Recent Advances in Intelligent Information Systems
ISBN 978-83-60434-59-8, pages 159–172

Phonetic Statistics from an Internet Articles Corpus of Polish Language

Bartosz Ziółko, Jakub Gałka, and Mariusz Ziółko

Department of Electronics
AGH University of Science and Technology
Kraków, Poland
{bziolko, jgalka, ziolko}@agh.edu.pl

Abstract

The statistics of Polish phonemes, biphones and triphones were collected from a large Internet articles corpus. The paper presents summarisation of the data and some phenomena in the statistics including a distribution of frequency of triphones occurring. Triphone statistics play an important role in automatic speech recognition systems. They are used to apply context-dependent speech units. The phonetic alphabet for Polish, SAMPA, and methods of providing phonetic transcriptions are described.

Keywords: phoneme statistics, triphone statistics, Polish

1 Computional Linguistic Research on Phonemes

Statistical research at the word and sentence level are popular for several languages (Agirre *et al.*, 2001; Bellegarda, 2000). Any similar research on phonemes is rare Denes (1962); Yannakoudakis and Hutton (1992); Kollmeier and Wesselkamp (1997). The frequency of phonetic units presence in speech is an interesting topic itself. It can also be used in several applications in speech processing, for example automatic speech recognition. It is very difficult to provide proper acoustic data for all possible triphones to represent them with audio parameters. There are methods to prepare models of triphones which did not appear in a training corpus of a speech recogniser. Data of other similar triphones and phonological similarities between different phonemes can be used (Young *et al.*, 2005). It means, that the list of possible triphones has to be provided for a particular language. The triphone statistics can be also used to generate hypotheses used in recognition of out-of-dictionary words like names.

We have already presented some similar statistics (Ziółko *et al.*, 2007), which were collected from around 10,000,000 words of mainly spoken language. Here we present statistical data collected from 94,000,000 words from a corpus containing Internet articles of encyclopedia type, reviewed by a human supervisor. As the result, we can expect many names and rare words. Experiments on different corpora will allow to compare these statistics to evaluate how representative and complete

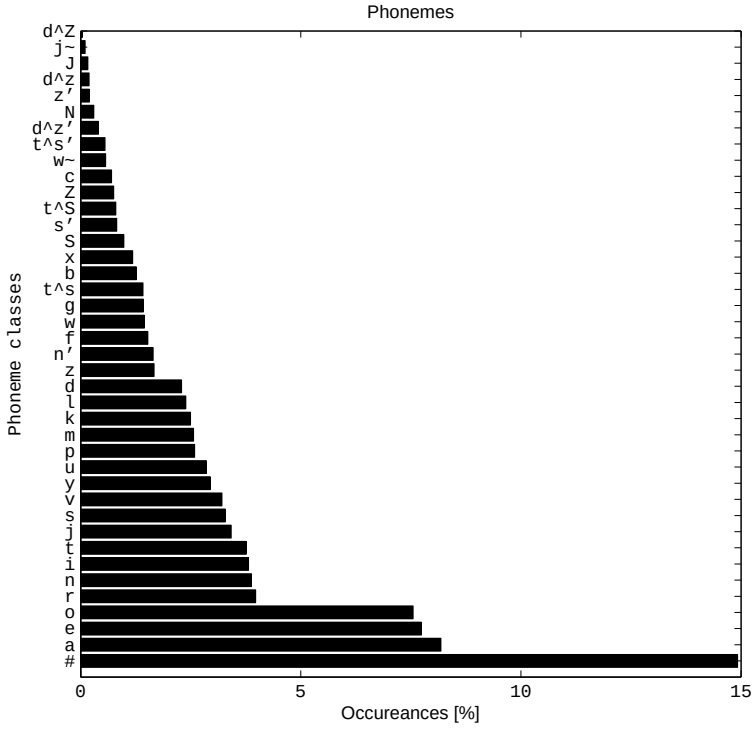

FIGURE 1: Phonemes in Polish in SAMPA alphabet

they are. We conduct similar experiments on literature and journal papers corpora, which were not published yet. We have found in research on semantics (Ziółko *et al.*, 2008) that often it is better to use formal written language corpus, rather then speech transcriptions for training. Even though, it is counter-intuitive, the written language has better quality, so it is a better base for linguistic rules. Speech transcriptions are quite random what can defect representing linguistic phenomena.

This paper describes several issues related to phoneme, biphone and triphone statistics and is divided as follows. Section 2 provides information about general scheme of our data acquisition method and standards we used. Section 3 describes the technically most difficult step which is changing the text corpus into a phonetic transcription. Section 4 contains a description of data we used and our results. Phenomena we uncovered are described as well. Section 5 presents opportunities of applying statistics we collected in natural language and speech processing for artificial intelligence tasks like automatic speech recognition. We sum up the paper with conclusions.

TABLE 1: Phoneme transcriptions in Polish — SAMPA (Demenko *et al.*, 2003)

SAMPA	example	transcriptions	occurrences	%
#		#	104,910,909	14.92
a	pat	pat	57,532,609	8.18
e	test	test	54,413,762	7.74
o	pot	pot	53,090,491	7.55
r	ryk	rIk	27,950,385	3.98
n	nasz	naS	27,258,894	3.88
i	PIT	pit	26,810,706	3.81
t	test	test	26,437,704	3.76
j	jak	jak	24,031,728	3.42
s	syk	sIk	23,087,313	3.28
v	wilk	vilk	22,559,385	3.21
I	typ	tIp	20,687,947	2.94
u	puk	puk	20,055,658	2.85
p	pik	pik	18,204,293	2.59
m	mysz	mIS	18,032,655	2.56
k	kit	kit	17,537,476	2.49
l	luk	luk	16,786,430	2.39
d	dym	dIm	16,124,146	2.29
z	zbir	zbir	11,714,062	1.67
n’	koń	kon’	11,588,014	1.65
f	fan	fan	10,739,351	1.53
w	łyk	wIk	10,160,908	1.44
g	gen	gen	9,997,347	1.42
ts	cyk	tsIk	9,963,254	1.42
b	bit	bit	8,878,611	1.26
x	hymn	xImn	8,314,327	1.18
S	szyk	SIk	6,910,431	0.98
s’	świt	s‘vit	5,751,440	0.82
tS	czyn	tSIn	5,606,551	0.80
Z	żyto	ZIto	5,272,166	0.75
c	kiedy	cjedy	4,924,129	0.70
w∼	ciąża	ts’ow∼Za	3,989,996	0.57

TABLE 2: Phoneme transcriptions in Polish (continued)

SAMPA	example	transcriptions	occurrences	%
ts'	ćma	ts'ma	3,893,733	0.55
dz'	dźwig	dz'vik	2,843,134	0.40
N	pęk	peNk	2,095,984	0.30
z'	źle	z'le	1,396,567	0.20
dz	dzwoń	dzvon'	1,337,900	0.19
J	giełda	Jjewda	1,146,444	0.16
j~	więź	vjej~s'	721,950	0.10
dZ	dżem	dZem	273,615	0.04

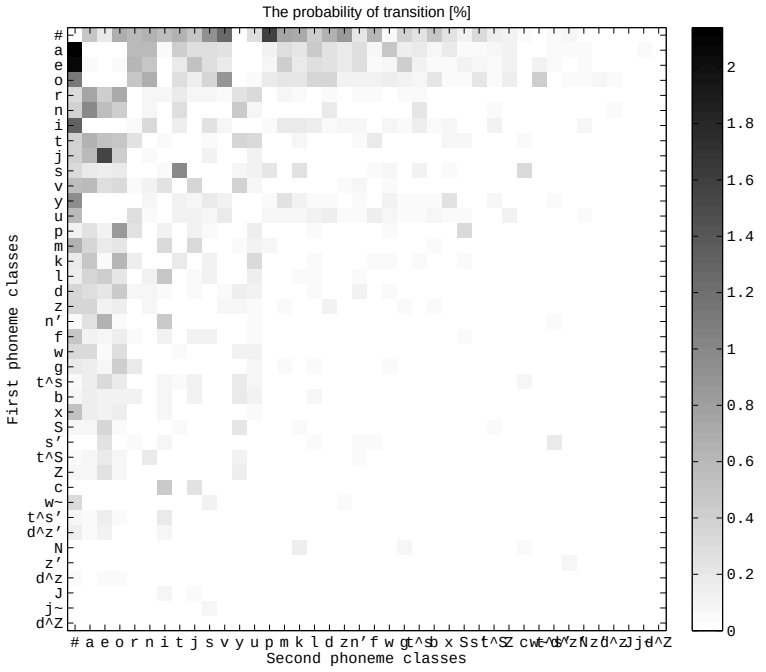

FIGURE 2: Phonemes in Polish in SAMPA alphabet

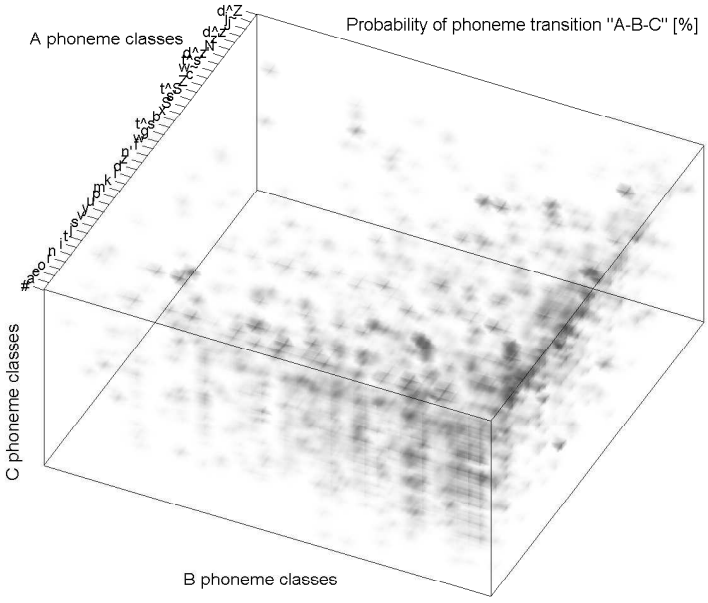

FIGURE 3: Phonemes in Polish in SAMPA alphabet

2 Method Description

Obtaining of phonetic information from an orthographic text-data is not straightforward (Holmes *et al.*, 1964; Ostaszewska and Tambor, 2000). Transcription of text into phonetic data has to be applied first (Oliver, 1998). We used PolPhone (Demenko *et al.*, 2003) software, which is described in the next section, for this aim. The SAMPA extended phonetic alphabet was applied with 39 symbols (plus space) and pronunciation rules typical for cities Kraków and Poznań as the standard options in PolPhone. For programming reasons we used our own single letter only symbols corresponding to SAMPA symbols, instead of typical ones, to distinguish phonemes easier while analysing received phonetic transcriptions. SED was used to change original phoneme transcriptions into a set of letters and digits designed by us to simplify calculations. Statistics can be now simply gathered by counting number of occurrences of each phoneme, phoneme pair, and phoneme triple in analysed text, where each phoneme is just a symbol (single letter or digit). The last, analysing, part was conducted in Matlab on a high performance computer in the Academic Computer Centre CYFRONET AGH. Mars, the computer used, has following specification: IBM Blade Center HS21 — 112 Intel Dual-core processors, 8GB RAM/core, 5 TB disk storage and 1192 Gflops. It operates using Red Hat Linux and allows every user to conduct 10 processes at once.

3 Text to Phonetic Transcription

Two main approaches are used for the automatic transcription of texts into phonemic forms. The classical approach is based on phonetic grammatical rules specified by human (Steffen-Batóg and Nowakowski, 1993) or machine learning process (Daelemans and van den Bosch, 1997). The second solution utilises graphemic-phonetic dictionaries. Both mentioned methods were used in PolPhone to cover typical and exceptional transcriptions. Polish phonetic transcription rules are relatively easy to formalise because of their regularity. In this experiment pronounciation variant for cities Kraków and Poznań was used. The foreign words and abbreviations were treated by applying general grapheme-to-phoneme rules for Polish. Numerals were skipped.

The necessity of investigating large text corpus pointed to the use of the Polish phonetic transcription system PolPhone (Jassem, 1996; Demenko *et al.*, 2003). In this system, strings of Polish characters are converted into their phonetic SAMPA representation. Extended SAMPA (Tables 1 and 1) is used, to deal with nuances of Polish phonetic system. The transcription process is performed by a table-based system, which implements the rules of transcription. Matrix $T[1..m][1..n]$ is a *transcription table* and its cells meet a set of requirements (Demenko *et al.*, 2003). The first element ($T[1][1]$) of each table contains currently processed character of the input string. For every character (or character substring) one table is defined. The first column of each table ($T[i][1]$, where $i = 1, ..., m$) contains all possible character strings that could preceed currently transcribed character. The first row ($T[1][j]$, where $j = 1, ..., m$) contains all possible character strings that can follow a currently transcribed character. All possible phonetic transcription results (in SAMPA) are stored in the remaining cells of the tables ($T[2..n][2..m]$). A particular element $T[i][j]$ is chosen as a transcription result, if $T[i][1]$ matches the substring preceeding $T[1][1]$ and $T[1][j]$ matches the substring proceeding $T[1][1]$. This basic scheme is extended to cover overlapping phonetic contexts. If more then one result is possible, then longer context is chosen for transcription, which increases its accuracy. Exceptions are handled by additional tables in the similar manner.

Specific transcription rules were designed by a human expert in an iterative process of testing and updating rules. Text corpora used in design process consisted of various sample texts (newspaper articles) and a few thousand words and phrases including special cases and exceptions.

4 Corpus and Results

Several hundreds of thousands of Internet articles in Polish were used as input data in our experiment. They are all from a high quality website where all content is reviewed and controlled by moderators. They are of encyclopedical type so they contain many names including foreign ones, what may influence the results slightly. In total, 754 Mbytes (around 94,000,000 words) were included in the process.

Total number of 703,032,406 phonemes were analysed. They are grouped into 40 categories (including space). Their distribution is presented in Tables 1 and

TABLE 3: Biphone statistics for Polish

biphone	occurrences	%	biphone	occurrences	%
a#	15,045,546	2.1405	v#	3,979,393	0.56615
e#	14,493,749	2.062	ne	3,936,303	0.56002
#p	11,164,873	1.5884	#i	3,842,357	0.54665
je	10,878,485	1.5477	te	3,746,928	0.53307
i#	9,268,528	1.3186	ej	3,661,116	0.52086
#v	8,726,541	1.2415	x#	3,635,418	0.51721
o#	7,929,368	1.1281	li	3,525,156	0.50152
na	6,950,150	0.98879	#a	3,525,085	0.50151
st	6,854,187	0.97514	f#	3,480,249	0.49513
y#	6,810,412	0.96891	ka	3,478,353	0.49486
#s	6,459,552	0.919	#b	3,476,948	0.49466
ov	6,278,554	0.89325	to	3,445,209	0.49015
po	5,956,505	0.84743	or	3,387,256	0.4819
#z	5,879,431	0.83646	en	3,369,798	0.47942
ra	5,291,504	0.75282	#j	3,359,347	0.47793
#m	5,269,923	0.74975	aw	3,332,003	0.47404
ro	5,061,958	0.72016	do	3,238,755	0.46078
#k	5,034,739	0.71629	ci	3,191,145	0.454
#o	4,906,760	0.69808	ny	3,156,592	0.44909
on	4,839,977	0.68858	n'i	3,156,070	0.44901
m#	4,684,961	0.66653	al	3,142,551	0.44709
#t	4,607,084	0.65545	le	3,036,888	0.43206
n'e	4,605,779	0.65526	at	3,032,032	0.43137
ta	4,585,711	0.65241	no	3,018,339	0.42942
#d	4,544,105	0.64649	go	2,989,946	0.42538
#n	4,501,621	0.64044	#l	2,986,668	0.42491
#f	4,457,240	0.63413	re	2,975,720	0.42335
er	4,436,631	0.6312	em	2,957,362	0.42074
ko	4,363,211	0.62075	eg	2,948,140	0.41943
va	4,210,423	0.59901	ow∼	2,857,801	0.40658
ar	4,149,362	0.59033	jo	2,838,947	0.4039
ja	4,116,780	0.58569	j#	2,800,720	0.39846
an	4,091,758	0.58213	vy	2,757,278	0.39228
u#	4,073,289	0.5795	#g	2,733,303	0.38887
#r	4,056,014	0.57705	t#	2,662,512	0.37879

Table 4: Biphone statistics for Polish (2nd part)

biphone	occurrences	%	biphone	occurrences	%
n#	2,648,122	0.37675	ot	2,011,588	0.28619
ma	2,579,022	0.36692	es	2,006,208	0.28542
vj	2,573,592	0.36614	#u	1,958,730	0.27867
ol	2,570,608	0.36572	an’	1,941,239	0.27618
za	2,562,161	0.36452	as	1,925,487	0.27394
ty	2,536,958	0.36093	aj	1,889,022	0.26875
od	2,471,143	0.35157	ur	1,884,386	0.26809
z#	2,428,847	0.34555	ad	1,871,173	0.26621
la	2,419,782	0.34426	pr	1,864,766	0.2653
Se	2,412,860	0.34328	ry	1,849,630	0.26315
os	2,408,470	0.34265	#x	1,845,924	0.26262
d#	2,388,259	0.33978	s’e	1,826,793	0.2599
s#	2,348,511	0.33412	is	1,792,576	0.25503
r#	2,343,088	0.33335	yx	1,782,674	0.25362
mj	2,342,027	0.3332	#t^s	1,775,430	0.25259
mi	2,326,378	0.33097	av	1,760,822	0.25051
wa	2,303,084	0.32766	sk	1,751,013	0.24912
vo	2,300,257	0.32726	ed	1,741,562	0.24777
t^se	2,292,301	0.32612	cj	1,722,926	0.24512
ru	2,284,982	0.32508	pa	1,710,691	0.24338
#s’	2,258,115	0.32126	tr	1,705,372	0.24262
pS	2,243,057	0.31912	vi	1,704,294	0.24247
w∼#	2,237,992	0.3184	ym	1,693,304	0.24091
in	2,196,720	0.31253	Ze	1,684,667	0.23968
ku	2,190,905	0.3117	n’a	1,678,440	0.23879
w#	2,179,393	0.31006	om	1,636,588	0.23284
sc	2,149,238	0.30577	ok	1,626,837	0.23145
tu	2,143,245	0.30492	lo	1,616,826	0.23002
el	2,090,592	0.29743	mo	1,602,709	0.22802
en’	2,082,946	0.29634	de	1,567,500	0.22301
wo	2,075,865	0.29533	ak	1,556,402	0.22143
da	2,063,451	0.29357	#n’	1,544,571	0.21975
nt	2,027,352	0.28843	ob	1,543,250	0.21956
ve	2,015,824	0.28679	nt^s	1,529,147	0.21755
am	2,012,911	0.28638	Sy	1,508,097	0.21456

TABLE 5: Triphone statistics for Polish

triphone	occurrences	%	triphone	occurrences	%
#po	4,651,870	0.66195	a#v	1,393,861	0.19834
#na	3,341,340	0.47547	#s'e	1,365,103	0.19425
na#	3,302,779	0.46998	s'e#	1,337,827	0.19037
#v#	2,690,086	0.38279	ka#	1,318,276	0.18759
n'e#	2,655,483	0.37787	#pr	1,274,496	0.18136
ej#	2,415,392	0.34371	sci	1,273,173	0.18117
#i#	2,299,241	0.32718	e#v	1,267,692	0.18039
ow∼#	2,228,643	0.31713	#n'e	1,184,756	0.16859
je#	2,219,065	0.31577	#te	1,176,184	0.16737
go#	2,170,585	0.30887	do#	1,168,435	0.16627
#f#	2,155,323	0.3067	ja#	1,165,087	0.16579
ego	2,151,570	0.30616	#mj	1,150,093	0.16366
sta	2,083,571	0.29649	jon	1,145,854	0.16305
ova	2,047,299	0.29133	ym#	1,114,268	0.15856
#za	2,035,676	0.28967	jej	1,112,770	0.15834
#do	2,001,403	0.2848	aw#	1,087,940	0.15481
#pS	1,992,629	0.28355	ku#	1,045,704	0.1488
ci#	1,927,816	0.27432	wa#	1,027,104	0.14615
vje	1,920,947	0.27335	ost	1,026,990	0.14614
#je	1,826,989	0.25998	#z#	1,011,932	0.144
ne#	1,755,954	0.24987	ent	1,000,758	0.14241
cje	1,707,782	0.24301	#li	994,834	0.14156
yx#	1,687,450	0.24012	i#p	994,391	0.1415
mje	1,667,373	0.23726	e#s	964,548	0.13725
ny#	1,650,197	0.23482	e#z	962,713	0.13699
a#p	1,635,167	0.23268	#pa	934,503	0.13298
#vy	1,607,278	0.22871	#ja	933,178	0.13279
#ro	1,589,023	0.22611	ont^s	921,679	0.13115
#st	1,582,732	0.22522	f#p	921,488	0.13113
pSe	1,542,211	0.21945	ove	919,707	0.13087
e#p	1,515,309	0.21563	ax#	907,397	0.12912
em#	1,492,716	0.21241	nyx	907,030	0.12907
#ko	1,488,319	0.21178	a#s	895,522	0.12743
#ma	1,460,864	0.20788	os't^s'	891,770	0.1269
n'a#	1,410,948	0.20077	#ka	891,704	0.12689

TABLE 6: Triphone statistics for Polish (2nd part)

triphone	occurrences	%	triphone	occurrences	%
#vj	888,656	0.12645	s'tˆs'i	738,031	0.10502
sto	884,894	0.12592	a#z	734,672	0.10454
ter	880,196	0.12525	str	728,361	0.10364
scj	876,048	0.12466	#tˆse	726,492	0.10338
jeg	867,120	0.12339	er#	723,151	0.1029
pol	863,858	0.12293	ste	721,009	0.1026
o#p	863,394	0.12286	i#z	716,534	0.10196
tˆse#	854,191	0.12155	va#	715,369	0.1018
tur	846,452	0.12045	e#n	703,914	0.10017
ist	844,202	0.12013	jow∼	700,644	0.0997
ovy	831,075	0.11826	o#v	698,196	0.099352
uv#	825,181	0.11742	#s#	697,785	0.099293
jer	824,115	0.11727	e#o	696,794	0.099152
#vo	814,771	0.11594	kov	694,941	0.098889
tˆs'i#	814,724	0.11593	pje	693,989	0.098753
#to	808,403	0.11503	ktu	690,887	0.098312
i#v	799,853	0.11382	pov	686,389	0.097672
van	797,368	0.11346	ast	685,080	0.097485
uf#	797,040	0.11342	im#	677,256	0.096372
to#	787,211	0.11202	e#d	675,538	0.096128
a#m	786,986	0.11199	le#	674,144	0.095929
est	784,805	0.11168	ona	672,802	0.095738
mi#	776,160	0.11045	#by	672,333	0.095672
ovo	774,574	0.11022	a#n	669,471	0.095264
ovj	768,663	0.10938	ale	664,494	0.094556
e#m	768,628	0.10937	a#o	659,578	0.093857
#re	766,182	0.10903	#ta	651,804	0.09275
pro	763,416	0.10863	yst	651,562	0.092716
ta#	760,823	0.10826	#gr	650,239	0.092528
y#p	760,190	0.10817	awa	647,742	0.092172
a#k	754,819	0.10741	ajo	646,304	0.091968
e#f	748,358	0.10649	a#t	644,090	0.091653
Ze#	747,594	0.10638	a#d	641,233	0.091246
any	747,339	0.10634	ko#	640,803	0.091185
#mo	742,502	0.10566	#sp	640,348	0.09112

Table 7: Triphone statistics for Polish (3rd part)

triphone	occurrences	%	triphone	occurrences	%
pra	639,526	0.091003	tˆs'e#	594,140	0.084545
ra#	634,601	0.090302	li#	592,447	0.084304
ali	634,426	0.090277	ve#	589,183	0.083839
#va	631,844	0.08991	rov	588,302	0.083714
tem	631,723	0.089893	#al	587,217	0.08356
pSy	630,562	0.089728	sko	585,499	0.083315
a#f	630,282	0.089688	a#r	584,427	0.083163
jed	628,568	0.089444	#in	582,925	0.082949
#od	626,681	0.089175	#zo	581,066	0.082684
nov	625,673	0.089032	an'e	580,511	0.082605
tˆsi#	625,387	0.088991	tra	579,760	0.082499
dov	614,157	0.087393	e#t	579,083	0.082402
gra	613,644	0.08732	ane	578,436	0.08231
tˆSne	608,173	0.086542	mja	576,316	0.082009
#la	603,222	0.085837	ejs	574,988	0.08182
#a#	599,813	0.085352	n'i#	573,251	0.081572
tˆsa#	596,945	0.084944	#sa	571,143	0.081272
tor	596,464	0.084876	tˆsy#	567,955	0.080819
zna	595,816	0.084783	e#k	566,787	0.080653
st#	595,323	0.084713	taw	566,630	0.08063

2 and in Fig. 1. 1,232 different biphones (Fig. 2 and Tables 3 and 4) for 1,560 possible combinations were found (79%). 20,139 different triphones (Fig. 3 and Tables 5, 6 and 7) were found. It has to be mentioned that all combinations like *#*, where * is any phoneme and # is space, were removed as we do not treat these triples as triphones. The reason for it, is that first phoneme * and the second one are actually in 2 different words, while we are interested in triphone statistics inside words. The list of the most common triphones is presented in Tables 5, 6 and 7. Assuming 40 different phonemes (including space) and subtracting mentioned *#* combinations, there are 62,479 possible triples. We found 20,139 different triphones. It leads to a conclusion that around 32% of possible combinations were actually found as triphones, which is more than in our previous experiment (Ziółko *et al.*, 2007). Space (noted as #) frequency was 14.92. Let us divide 100 by 14.92 to receive an average length of words in phonemes as 6.7. It includes one space after each word, which means that the real average length is less then 6.

Fig. 2 shows some symmetry. Of course, the probability of biphone $\alpha\beta$ is usually different than probability of $\beta\alpha$. Some symmetry results from the fact that high values of α probability and β probability gives usually high probability of product $\alpha\beta$ and $\beta\alpha$ as well. Similar effects can be observed for triphones.

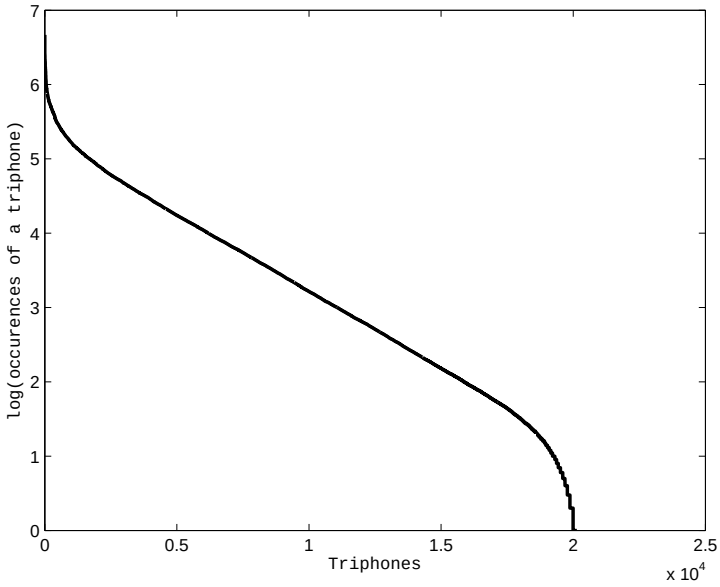

FIGURE 4: Polish Phonemes in SAMPA alphabet

Data presented in this paper illustrate the well-known fact that probabilities of triphones (presented in Tables 5, 6 and 7) cannot be calculated from the biphone probabilities (some of them are presented in Tables 3 and 4). The reason for this is that the conditional probabilities have to be known.

We observed that all statistics, even the phoneme one (Tables 1, 2 and Fig. 1), are different in this experiment then in the previous one (Ziółko *et al.*, 2007). They also differ slightly from statistics we collected from a literature corpus, which are under review now. We used a slightly different version of SAMPA alphabet there, but the differences between experiments, in order of phonemes can be easily spotted. In (Ziółko *et al.*, 2007) phonemes were ordered by frequency in the list: a, e, o, s, t, r, p, v, j, i, I, n, l, u, k, z, m, d, n', f, ts, g, S, b, x, tS, dz, ts', dz', Z, s', o~, N, w, z', dZ, e~. It leads to a conclusion that the results are not fully representative and even more data should be analysed to provide the frequency of phonemes as proper linguistic data. Even though, our results are very useful for several engineering tasks. Some of them are presented in the next section.

Besides the frequency of triphones occurring, we are also interested in distributions of different frequencies, which is presented in logarithmic scale in Fig. 4. We received another distribution than in the previous experiment (Ziółko *et al.*, 2007) because very large number of words were analysed. We have few triphones which occurred rarely. We noted around 200 triphones which occurred just once and around 100 with occurrences 2 to 7. It supports a hypothesis that one can reach a situation, when new triphones do not appear and a distribution of occurrences is changing as a result of more data being analysed. Then a threshold can be set and

the rarliest triphones can be removed as errors. Some triphones with very small occurrence are non-Polish triphones which should be excluded from the statistics. The rare triphones come from unusual Polish word combinations, slang and other variations of dictionary words, onomatopoeic words, foreign words, errors in phonisation and typos in the text corpus. In our further works we will assume that from statistical point of view it is not important, especially when smoothing operation is applied in order to eliminate disturbances caused by lack of text data (Rabiner, 1989; Altwarg, 2000).

5 N-gram Probability Model

N-gram models are the most often applied for word statistics. However, the same mathematical tool can be applied for sequences of phonemes. Context-dependent units can improve recognition highly. In speech, applying biphones and triphones gives this opportunity. Phonemes vary slightly depending on the context — neighbouring phonemes due to a natural phenomena of coarticulation. There are no clear boundaries between phonemes. They rather overlap each other which results in starts and ends being dependant on other phonemes. Speech recognisers based on triphone models rather than phoneme ones are much more complex but give better results (Rabiner and Juang, 1993). Examples of transcribing word *above*: phoneme model *ax b ah v* and triphone model **-ax+b ax-b+ah b-ah+v ah-v+**. In case a specific triphone is not present, it can be replaced by a phonetically similar triphone (phonemes of the same phonetic group interfere in similar way with their neighbours) or a biphone (applying only left or right context).

6 Conclusions

94,000,000 words in Internet, encyclopedia articles were analysed and statistics of Polish phonemes, biphones and triphones were created in this way. They are not fully complete but the corpus was large enough, that they can be successfully used for language modelling. 32% of possible triples were detected as triphones, most of them at least several times. The full statistics are available on request by an email.

7 Acknowledgements

This work was supported by MNISW grant number OR00001905. We would like to thank Institute of Linguistics, Adam Mickiewicz University for providing PolPhone — a software tool to make a phonetic transcription for Polish.

References

E. Agirre, O. Ansa, D. Martínez, and E. Hovy (2001), Enriching WordNet concepts with topic signatures, *Procceedings of the SIGLEX Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations.*

R. Altwarg (2000), Language Models in Speech Recognition, www.shlrc.mq.edu.au/masters/students/raltwarg/lmtoc.htm.

J. R. Bellegarda (2000), Large Vocabulary Speech Recognition with multispan Statistical Language Models, *IEEE Transactions on Speech and Audio Processing*, 8(1):76–84.

W. Daelemans and A. van den Bosch (1997), Language-Independent Data-Oriented Grapheme-to-Phoneme Conversion, *Progress in Speech Synthesis, New York: Springer-Verlag.*

G. Demenko, M. Wypych, and E. Baranowska (2003), Implementation of Grapheme-to-phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-speech Synthesis, *Speech and Language Technology, PTFon, Poznań*, 7(17).

P. B. Denes (1962), Statistics of Spoken English, *The Journal of the Acoustical Society of America*, 34:1978–1979.

J.N Holmes, I.G. Mattingley, and J.N. Shearme (1964), Speech synthesis by rule, *Language and Speech*, 7:127–143.

K. Jassem (1996), A phonemic transcription and syllable division rule engine, *Onomastica-Copernicus Research Colloquium, Edinburgh.*

B. Kollmeier and M. Wesselkamp (1997), Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment, *The Journal of the Acoustical Society of America*, 102:2412–2421.

D. Oliver (1998), *Polish Text to Speech Synthesis, MSc. Thesis in Speech and Language Processing*, Edinburgh University, Edinburgh.

D. Ostaszewska and J. Tambor (2000), *Fonetyka i fonologia współczesnego języka Polskiego (eng. Phonetics and phonology of modern Polish language)*, PWN.

L. Rabiner and B. H. Juang (1993), *Fundamentals of speech recognition*, PTR Prentice-Hall, Inc., New Jersey.

L. R. Rabiner (1989), A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, 77(2):257–286.

M. Steffen-Batóg and P. Nowakowski (1993), An algorithm for phonetic transcription of ortographic texts in Polish, *Studia Phonetica Posnaniensia*, 3.

E. J. Yannakoudakis and P. J. Hutton (1992), An assessment of N-phoneme statistics in phoneme guessing algorithms which aim to incorporate phonotactic constraints, *Speech Communication*, 11:581 – 602.

S. Young, G. Evermann, M. Gales, Th. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland (2005), *HTK Book*, Cambridge University Engineering Department, UK.

B. Ziółko, J. Gałka, S. Manandhar, R.C. Wilson, and M. Ziółko (2007), Triphone Statistics for Polish Language, *Proceedings of 3rd Language and Technology Conference, Poznań.*

B. Ziółko, S. Manandhar, R. C. Wilson, and M. Ziółko (2008), Semantic Modelling for Speech Recognition, *Proceedings of Speech Analysis, Synthesis and Recognition. Applications in Systems for Homeland Security, Piechowice, Poland.*