Diphone Analysis of the Macedonian Language for the Purpose of Text-to-Speech Synthesis

Branislav Gerazov[1] and Zoran Ivanovski[2]

Abstract – **The paper presents the results obtained from a thorough phonetic analysis of the Macedonian language for the purpose of diphone-based Text-to-Speech (TTS) synthesis. A Body of Text containing more than 2300000 words was compiled and analyzed. From it a set of 707 unique diphones was extracted, and is presented in this paper. The set is a necessary starting point for the development of diphone-based TTS for Macedonian. It reduces the man-hour cost involved in building the TTS system's unit inventory by 40%. This is the first paper to present such a thorough analysis of Macedonian diphone characteristics, and is of importance not only for TTS development in Macedonia, but also on the international level for multi-lingual TTS systems.**

Keywords – **Text-to-Speech, diphone, phonetics, Macedonian.**

I. Introduction

There are three paradigms of Text-to-Speech (TTS) synthesis: articulatory, formant and concatenative synthesis. Up until the 1990's, TTS synthesizers predominately used articulatory and formant synthesis. Both of these are based on a model of the speech production process. Because of the cost involved in building the models, TTS synthesizers based on these two paradigms were restricted only to the world's most wide-spoken languages, i.e. English, French, German, Chinese etc.

The third paradigm, concatenative synthesis, abandons modeling speech production altogether. It uses a data base of prerecorded segments of natural speech, which it concatenates one after the other, to generate the requested speech output. Concatenation of prerecorded (natural) segments, also called units, gives the synthetic speech a very natural sound. Various unit lengths are used in concatenative synthesis, the longer the units the more natural the output speech, but the bigger the database. For example a given language may have only 30 – 40 phones, which are the shortest phonetic (acoustic) units of a language; but more than 200000 words. Most systems rely on a compromise between the two – the diphone, i.e. two half-phones including the transition between them. Diphones usually number around 1000 and provide for reasonable speech quality. The most popular diphone based systems are AT&T's diphone based TTS system and Dutoit's MBROLA synthesizer [1, 2].

Owing to the simplicity of concatenative synthesis and the widespread availability of computing power, in 1990's TTS systems began to be developed in languages that lacked world domination. In the Slavic speaking part of the Balkans various systems have appeared, most of them based on diphones. The most prominent of these is the "AlphaNum" unit-selection TTS system for Serbian and later Croatian, [3]. Other systems include "SpeechLab" for Bulgarian and "Govorec" and "Proteus" for Slovenian.

Several attempts have been made to synthesize speech in Macedonian. The first attempts date back to 1997, when a concept solution was found, but never became fully operational, [4]. Later attempts included emulating Macedonian using the Croatian diphone inventory under the MBROLA framework, [5], with a similar approach by the AlfaNum team, this time using a Serbian diphone inventory, [6]. Currently, two high quality systems are under development: TTS-MK, a diphone based system in development at FON University, [7], and "Speak Macedonian", our system, which is based on modified diphone units, [8].

The set of diphones found in a language is an essential starting point for developing a diphone-based TTS system for that particular language. This is highly specific to the language at hand, and calls for research of its phonetic characteristics. Although TTS systems in Macedonian have been in development for some time, results of this kind have not been published yet. This paper presents the results obtained from a thorough phonetic analysis of a large body of text in Macedonian, totaling more than 2300000 words. From the analysis data, the diphone set needed for TTS in Macedonian was extracted and statistically processed. We have found 707 unique diphones which form a practically complete basis for a diphone based concatenative synthesis TTS system. The results presented herein can be of value not only for TTS development efforts in Macedonia, but also for international multi-lingual TTS system development.

II. Diphones

Although the phone is the basic acoustic unit of speech, synthesizing speech by concatenating phones yields bad results. The quality of the synthesized speech is poor, with mediocre intelligibility and no naturalness, [9]. This is because of the great difficulty in simulating the gradual change of acoustic information from one phone to the next, called an interphone transition. Because of this systems rarely rely on phone based units.

The more common approach is to use a unit that includes half of each phone and the transition between them, called a diphone. This way the concatenation point is no longer the transition interval, but rather the middle, somewhat, steady-state of the phone.

Theoretically the number of diphones is given by the number of phones squared. However, due to phonotactic constraints, i.e. restrictions on phone pairing in language, this is not the case. For example, English does not endorse an "sh-t" transition, in IPA (International Phonetic Alphabet) terms /ʃ/-/t/, which is common in Macedonian and other Slavic languages. The actual number of

[1]Branislav Gerazov is with the Faculty of Electrical Engineering and Information Technology at University of Ss. Cyril and Methodius, Skopje, Macedonia, E-mail: gerazov@feit.ukim.edu.mk

[2]Zoran Ivanovski is with the Faculty of Electrical Engineering and Information Technology at University of Ss. Cyril and Methodius, Skopje, Macedonia, E-mail: mars@feit.ukim.edu.mk

diphones varies between languages and is usually far from this theoretical value. For example English has 43 phones and 1162 diphones, as used by AT&T's TTS system. This is far from the theoretical 1849 possible diphone combinations, [1].

III. MACEDONIAN PHONES

Macedonian is comprised of 33 basic phones, 28 of which are represented a unique letter in the alphabet. The letters together with their IPA (International Phonetic Alphabet) equivalents are presented in Table I. Of these, five are vowels (/a/, /ɛ/, /i/, /ɔ/ and /u/), and the rest are consonants. Grouped according to the manner of articulation the consonants are comprised of: 6 plosives (/b/, /p/, /g/, /k/, /d/, /t/), 3 approximants (/j/, /r/), 7 fricatives (/v/, /f/, /z/, /s/, /ʒ/, /ʃ/, /h/), 6 affricates (/dʒ/, /tʃ/, /dz/, /ts/, /ɟ/, /c/) and 3 nasals (/m/, /n/, /ɲ/) [10]. The letter "л" reads /ɬ/ but also /l/ when preceding the front vowels /ɛ/ and /i/, and the approximant /j/. The letter "љ" reads /lj/. The cluster /lj/ was treated as a phone, due to its compactness and the tendency to palatize it in common speech. The phone "r" can become syllabic /r̩ / when enclosed by consonants, as it is in: "првиот" /pr̩viɔt/ (eng. the first), " 'рбет" /r̩bɛt/ (eng. spine) etc. Finally "n" is velarized to /ŋ/ before /k, g/, as in "банка" /baŋka/ or "англиски" /aŋgliski/, [11]. The phones /l/, /r̩ / and /ŋ/ were coded with "q", "w" and "n" in the text, in the phonetic processing step.

TABLE I
PHONE INVENTORY OF MACEDONIAN WITH CORRESPONDING IPA EQUIVALENTS

А	/ a /	И	/ i /	С	/ s /
Б	/ b /	Ј	/ j /	Т	/ t /
В	/ v /	К	/ k /	Ќ	/ c /
Г	/ g /	Л	/ ɬ /, / l /	У	/ u /
Д	/ d /	Љ	/ lj /	Ф	/ f /
Ѓ	/ ɟ /	М	/ m /	Х	/ x/
Е	/ ɛ /	Н	/ n /, / ŋ /	Ц	/ ʦ /
Ж	/ ʒ /	Њ	/ ɲ /	Ч	/ ʧ /
З	/ z /	О	/ ɔ /	Џ	/ ʤ /
S	/ ʣ/	П	/ p /	Ш	/ ʃ /
		Р	/ r /, / r̩ /		

IV. MACEDONIAN DIPHONE ANALYSIS

With a phone count of 31 + /r̩ / + / l / + / ŋ / + /silence/ = 35, the theoretical maximum of diphones for Macedonian is 1225. Again due to Macedonian phonotactics the diphone inventory size is significantly less that the theoretical maximum. To find the exact number of diphones in Macedonian, a large body of text was assembled from various sources. This text was phonetically processed and analyzed in two main process lines shown in Fig. 1. They will be further elaborated in the paper.

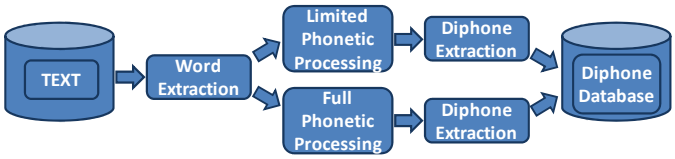

Fig. 1. Two processing lines for Macedonian diphone set extraction

A. Body of Text

The body of text analyzed in our work has a total of more than 2300000 words. It consists of three major parts. The first part, termed the *Reference Body*, was compiled from works of classic and contemporary literature in Macedonian, that were readily accessible on-line. Special care was taken to use a representative cross section of both standard and contemporary trends in the Macedonian language. In addition the Macedonian Constitution as well as various acts of law were included.

The second part of the body was assembled from the Old and New Testament. It's word-count is over 800000. This text was processed separately because the Bible contains names of archaic places and people not in general use today. These distinctive words were later necessarily excluded from the diphone analysis.

The third part of the body comprises a 260000 word Macedonian Dictionary made available by *OpenOffice.org*. Due to the lack of valid word-counts in it, a relevant statistical distribution of diphones could not be extracted. However, the Dictionary was invaluable due to its wide coverage of words in Macedonian. This kind of coverage would not be possible even by making the body of text become arbitrarily large.

B. Word Extraction

The starting point of the diphone analysis process was word extraction from the bodies of text. In total, from the approximately 1300000-word Reference Body, close to 80000 unique words were extracted. Of these, about 44000 occurred only once in the entire body of text. The 404 most common words accounted for 50% of the total word count. The analysis of the Bible text yielded around 40000 unique words, of which 17062 appeared only once, and 94 most common accounted for 50% of words in the text. This data is summarized in Table II.

Here we can see why the Dictionary was invaluable to our analysis. Of about 2100000 words from the Reference Body and the Bible, only 120000 are unique and 55000 of these appeared only once in the whole text. This means that no matter how large the assembled body of text is, it can never cover all the words in a given language. The inclusion of the Dictionary, with its wide coverage of words, reduces this problem.

Here we can see why the Dictionary was invaluable to our analysis. Of about 2100000 words from the Reference Body and the Bible, only 120000 are unique and 55000 of these appeared only once in the whole text. This means that no matter how large the assembled body of text is, it can never cover all the words in a given language. The inclusion of the Dictionary, with its wide coverage of words, reduces this problem.

Table III gives the 10 most common words in the Reference Body and the Bible. It can be seen that the two line up well, i.e. 8 of the words are the same, with the first 5 being almost

equal. This data can be useful should the diphone database be expanded with whole words.

TABLE II
WORD STATISTICS FOR THE THREE PARTS OF THE BODY OF TEXT USED IN THE DIPHONE ANALYSIS

	Word count	Unique	Appeared once	Acc. for 50%	Acc. for 90%
Reference Body	1287513	79065	38593	404	21972
Bible	813878	38572	17062	94	5795
Dictionary	261460	n/a	n/a	n/a	n/a

TABLE III
10 MOST COMMON WORDS IN THE REFERENCE BODY AND THE BIBLE

	Reference Body			Bible		
	Word	Occurences	%	Word	Occurences	%
1	на	59953	4,66	и	45340	5,57
2	и	44937	3,49	на	29195	3,59
3	се	33420	2,60	се	18081	2,22
4	да	33020	2,56	да	17127	2,10
5	во	26959	2,09	во	16306	2,00
6	од	24400	1,90	го	15902	1,95
7	за	21733	1,69	од	15264	1,88
8	со	19494	1,51	ќе	14437	1,77
9	го	17508	1,36	за	11188	1,37
10	не	17496	1,36	што	10957	1,35

A statistical curve, in log-log scale, of the distribution of word appearance against word rank in the frequency distribution table is given in Fig. 2. From the linearity of the plot, we can see that it clearly abides to Zipf's power law, i.e. the frequency of a word is inversely proportional to its rank in the occurrence table, [12].

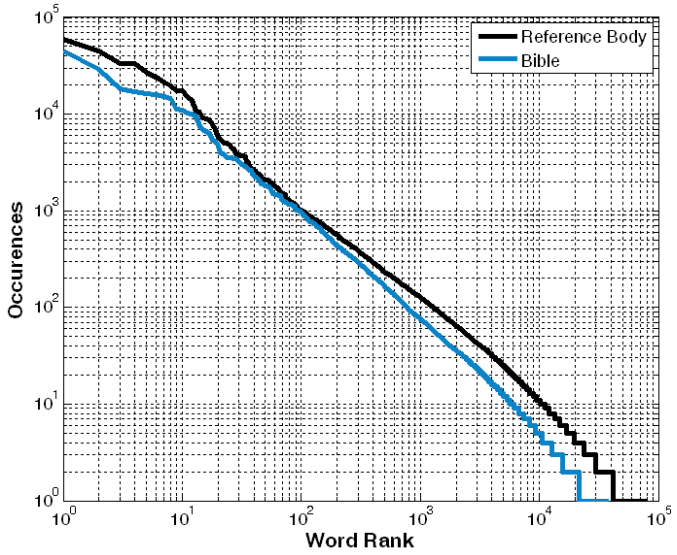

Fig. 2. Word distribution plotted in log-log scale in the Reference Body and the Bible

C. Phonetic Processing

The phonetic processing step compensates for the difference in the rules of orthography (standard writing) vs. those of orthoepy (standard speaking). It was done on two levels: a *limited* ("light") and a *expanded phonetic processing level*.

- *limited phonetic processing level* – singles out the appearance of / r̩ /, / l / and / ŋ /, and codes (substitutes) them with "q", "w" and "n", respectively. The cluster /lj/ represented by "љ", was left intact.
- *expanded phonetic processing level* – builds on the previous level and additionally implements an expanded set of Macedonian orthoepy rules, comprising: voicing assimilation in consonant pairs, reduction of consonants due to assimilation, and devoicing of voiced consonants at the end of words, [10]. This is expected to decrease the total diphone count.

D. Diphone Extraction

Diphone extraction is then carried out on the words that have been phonetically processed. All diphones are extracted and a diphone data base is compiled together with their number of occurrences.

V. RESULTS

This section summarizes the results obtained from the diphone analysis of Macedonian. The total number of diphones from the Reference Body, the Bible, and the Dictionary are given in Table IV. Due to the uniqueness of the Bible texts, two statistics were drawn out, one from the complete word-count, and another from a reduced word-count that excluded the diphones unique to biblical places/people.

Three of the texts: the Reference Body, the Bible Reduced and the Dictionary, were used to generate a single consolidated list of diphones. This *Master List* gives the total count of 728 unique diphones found in the Macedonian language. This is 59,4% of the theoretical maximum of 1225, for the 35 phones used. Its statistics are also given in Table IV. The second processing line, which implemented expanded phonetic processing, yielded a smaller orthoepy based diphone set totaling 707 diphones, or 57.7% of the theoretical maximum. This set is invaluable to the task of building a unit inventory for a diphone-based TTS system for Macedonian, because it provides for a considerable 40% decrease of the man-hour cost involved in the process!

TABLE IV
DIPHONE STATISTICS FOR THE THREE PARTS OF THE ANALYZED TEXT

	Diphone count	Unique	Appeared in a single word	Acc. for 50%	Acc. for 90%
Reference Body	7508399	708	16	41	192
Bible	4407721	680	41	39	170
Bible Reduced	4402611	622	23	39	170
Dictionary	2665939	705	16	47	213
Master List	14582781	728	6	43	195

It should be mentioned that there is a possibility that words in the Macedonian language have not appeared at all in the Body of Text used for the analysis. There is a chance some of these words to contain a diphone not found in this analysis. However, the authors consider this chance as being minute.

Table V shows the 10 most common diphones in the three texts. It can be seen that there is large correlation between the Reference Body and the Reduced Bible texts. The Dictionary most common diphone set differs, because this text contains every word only once (even the most common), while the first two contain natural language word distributions. Even so, a small degree of correlation can still be found.

TABLE V
MOST COMMON DIPHONES IN THE ANALYZED TEXTS

	Reference Body			Bible Reduced			Dictionary		
	Diph.	Occ.	%	Diph.	Occ.	%	Diph.	Occ.	%
1	а#	358709	4,78	а#	195037	4,43	а#	67485	2,53
2	е#	221411	2,95	е#	148312	3,37	ва	67411	2,53
3	о#	197631	2,63	и#	136874	3,11	е#	58687	2,20
4	и#	185966	2,48	о#	110728	2,52	на	51552	1,93
5	#с	154224	2,05	#с	93976	2,13	#п	48735	1,83
6	#н	152863	2,04	#н	91613	2,08	ан	46544	1,75
7	на	152414	2,03	на	73459	1,67	ув	45749	1,72
8	#п	108475	1,44	#и	72787	1,65	ра	41678	1,56
9	#д	97997	1,31	#п	61978	1,41	ни	41006	1,54
10	то	89878	1,20	то	54317	1,23	о#	38590	1,45

Fig. 3 gives the log-log diphone distribution against diphone rank for the three texts, and the consolidated Master List. It can be seen they only partially follow Zipf's law with a sharp roll-off near the end.

The diphone set extracted is given in matrix form in Table VI. Phones are given in Cyrillic as well as IPA characters. The rows represent the first phone in the diphone, while the columns represent the second. Diphones that appeared in both the texts with limited and expanded phonetic processing were marked with an "x". Diphones unique for the text with limited phonetic processing were marked with a "g" (orthography), and the ones unique to the text with expanded phonetic processing with an "e" (orthoepy). The total number of diphones is 732, 728 of these abide orthography and 707 orthoepy.

VI. CONCLUSION

We have presented the results of a phonetic analysis of the structure of the Macedonian language in the respect of diphones. These are the first results from phonetic diphone analysis of Macedonian, for the purpose of TTS synthesis.

The results are of great value for the development of TTS systems in this language, whether in Macedonia or on an international level. The set of 707 diphones acquired through this analysis is a necessary starting point for building unit databases for diphone-based TTS systems in Macedonian. Using the set cuts the man-hour cost of this process down by 40 %.

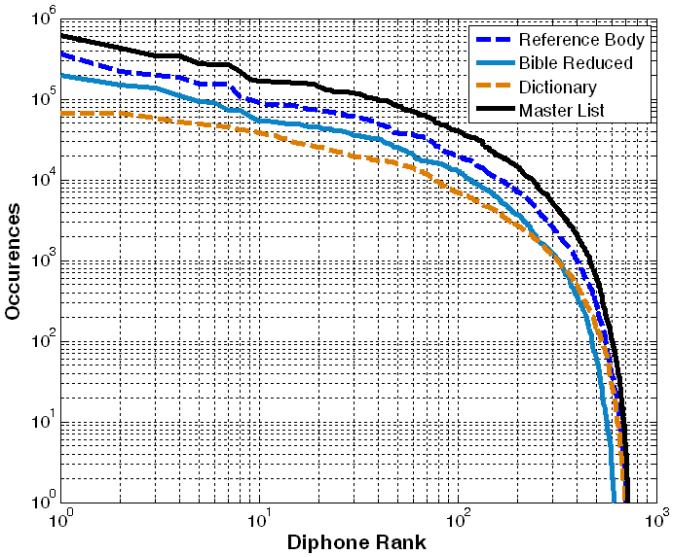

Fig. 3. Log-log scale graph of diphone occurrences vs. rank in the distribution table in the three analyzed texts

REFERENCES

[1] J. Olive, J. van Santen, B. Möbius and C. Shih. "Synthesis". In R. Sproat (Ed.), *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, pp. 191-228, Kluwer, 1998.

[2] T. Dutoit, V. Pagel, N. Pierret, F. Bataille and O. Vrecken, "The MBROLA Project: Towards a Set of High Quality Speech Synthesizers Free of Use for Non-Commercial Purposes", Proceedings of ICSLP 96, 1996.

[3] M. Sečujski, R. Obradović, D. Pekar, Lj. Jovanov and V. Delić, "AlfaNum System for Speech Synthesis in Serbian Language", In Proc. of the 5th Conf. Text, Speech and Dialogue, Brno, 2002.

[4] L. Josifovski, D. Mihajlov and D. Gorgevik, "Speech Synthesizer Based on Time Domain Syllable Concatenation", SPECOM '97 Cluj-Napoca, Oct. 27-30, 1997.

[5] М. Зрмановска, *Синтеза на македонски говор врз база на текст со прозодија*, Магистерска Теза, Електротехнички Факултет, Скопје, 2005.

[6] V. Delić, M. Sečujski, D. Pekar, N. Jakovljević and D. Mishković, "A Review of AlfaNum Speech Technologies for Serbian, Croatian and Macedonian", IS-LTC 06, Ljubljana, Slovenia, 9. - 10 october, 2006

[7] S. Chungurski, I. Kraljevski, D. Mihajlov and S. Arsenovski, "Concatenative Speech Synthesizers and Speech Corpus for Macedonian Language", 30th International Conference ITI Cavtat/Dubrovnik, Croatia, Jun. 23-26, 2008.

[8] B. Gerazov, G. Shutinoski and G. Arsov, "A Novel Quasi-Diphone Inventory Approach to Text-To-Speech Synthesis", MELECON '08, Ajaccio, France, May 5-7, 2008

[9] Б. Геразов и Г. Шутиноски, "Еден Пристап кон Преобразба на Текст во Говор на Македонски", ЕТАИ 2007, Охрид, Македонија, 19 – 21 сеп., 2007.

[10] С. Бојковска, Л. Минова-Ѓуркова, Д. Пандев и Ж. Цветковски, *Опита Граматика на Македонскиот Јазик*, Просветно Дело, 2008.

[11] V. Friedman, "Macedonian", SEELRC, 2001.

[12] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, p. 23, MIT Press, 1999.

TABLE VI
DIPHONE MATRIX FOR MACEDONIAN
DIPHONES FOUND IN: BOTH ORTHOGRAPHY AND ORTHOEPY (x) ; ORTHOGRAPHY ONLY (g) ; ORTHOEPY ONLY (e)

Cyrillic		а	б	в	г	д	ѓ	е	ж	з	ѕ	и	ј	к	л	љ	м	н	њ	о	п	р	с	т	ќ	у	ф	х	ц	ч	џ	ш	#	q	w	n
	IPA	a	b	v	g	d	ɟ	ɛ	ʒ	z	ʣ	i	j	k	ɫ	lj	m	n	ɲ	ɔ	p	r	s	t	c	u	f	x	ʦ	ʧ	ʤ	ʃ	#	r,	l	ŋ
а	a	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x
б	b	x		x	x	x		x		x	x	x	x		x	x	x	x		x		x		g		x				g			g	x	x	
в	v	x	x	x	x	x		x	x	x	x	x	x	g	x	x	x	x		x	g	x	g	g		x			g	g	x	g	g	x	x	
г	g	x	x	x		x		x		x		x			x		x	x	x	x	g	x	g	x		x				g		g	g	x	x	
д	d	x	x	x	x	x		x	x	x	x	x	x		x	x	x	x	x	x	g	x	g	g		x		g	g		x		g	x	x	
ѓ	ɟ	x						x				x						x		x						x							g			
е	ɛ	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x
ж	ʒ	x	x	x	x	x		x				x	x		x		x	x		x		x		x		x							g	x	x	
з	z	x	x	x	x	x		x	x	x	x	x	x		x	x	x	x		x		x		x		x					x		g	x	x	
ѕ	ʣ	x		x				x				x										x				x								x		
и	i	x	x	x	x	x	x	x	x	x		x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x
ј	j	x	x	x	x	x		x	x	x		x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	
к	k	x		x				x				x			x		x	x	x	x	e	x	x	x		x			x	x		x	x	x	x	
л	ɫ	x	x	x	x	x			x	x				x			x	x	x	x	x		x	x		x	x	x	x	x	x	x	x			x
љ	lj	x										x								x				x		x											
м	m	x	x	x	x	x		x				x	x	x	x		x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	
н	n	x	x	x		x	x	x	x	x	x	x			x					x	x	x	x	x		x	x	x	x	x	x	x	x		x	
њ	ɲ	x						x				x								x			x			x				x				x			
о	ɔ	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x
п	p	x						x				x	x	x	x	x		x		x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	
р	r	x	x	x	x	x	x	x	x	x		x	x	x	x		x	x		x	x		x	x		x	x	x	x	x	x	x	x		x	
с	s	x		x		x		x				x	x	x	x		x	x		x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	
т	t	x		x	x	x		x				x	x	x	x	x	x	x		x	x	x	x	x		x	x	x	x	x		x	x	x	x	
ќ	c	x		x				x				x		x				x		x				x		x								x			
у	u	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x
ф	f	x						x				x		x	x			x		x	e	x	x	x		x			x	x		e	x	x	x	
х	x	x		x				x				x			x		x	x		x		x		x		x		x		x			x	x		
ц	ʦ	x		x				x				x		x	x		x	x		x		x				x							x	x	x	
ч	ʧ	x	x	x				x				x	x	x	x		x	x	x	x	x			x		x							x	x	x	
џ	ʤ	x		x	x			x				x					x			x						x							g			
ш	ʃ	x		x				x				x	x	x	x		x	x	x	x	x	x		x		x			x	x		x	x		x	
#	#	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	
q	r,		x	x	x	x	x		x	x				x	x	x	x	x	x		x	x	x	x			e	x	x	x	x	x			x	x
w	l							x				x	x																							
n	ŋ				x									x																						