

Word Analysis Results

Some word analysis results after filtering out numbers, abbreviations and acronyms, detecting foreign words, and normalising words consisting of a combination of definite articles, prepositions and nouns. The aim of the end result is to arrive at a list of distinct Maltese words (This is saved to file "<OUTPUT_DIR>\word_list".)

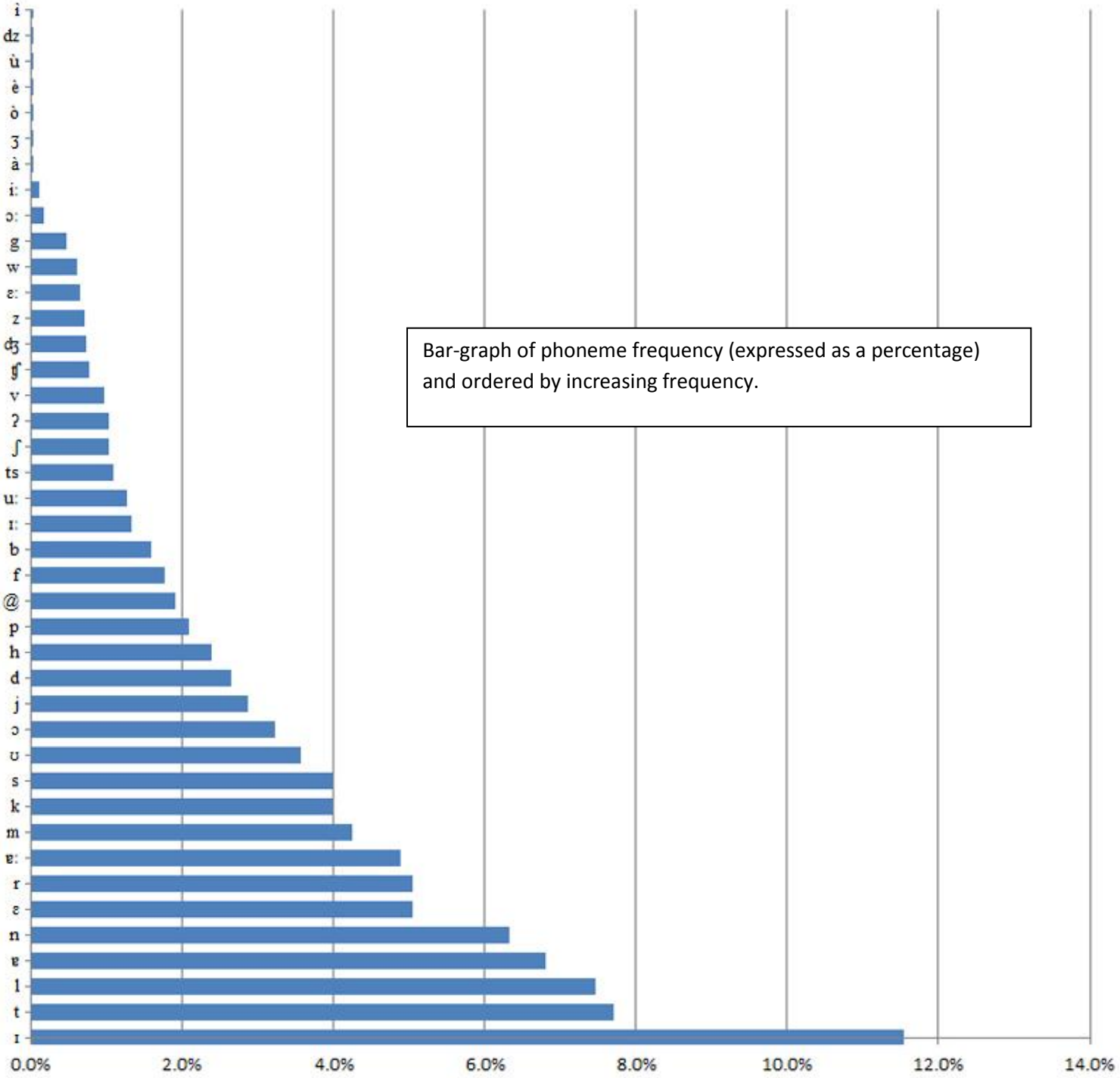
Initial number of words in corpus	21,147,905		
Total # of "numbers" words filtered out	576,939	Of which 45,554 are distinct	
Total # of abbreviations/acronyms filtered out	190,900	Of which 5191 are distinct	
Total # of foreign words detected and filtered out	305,134	Of which 5653 are distinct	
Remaining number of valid Maltese words	20,016,021		
After splitting hyphenated words where one of the sub-words is a definite article and/or preposition combination.	24,912,656	Of which 182,419 are distinct	
Total # of filtered out hyphenated words that do not match any of the patterns given in step 3.3.	58,911	Of which 8371 are distinct	Mostly consisting of unrecognised mixed alpha numeric symbols, e.g., 999'N, ax2, CO_2, H1N1, il-166A, etc.
Final number of Maltese words	**24,853,745**	**Of which 174,048 are distinct**	

Phoneme Analysis Results

A table of phoneme frequency obtained from the corpus is given below. In total 114,774,751 phonemes were found in the text corpus.

phoneme	phonetic class	count	%
ɪ	Vowel	13256494	11.6%
t	Plosive	8851052	7.7%
l	Lateral Approximant	8560271	7.5%
ɐ	Vowel	7809202	6.8%
n	Nasal	7250758	6.3%
ɛ	Vowel	5798085	5.1%
r	Retroflex	5791569	5.0%
ɐː	Vowel	5597687	4.9%
m	Nasal	4864204	4.2%
k	Plosive	4569808	4.0%
s	Fricative	4566095	4.0%
ʊ	Vowel	4095668	3.6%
ɔ	Vowel	3710980	3.2%
j	Approximant	3301290	2.9%
d	Plosive	3027915	2.6%
h	Fricative	2733179	2.4%
p	Plosive	2389673	2.1%
@	Silence	2200179	1.9%
f	Fricative	2028929	1.8%
b	Plosive	1816361	1.6%
ɪː	Vowel	1525638	1.3%

phoneme	phonetic class	count	%
uː	Vowel	1464086	1.3%
ts	Affricate	1258319	1.1%
ʃ	Fricative	1182263	1.0%
ʔ	Plosive	1172841	1.0%
v	Fricative	1109315	1.0%
ʧ	Affricate	877425	0.8%
ʤ	Affricate	826605	0.7%
z	Fricative	811555	0.7%
ɛː	Vowel	738223	0.6%
w	Approximant	692521	0.6%
g	Plosive	536608	0.5%
ɔː	Vowel	183376	0.2%
iː	Vowel	120028	0.1%
à	Accented Vowel	40801	0.0%
ʒ	Fricative	7390	0.0%
ò	Accented Vowel	5476	0.0%
è	Accented Vowel	1390	0.0%
ù	Accented Vowel	864	0.0%
dz	Affricate	546	0.0%
ì	Accented Vowel	82	0.0%



Bar-graph of phoneme frequency (expressed as a percentage) and ordered by increasing frequency.

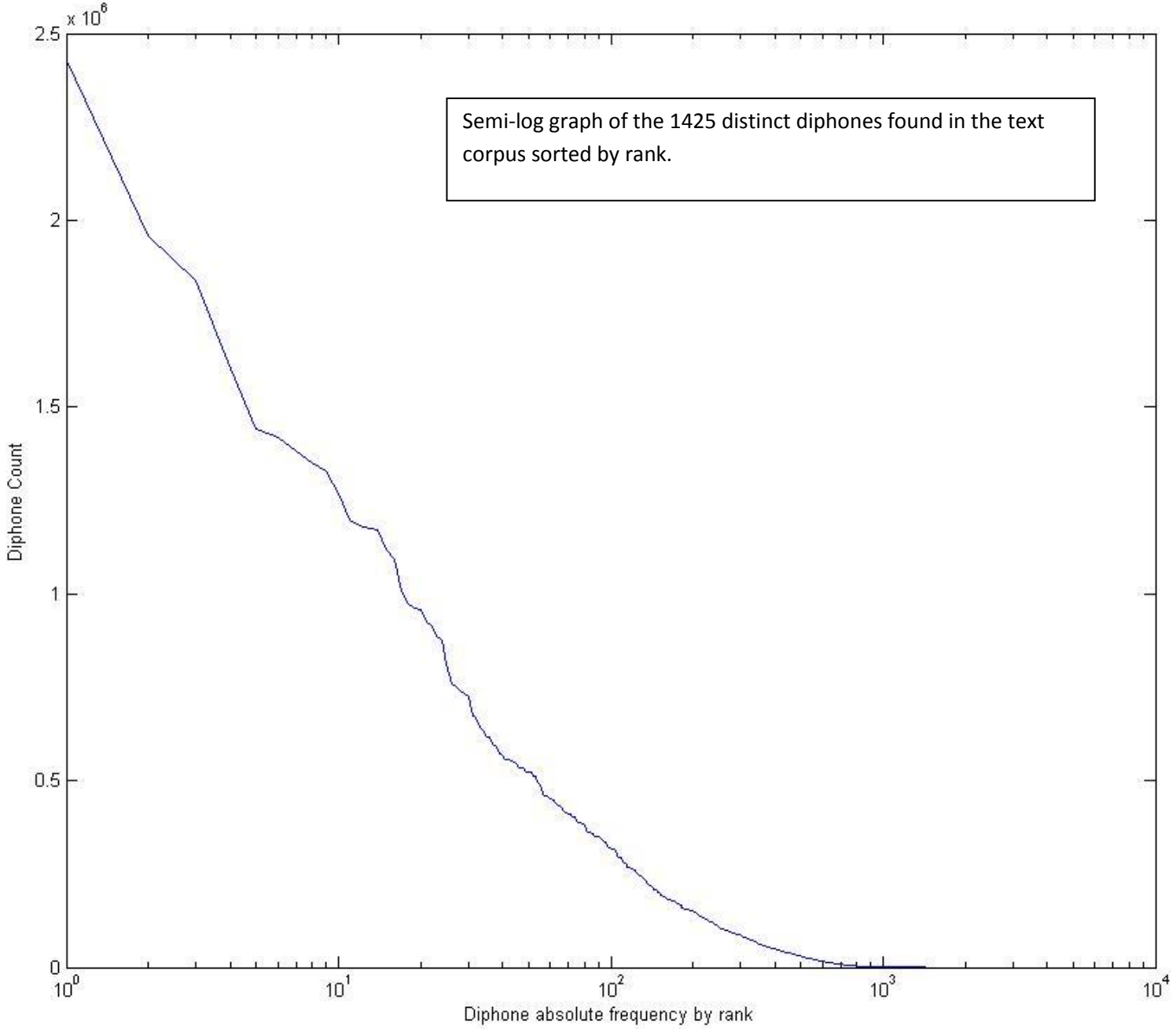
Diphone Statistics

Total diphone occurrences found in corpus: 113,860,068.

Maximum possible number of diphones (phoneme combinations): 43 x 43 = 1849 different diphones.

Actual number of distinct diphones found in the text corpus: 1425 (77% of all possible combinations). That is, 23% of possible diphones didn't occur even once in the around 113 million diphone occurrences found in the text corpus.

Note: Diphones consisting of *@ + <phoneme>*, and *<phoneme> + @* (where @ stands for silence), indicate the start and end of a sentence respectively.


x 10^6
2.5
2
1.5
1
0.5
0
Diphone Count
10^0
10^1
10^2
10^3
10^4
Diphone absolute frequency by rank
Semi-log graph of the 1425 distinct diphones found in the text corpus sorted by rank.

Of the 1425 distinct diphones found, the first 347 diphones account for 90% of all diphone occurrences.

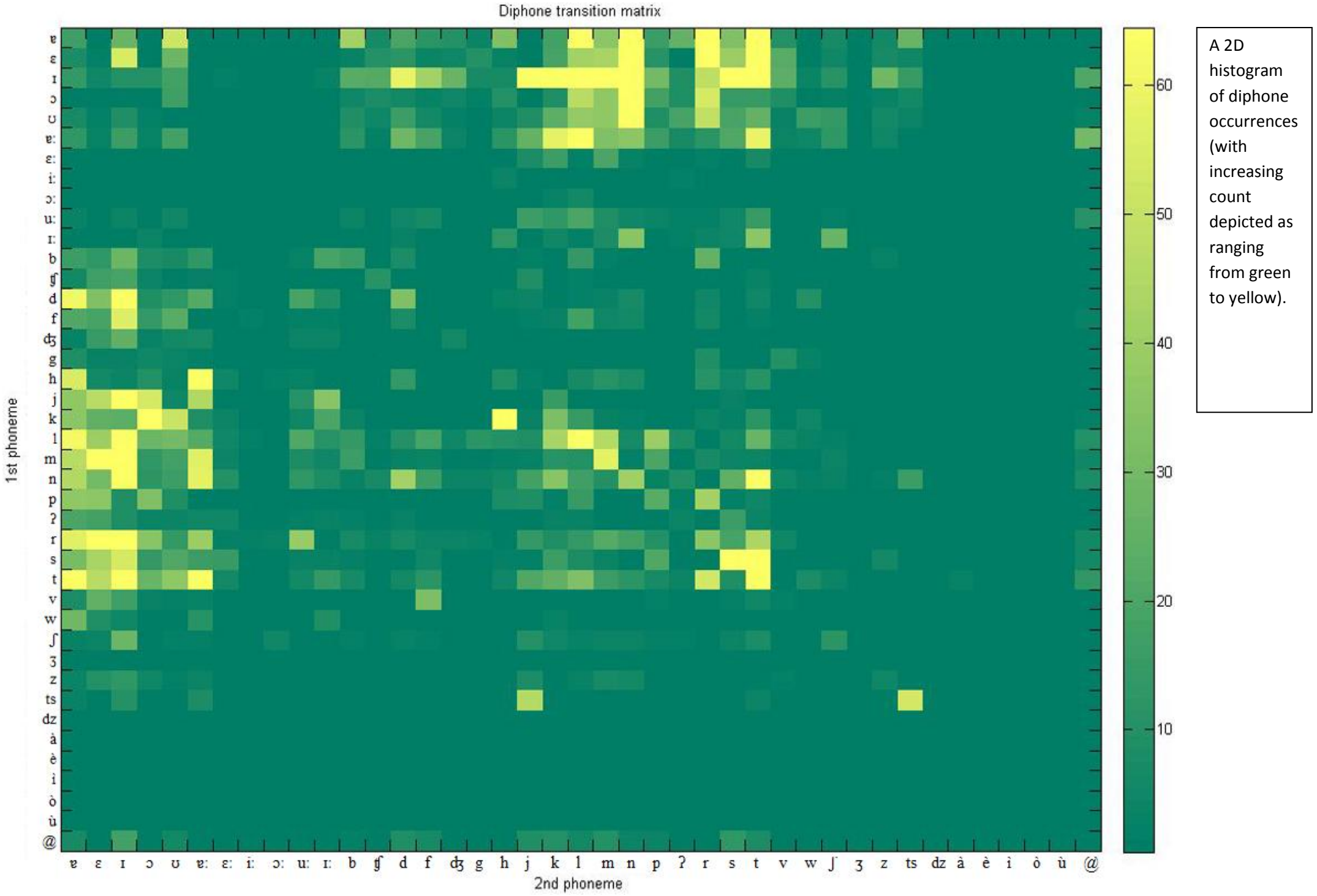
And the first 74 diphones account for 50% of all diphone occurrences; these are given in the table below:

diphone	phonetic class	count	%	cum. %
l + ɪ	Lateral Approximant + Vowel	2430652	2.13%	2.13%
ɪ + n	Vowel + Nasal	1957405	1.72%	3.85%
ɪ + l	Vowel + Lateral Approximant	1837473	1.61%	5.47%
n + ɪ	Nasal + Vowel	1602938	1.41%	6.88%
ɪ + s	Vowel + Fricative	1441802	1.27%	8.14%
t + ɪ	Plosive + Vowel	1419699	1.25%	9.39%
t + ɐː	Plosive + Vowel	1382066	1.21%	10.60%
t + ɐ	Plosive + Vowel	1350699	1.19%	11.79%
k + h	Plosive + Fricative	1329275	1.17%	12.96%
ɪ + t	Vowel + Plosive	1270346	1.12%	14.07%
l + l	Lateral Approximant + Lateral Approximant	1197676	1.05%	15.12%
ɐ + l	Vowel + Lateral Approximant	1182159	1.04%	16.16%
ɐ + r	Vowel + Retroflex	1174565	1.03%	17.19%
s + t	Fricative + Plosive	1171378	1.03%	18.22%
m + ɪ	Nasal + Vowel	1120498	0.98%	19.21%
ɐː + l	Vowel + Lateral Approximant	1091452	0.96%	20.17%
n + t	Nasal + Plosive	1007284	0.88%	21.05%
ɪ + j	Vowel + Approximant	971687	0.85%	21.90%
h + ɐː	Fricative + Vowel	962741	0.85%	22.75%
r + ɪ	Retroflex + Vowel	958064	0.84%	23.59%
t + t	Plosive + Plosive	924493	0.81%	24.40%
ɔ + n	Vowel + Nasal	911626	0.80%	25.20%
ɛ + n	Vowel + Nasal	884939	0.78%	25.98%
ɛ + r	Vowel + Retroflex	874991	0.77%	26.75%
ɪ + k	Vowel + Plosive	810539	0.71%	27.46%
ɐ + n	Vowel + Nasal	757979	0.67%	28.13%
j + ɪ	Approximant + Vowel	751180	0.66%	28.79%
ɐ + t	Vowel + Plosive	740833	0.65%	29.44%
ɛ + t	Vowel + Plosive	733403	0.64%	30.08%

ʊ + n	Vowel + Nasal	725820	0.64%	30.72%
ɪ + m	Vowel + Nasal	673707	0.59%	31.31%
m + ɛ	Nasal + Vowel	670032	0.59%	31.90%
r + ɛ	Retroflex + Vowel	641150	0.56%	32.46%
s + s	Fricative + Fricative	633704	0.56%	33.02%
d + ɪ	Plosive + Vowel	617072	0.54%	33.56%
k + ɔ	Plosive + Vowel	615926	0.54%	34.10%
l + ɐ	Lateral Approximant + Vowel	595747	0.52%	34.62%
d + ɐ	Plosive + Vowel	588958	0.52%	35.14%
ɐː + t	Vowel + Plosive	571046	0.50%	35.64%
ɪ + d	Vowel + Plosive	567137	0.50%	36.14%
ɐː + k	Vowel + Plosive	556854	0.49%	36.63%
n + ɐː	Nasal + Vowel	555128	0.49%	37.12%
m + m	Nasal + Nasal	554230	0.49%	37.60%
m + ɐː	Nasal + Vowel	550585	0.48%	38.09%
r + ɐ	Retroflex + Vowel	546667	0.48%	38.57%
f + ɪ	Fricative + Vowel	534366	0.47%	39.04%
h + ɐ	Fricative + Vowel	533633	0.47%	39.51%
ɛ + ɪ	Vowel + Vowel	528830	0.46%	39.97%
s + ɪ	Fricative + Vowel	523575	0.46%	40.43%
ɔ + r	Vowel + Retroflex	523423	0.46%	40.89%
j + ɔ	Approximant + Vowel	522581	0.46%	41.35%
t + r	Plosive + Retroflex	512718	0.45%	41.80%
ts + ts	Affricate + Affricate	509910	0.45%	42.25%
ɐ + ʊ	Vowel + Vowel	495873	0.44%	42.68%
k + ʊ	Plosive + Vowel	487288	0.43%	43.11%
ʊ + r	Vowel + Retroflex	467600	0.41%	43.52%
ɪ + r	Vowel + Retroflex	460601	0.40%	43.93%
j + ɛ	Approximant + Vowel	457334	0.40%	44.33%
m + ɐ	Nasal + Vowel	453731	0.40%	44.73%
ɔ + l	Vowel + Lateral Approximant	451481	0.40%	45.12%
t + ɛ	Plosive + Vowel	449541	0.39%	45.52%
l + m	Lateral Approximant + Nasal	447922	0.39%	45.91%
ts + j	Affricate + Approximant	441482	0.39%	46.30%
j + ɐː	Approximant + Vowel	433452	0.38%	46.68%
n + ɐ	Nasal + Vowel	433012	0.38%	47.06%

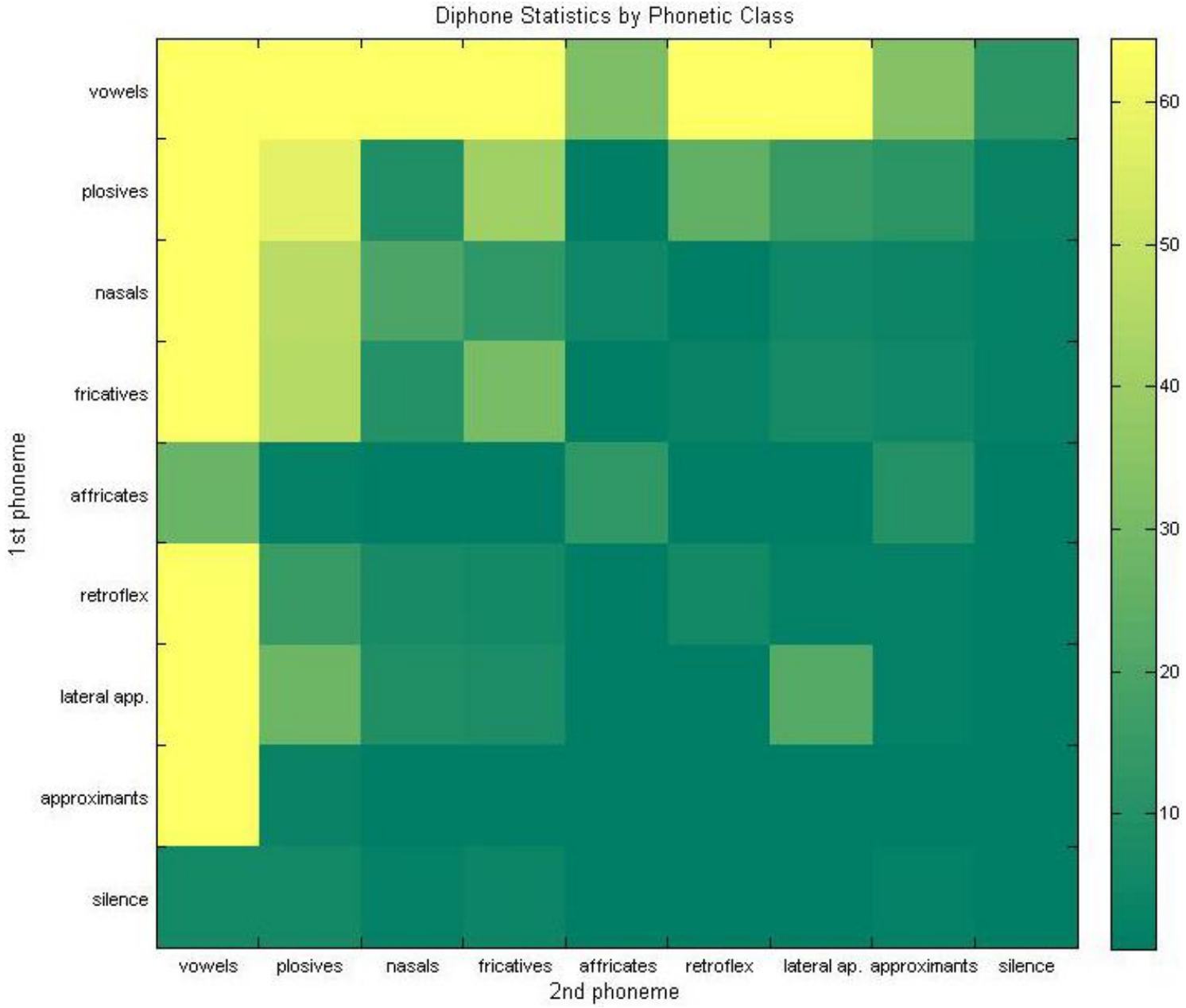
r + t	Retroflex + Plosive	427668	0.38%	47.44%
s + ɛ	Fricative + Vowel	418469	0.37%	47.80%
ɛ + m	Vowel + Nasal	414852	0.36%	48.17%
l + k	Lateral Approximant + Plosive	413222	0.36%	48.53%
ɛ + l	Vowel + Lateral Approximant	409769	0.36%	48.89%
n + d	Nasal + Plosive	409338	0.36%	49.25%
p + r	Plosive + Retroflex	404583	0.36%	49.60%
ɐ + b	Vowel + Plosive	402896	0.35%	49.96%
ɪ + f	Vowel + Fricative	402099	0.35%	50.31%

...rest of diphones not shown. The full list of diphones found in the text corpus is saved to the file "<OUTPUT_DIR>\diphone_list".



A 2D histogram of diphone occurrences (with increasing count depicted as ranging from green to yellow).

Same 2D histogram of diphone occurrences (as previous graph) but with diphones grouped by their respective phonetic class:



The previous graph in tabular form (showing frequency counts):

# of diphones	Vowel	Plosive	Nasal	Fricative	Affricate	Retroflex	Lateral-Approximant	Approximant	Silence	Total
Vowel	3827813	11331819	8860612	6754639	1691740	3713349	5647314	1840398	680396	44348080
Plosive	13468614	3075597	522769	2206049	41477	1370737	820457	684022	174536	22364258
Nasal	6778804	2506804	1073562	730422	284360	94268	269890	242519	134333	12114962
Fricative	6704707	2453477	567183	1652117	50537	191208	398745	295266	125486	12438726
Affricate	1459882	144119	25486	25767	701020	25845	29008	541838	9930	2962895
Retroflex	3485723	793484	388996	328048	100959	350066	136491	149312	58490	5791569
Lateral-Approxim.	4592729	1490977	519014	445407	78833	12061	1197676	121561	102013	8560271
Approximant	3687973	198905	12761	56224	4213	20084	1557	11782	312	3993811
Silence	341835	369076	144579	240053	9756	13951	59133	107113	0	1285496
Total	44348080	22364258	12114962	12438726	2962895	5791569	8560271	3993811	1285496	113860068

And the same graph with frequency expressed as a %:

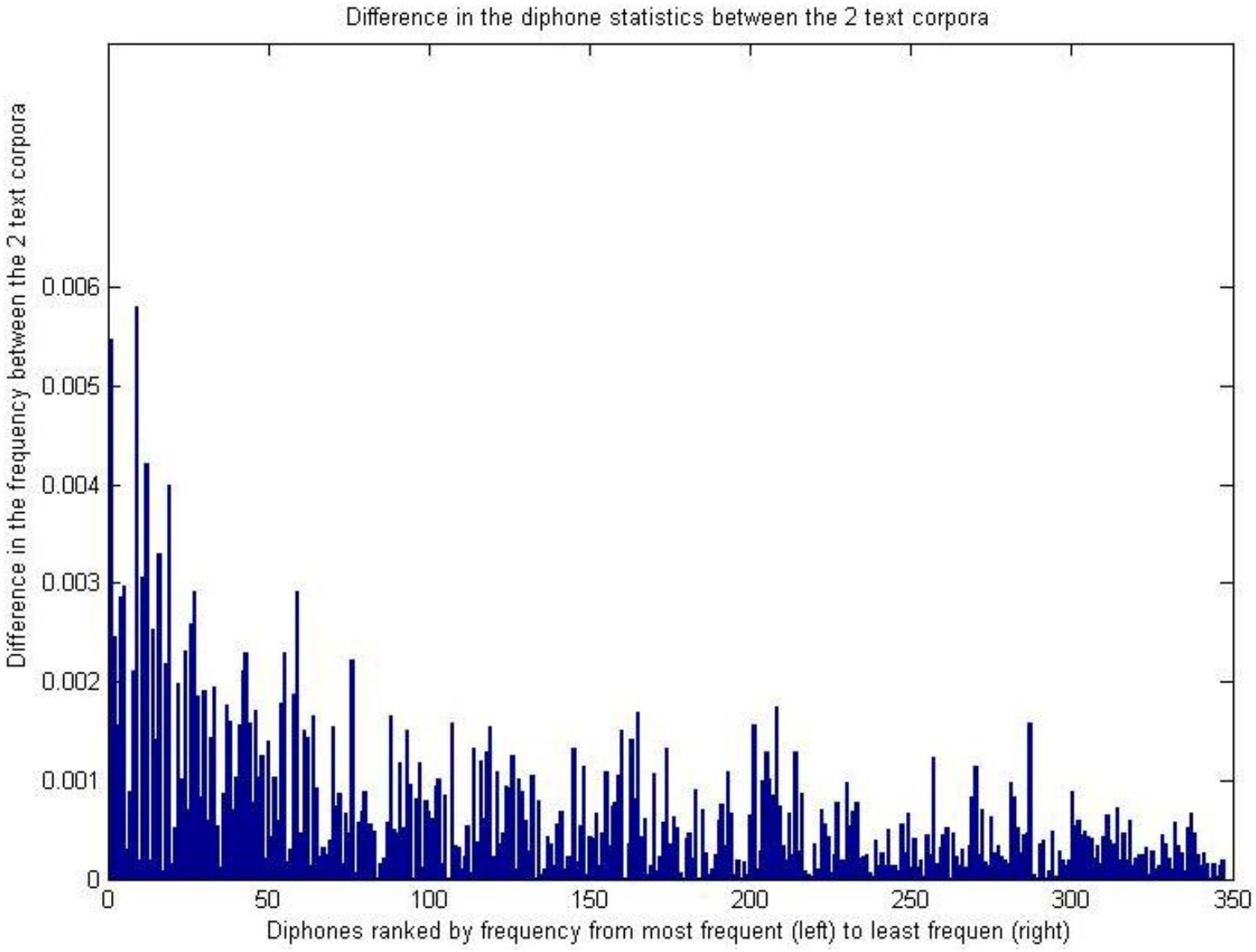
%	Vowel	Plosive	Nasal	Fricative	Affricate	Retroflex	Lateral-Approximant	Approximant	Silence	Total
Vowel	3.36%	9.95%	7.78%	5.93%	1.49%	3.26%	4.96%	1.62%	0.60%	38.95%
Plosive	11.83%	2.70%	0.46%	1.94%	0.04%	1.20%	0.72%	0.60%	0.15%	19.64%
Nasal	5.95%	2.20%	0.94%	0.64%	0.25%	0.08%	0.24%	0.21%	0.12%	10.64%
Fricative	5.89%	2.15%	0.50%	1.45%	0.04%	0.17%	0.35%	0.26%	0.11%	10.92%
Affricate	1.28%	0.13%	0.02%	0.02%	0.62%	0.02%	0.03%	0.48%	0.01%	2.60%
Retroflex	3.06%	0.70%	0.34%	0.29%	0.09%	0.31%	0.12%	0.13%	0.05%	5.09%
Lateral-Approxim.	4.03%	1.31%	0.46%	0.39%	0.07%	0.01%	1.05%	0.11%	0.09%	7.52%
Approximant	3.24%	0.17%	0.01%	0.05%	0.00%	0.02%	0.00%	0.01%	0.00%	3.51%
Silence	0.30%	0.32%	0.13%	0.21%	0.01%	0.01%	0.05%	0.09%	0.00%	1.13%
Total	38.95%	19.64%	10.64%	10.92%	2.60%	5.09%	7.52%	3.51%	1.13%	100.0%

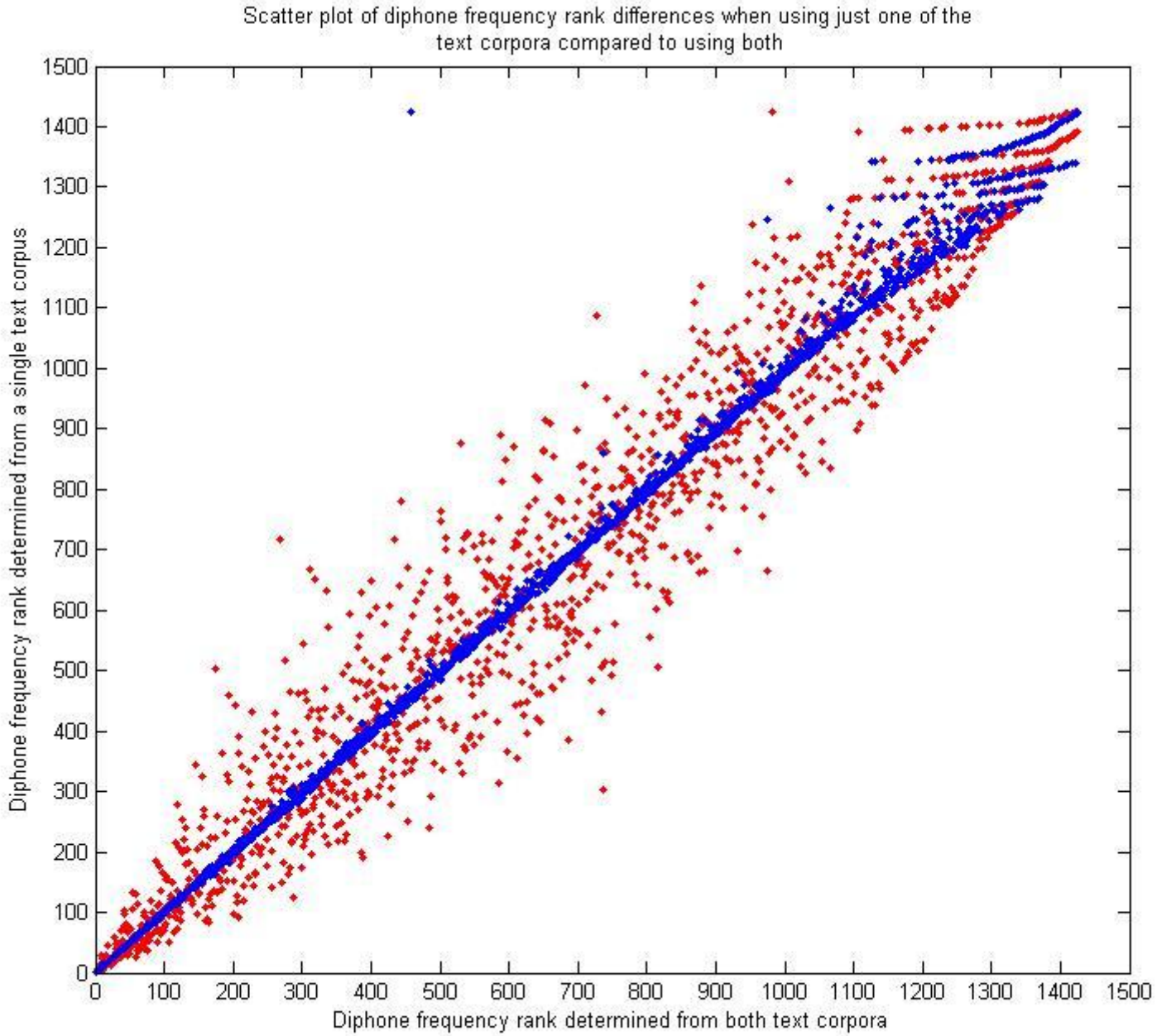
The diphone frequency grouped by consonant (C) and vowel (V) classes (e.g., C+V means diphones consisting of a consonant followed by a vowel):

	# of diphones	%
V+V	3827813	3.36%
V+C	39839871	34.99%
C+V	40178432	35.29%
C+C	27442960	24.10%

Diphone Statistics Variation between Text Corpora

A simple analysis was performed to get an idea of how much diphone statistics can vary between different text corpora. In this case, diphone statistics were individually calculated for each of the 2 text corpora used here and normalised (to counter for the difference in sizes of the 2 text corpora). Then the differences between the diphone statistics were calculated. The following graph shows the differences for the first 347 diphones (i.e., covering 90% of all diphone occurrences).





Red = text corpus 1

Blue = text corpus 2