Computer Science • Vol. 10 • 2009

Bartosz Ziółko*, Jakub Gałka*, Mariusz Ziółko*

POLISH PHONEME STATISTICS OBTAINED ON LARGE SET OF WRITTEN TEXTS

The phonetical statistics were collected from several Polish corpora. The paper is a summary of the data which are phoneme n-grams and some phenomena in the statistics. Triphone statistics apply context-dependent speech units which have an important role in speech recognition systems and were never calculated for a large set of Polish written texts. The standard phonetic alphabet for Polish, SAMPA, and methods of providing phonetic transcriptions are described.

Keywords: *NLP, triphone statistics, speech processing, Polish*

STATYSTYKI POLSKICH FONEMÓW UZYSKANE Z DUŻYCH ZBIORÓW TEKSTÓW

W niniejszej pracy zaprezentowano opis statystyk głosek języka polskiego zebranych z dużej liczby tekstów. Triady głosek pełnia istotną rolę w rozpoznawaniu mowy. Omówione obserwacje dotyczące zebranych statystyk i przedstawiono listy najpopularniejszych elementów.

Słowa kluczowe: *przetwarzanie języka naturalnego, statystyki głosek, przetwarzanie mowy*

1. Introduction

The authors uses the Cyfronet, high performance computers to process linguistic data in aim to construct the Polish language models. The results will be applied to a large vocabulary continuous speech recognition system (LVCSR). Natural language processing (NLP) faces problems of data sparsity very often. The quality of language models is strongly dependant on the amount of text corpora available during the training. This is why, there is a trade-off of quality and time spent on calculations. The high performance computers facilitate obtaining the linguistic rules from the huge amount of texts.

Statistical linguistics at the word and sentence level were under considerations for several languages [1, 2]. However, similar research on phonemes is rare [3, 4, 5]. The frequency of phonetic units appearance is an important topic itself for every

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

language. It can also be used in several speech processing applications, for example modelling in LVCSR or coding and compression. Models of triphones which are not present in a training corpus of a speech recogniser can be prepared using phonetic decision trees [6]. The list of possible triphones has to be provided for a particular language along with phonemes' categorisation. The triphone statistics can also be used to generate hypotheses used in recognition of out-of-dictionary words including names and addresses.

We have already presented some similar statistics [7], which were collected from around 10 000 000 words of mainly spoken language. Data collected from a few much larger corpora: Rzeczpospolita corpus (containing articles from a well known in Poland, everyday newspaper of quality and type like Times or Guardian), literature corpus and Internet encyclopedia corpus are presented in this work combined statistical. The presented statistics are the biggest and most representative statistics of phonemes for Polish. They were collected from over 250 000 000 words.

2. Description of a problem solution

The problem is to find triphone statistics for Polish language. Our first attempt to this task was already published [7]. The task was conducted on a corpus containing Parliament transcriptions mainly (around 50 megabytes of text). It was repeated on Mars, a Cyfronet computer cluster, for data of around 2 gigabytes.

Context-dependent modelling can significantly improve speech recognition quality. Each phoneme varies slightly depending on its context, namely neighbouring phonemes due to a natural phenomena of coarticulation. It means that there are no clear boundaries between phonemes and they overlap each other. It results in interference of acoustical properties. Speech recognisers based on triphone models rather than phoneme ones are much more complex but give better results [9]. Let us present examples of different ways of transcribing word *above*. Phoneme model is *ax b ah v* while the triphone one is **-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) using phonetic decision trees [6] or diphones (applying only left or right context) [10].

3. Methods, software and hardware

Sophisticated rules and methods are necessary to obtain the phonetic information from an orthographic text-data. Simplifications could cause errors [11]. Transcription of text into phonetic data was applied first by PolPhone [8]. The extended SAMPA phonetic alphabet was applied with 39 symbols (plus space) and pronunciation rules for cities Poznań and Kraków. We used our own digit symbols corresponding to SAMPA symbols, instead of typical ones, to distinguish phonemes easier while analysing received phonetic transcriptions.

Table 1
Phonemes in Polish (SAMPA [8])

SAMPA	example	transcr.	occurr.	%	% [5]
#		#	283 296 436	15.256	4.7
a	pat	pat	151 160 947	8.141	9.7
e	test	test	146 364 208	7.882	10.6
o	pot	pot	141 975 325	7.646	8.0
t	test	test	68 851 605	3.708	4.8
r	ryk	rIk	68 797 073	3.705	3.2
n	nasz	naS	68 056 439	3.665	4.0
i	PIT	pit	67 212 728	3.620	3.4
j	jak	jak	61 265 911	3.299	4.4
I	typ	tIp	58 930 672	3.174	3.8
v	wilk	vilk	58 247 951	3.137	2.9
s	syk	sIk	54 359 454	2.927	2.8
u	puk	puk	51 503 621	2.774	2.8
p	pik	pik	51 228 649	2.759	3.0
m	mysz	mIS	48 760 010	2.626	3.2
k	kit	kit	44 892 420	2.418	2.5
d	dym	dIm	44 406 412	2.391	2.1
l	luk	luk	40 189 121	2.164	1.9
n’	koń	kon’	34 092 610	1.84	2.4
z	zbir	zbir	30 924 282	1.665	1.5
w	łyk	wIk	30 194 178	1.626	1.8
f	fan	fan	25 308 167	1.363	1.3
g	gen	gen	24 910 462	1.341	1.3
t^s	cyk	t^sIk	24 789 080	1.335	1.2
b	bit	bit	24 212 663	1.304	1.5
x	hymn	xImn	21 407 209	1.153	1.0
S	szyk	SIk	20 756 164	1.118	1.9
s’	świt	s’vit	17 220 321	0.927	1.6
Z	żyto	ZIto	16 409 930	0.884	1.3
t^S	czyn	t^SIn	15 429 711	0.831	1.2
t^s’	ćma	t^s’ma	11 945 381	0.643	1.2
w~	ciąża	ts’ow~Za	10 814 216	0.582	0.6
c	kiedy	cjedy	10 581 296	0.570	0.7
d^z’	dźwig	d^z’vik	9 995 596	0.538	0.7
N	pęk	peNk	4 880 260	0.262	0.1
d^z	dzwoń	d^zvon’	4 212 857	0.227	0.2
J	giełda	Jjewda	3 680 888	0.198	0.1
z’	źle	z’le	3 390 372	0.183	0.2
j~	więź	vjej~s’	1 527 778	0.082	0.1
d^Z	dżem	d^Zem	693 838	0.037	0.1

Stream editor (SED) was applied to change original phoneme transcriptions into digits with the following script:

s/##/#/g	s/w~/2/g	s/d^z/6/g
s/t^s'/8/g	s/s'/5/g	s/t^S/0/g
s/d^z'/X/g	s/z'/4/g	s/d^Z/9/g
s/j~/1/g	s/t^s/7/g	s/n'/3/g.

Statistics can now be simply collected by counting the number of occurrences of each phoneme, phoneme pair, and phoneme triple in an analysed text, where each phoneme is just a symbol (single letter or a digit). Matlab was used to analyse the phonetic transcription of the text corpora. The calculations were conducted on Mars in Cyfronet, Krakow. We analysed more than 2 gigabytes of data. Text data for Polish are still being collected and will be included in the statistics in the future.

Mars is a cluster for calculations with following specification: IBM Blade Center HS21 – 112 Intel Dual-core processors, 8 GB RAM/core, 5 TB disk storage and 1192 Gflops. It operates using Red Hat Linux. Mars uses Portable Batch System (PBS) to queue tasks and split calculation power to optimise times for all users. A user have to declare expected time of every task. In example, a short time is up to 24 hours of calculations and a long one is up to 300 hours. Tasks can be submitted by simple commands with scripts and the cluster starts particular tasks when calculation resources are available. One process needs around 100 hours to analyse 45 megabytes text file.

3.1. Grapheme to phoneme 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 [12] or machine learning process [13]. The second solution utilises graphemic-phonetic dictionaries. Both methods were used in PolPhone to cover typical and exceptional transcriptions. Polish phonetic transcription rules are relatively easy to formalise because of their regularity.

The necessity of investigating large text corpus pointed to the use of the Polish phonetic transcription system PolPhone [14, 8]. In this system, strings of Polish characters are converted into their phonetic SAMPA representations. Extended SAMPA (Table 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 $T \in S^{m \times n}$ is a *transcription table*, where S is a set of strings and the cells meet the requirements listed precisely in [8]. 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 $\{t_{i,1}\}_{i=1}^{m}$ contains all possible character strings that could precede currently transcribed character. The first row $\{t_{1,j}\}_{j=1}^{n}$ contains all possible character strings that can proceed a currently transcribed character. All possible phonetic transcription results are stored in the remaining cells $\{t_{i,j}\}_{i=2,j=2}^{m,n}$. A particular element $t_{i,j}$ is chosen as a transcription result, if $t_{i,1}$ matches the substring preceding $t_{1,1}$ 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.

3.2. Corpora used

Several newspaper articles in Polish were used as input data in our experiment. They are from Rzeczpospolita newspaper from years 1993–2002. They cover mainly political and economic issues, so they contain quite many names and places including foreign ones, what may influence the results slightly. In example, q appeared once, even though it does not exist in Polish. In total, 879 megabytes of text, which corresponds to around 110 000 000 words, were included in the process.

Several hundreds of thousands of Internet articles in Polish made another corpus. They are all from a high quality website, where all content is reviewed and controlled by moderators. They are of encyclopedia type, so they also contain many names including foreign ones. In total, 754 megabytes (around 94 000 000 words) were included in the process.

The third corpus consists of several literature books in Polish. Some of them are translations from other languages, so they also contain foreign words. The corpus includes 490 megabytes (around 61 000 000 words) of text.

4. Results

The total number of around 1856 900 000 phonemes were analysed. They are grouped into 40 categories (including space). Actually, one more, namely q, was detected, which appeared in a foreign name. Since q is not a part of the Polish alphabet, it was not included in the phoneme distribution presented in Table 1. Space (noted as #) frequency was 15.26 %. An average number of phonemes in words is 6.6 including one space. Exactly 1271 different diphones (Fig. 1 and Table 2) for 1560 possible combinations were found, which constitutes 81%.

21 961 different triphones (see Table 3) were detected. Combinations like *#*, where * is any phoneme and # is a space were removed. These triples should not be considered as triphones because the first and the second * are in two different words. The list of the most common triphones is presented in Table 3. Assuming 40 different phonemes (including space) and subtracting mentioned *#* combinations, there are 62 479 possible triples. We found 21 961 different triphones. It leads to a conclusion that around 35% of possible triples were detected as triphones, the very most of them at least 10 times.

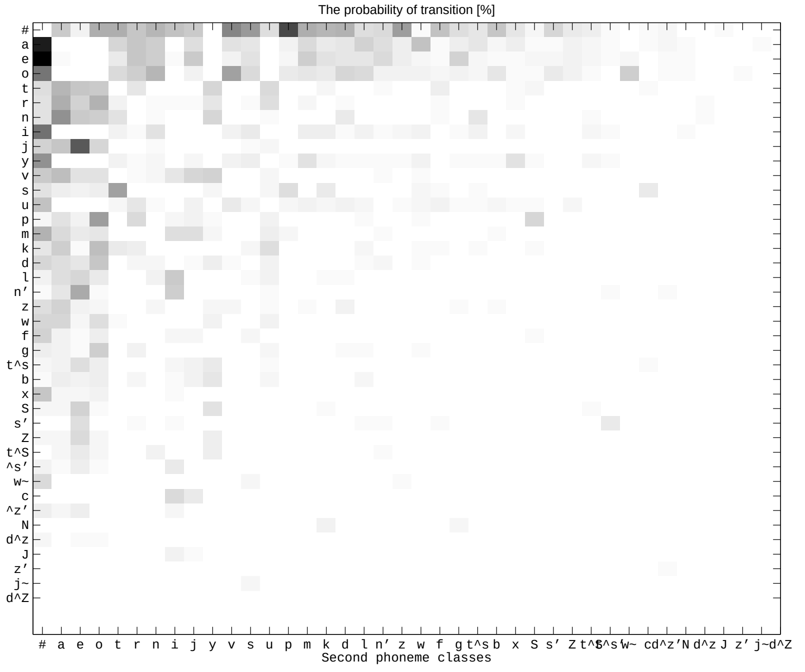

Fig. 1. Frequency of diphones in Polish (each phoneme separately)

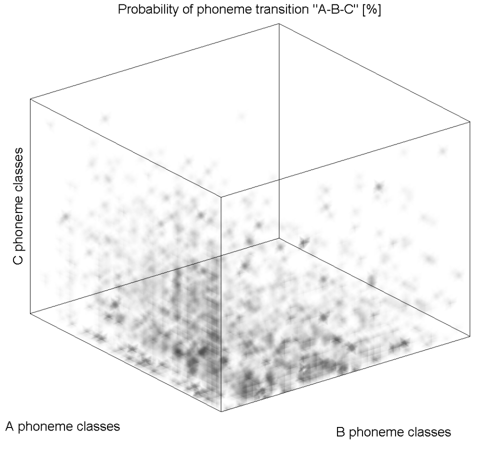

Fig. 2. Space of triphones in Polish

Table 2
Most common Polish diphones

diphone	no. of occurr.	%	diphone	no. of occurr.	%
e#	43 557 832	2.346	on	12 854 255	0.692
a#	38 690 469	2.084	#k	12 529 124	0.675
#p	31 014 275	1.671	ta	12 449 178	0.671
je	28 499 593	1.535	#n	12 316 393	0.663
i#	24 271 474	1.307	va	11 413 878	0.615
o#	23 552 591	1.269	ko	11 168 294	0.602
#v	20 678 007	1.114	#i	10 515 253	0.566
y#	19 018 563	1.024	aw	10 514 514	0.566
na	18 384 584	0.990	u#	10 379 234	0.559
#s	17 321 614	0.933	#f	10 265 162	0.553
po	16 870 118	0.909	#b	10 167 482	0.548
#z	16 619 556	0.895	#r	10 137 129	0.546
ov	16 206 857	0.873	ja	10 097 444	0.544
st	15 895 694	0.856	ar	9 818 127	0.529
n'e	14 851 771	0.800	x#	9 811 211	0.528
#o	14 104 742	0.760	do	9 779 666	0.527
#t	13 910 147	0.749	er	9 724 692	0.524
ra	13 713 928	0.739	te	9 618 998	0.518
#m	13 657 073	0.736	#j	9 398 210	0.506
ro	13 597 891	0.732	v#	9 251 288	0.498
#d	13 103 398	0.706	#a	9 143 021	0.492
m#	12 968 346	0.698	to	9 043 529	0.487

Young [9], estimates that in English, 60–70% of possible triples exist as triphones. However, in his estimation there is no space between words, what changes the distribution a lot. Some triphones may not occur inside words but may occur at combinations of an end of one word and the beginning of another. We started to calculate such statistics without an empty space as the next step of our research. It is also expected that there are different numbers of triphones for different languages. Some values are similar to statistics given by Jassem a few decades ago and reprinted in [5]. We applied computer clusters so our statistics were calculated for much more data and they are more represantative.

Fig. 1 shows some symmetry but the probability of diphone $\alpha\beta$ is usually different than probability of $\beta\alpha$. The mentioned quasi symmetry results from the fact that high values of α probability and (or) β probability often gives high probability of products $\alpha\beta$ and $\beta\alpha$ as well. Similar effects can be observed for triphones. Data presented in this paper illustrate the well-known fact that probabilities of triphones (see Table 3) cannot be calculated from the diphone probabilities (see Table 2). The conditional probabilities between diphones have to be known.

Table 3
Most common Polish triphones

triphone	no. of occurr.	%	triphone	no. of occurr.	%
#po	12 531 515	0.675	wa#	3 262 204	0.176
#na	9 587 483	0.516	do#	3 210 532	0.173
n'e#	9 178 080	0.494	#ma	3 209 675	0.173
na#	8 588 806	0.463	jon	3 082 879	0.166
ow~#	6 778 259	0.365	e#z	3 054 967	0.165
#do	6 751 495	0.364	a#v	3 028 787	0.163
#za	6 429 379	0.346	#z#	2 928 164	0.158
ej#	6 390 911	0.344	ka#	2 871 230	0.155
je#	6 388 032	0.344	#sp	2 818 515	0.152
#pS	6 173 458	0.333	ontˆs	2 754 934	0.148
go#	5 990 895	0.323	e#s	2 737 210	0.147
#i#	5 945 409	0.320	i#p	2 725 414	0.147
ego	5 742 711	0.309	o#p	2 719 121	0.146
ova	5 560 749	0.300	#Ze	2 701 194	0.145
vje	5 433 154	0.293	#ja	2 670 034	0.144
#v#	5 317 078	0.286	ta#	2 618 595	0.141
#je	5 311 716	0.286	ent	2 612 166	0.141
#n'e	5 292 103	0.285	#to	2 567 269	0.138
sta	4 983 295	0.268	to#	2 557 630	0.138
#s'e	4 861 117	0.262	pro	2 548 979	0.137
yx#	4 858 960	0.262	pra	2 539 424	0.137
#vy	4 763 697	0.257	#pa	2 503 153	0.135
s'e#	4 746 280	0.256	#re	2 502 443	0.135
pSe	4 728 565	0.255	ost	2 490 304	0.134
e#p	4 727 840	0.255	#ty	2 452 830	0.132
#f#	4 660 745	0.251	tˆse#	2 436 864	0.131
em#	4 514 478	0.243	#mj	2 397 741	0.129
#pr	4 428 341	0.239	ku#	2 383 231	0.128
#ko	4 216 459	0.227	e#m	2 379 510	0.128
a#p	4 155 732	0.224	ja#	2 353 638	0.127
ci#	3 965 693	0.214	e#o	2 343 622	0.126
ne#	3 958 262	0.213	a#s	2 336 272	0.126
cje	3 916 595	0.211	#vj	2 329 962	0.125
n'a#	3 888 279	0.209	#mo	2 320 091	0.125
#ro	3 785 754	0.204	nyx	2 299 719	0.124
mje	3 760 340	0.203	os'tˆs'	2 295 365	0.124
#st	3 745 320	0.202	ovy	2 284 782	0.123
aw#	3 596 680	0.194	sci	2 282 887	0.123
ny#	3 580 425	0.193	ove	2 262 277	0.122
#te	3 449 304	0.186	li#	2 255 403	0.121
e#v	3 313 798	0.178	ovj	2 251 294	0.121
Ze#	3 309 352	0.178	mi#	2 243 432	0.121
ym#	3 300 273	0.178	uv#	2 236 507	0.120

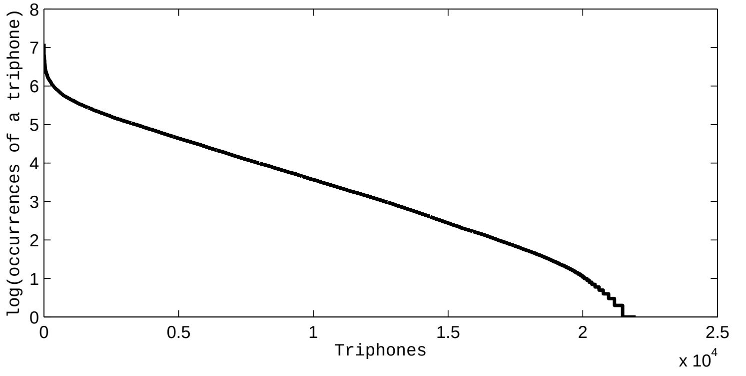

Fig. 3. Phoneme occurrences distribution

Besides the frequency of triphone occurrence, we are also interested in distributions of their frequencies. This is presented in logarithmic scale in Fig. 3. We received another distribution than in the previous experiment [7] because larger number of words were analysed. We have found around 500 triphones which occurred once and around 300 which occurred two or three times. Then every occurrence up to 10 happened for 100 to 150 triphones. 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. Some threshold can be set and the rarliest triphones can be removed as errors caused by unusual Polish word combinations, acronyms, slang and other variations of dictionary words, onomatopoeic words, foreign words, errors in phonisation and typographical errors in the text corpus.

Entropy:

$$H = -\sum_{i=1}^{40} p(i) \log_2 p(i), \tag{1}$$

where $p(i)$ is a probability of a particular phoneme, is used as a measure of the disorder of a linguistic system. It describes how many bits in average are needed to describe phonemes. According to Jassem in [5] entropy for Polish is 4.7506 bits/phoneme. From our calculations entropy for phonemes is 4.6335, for diphones 8.3782 and 11.5801 for triphones.

5. Conclusions

250 000 000 words from different corpora: newspaper articles, Internet and literature were analysed. Statistics of Polish phonemes, diphones and triphones were created. They are not fully complete, but the corpora were large enough, that they can be successfully applied in NLP applications and speech processing. The collected statistics are the biggest for Polish of this type of linguistic computational knowledge. Polish is

one of most common Slavic languages. It has several different phonemes than English and the statistics of phonemes are also different.

Acknowledgements

This work was supported by MNISW OR00001905.

References

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

[2] Bellegarda J. R.: *Large vocabulary speech recognition with multispan statistical language models*, IEEE Transactions on Speech and Audio Processing, vol. 8, no. 1, pp. 76–84, 2000

[3] Denes P. B.: *Statistics of spoken English*, The Journal of the Acoustical Society of America, vol. 34, pp. 1978–1979, 1962

[4] Yannakoudakis E. J., Hutton P. J.: *An assessment of n-phoneme statistics in phoneme guessing algorithms which aim to incorporate phonotactic constraints*, *Speech Communication*, vol. 11, pp. 581–602, 1992

[5] Basztura C.: *Rozmawiać z komputerem*, (Eng. *To speak with computers*). Format, 1992

[6] Young S., Evermann G., Gales M., Hain T., Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland P.: *HTK Book*. UK: Cambridge University Engineering Department, 2005

[7] Ziółko B., Gałka J., Manandhar S., Wilson R., Ziółko M.: *Triphone statistics for polish language*, Proceedings of 3rd Language and Technology Conference, 2007

[8] Demenko G., Wypych M., Baranowska E.: *Implementation of grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis*, Speech and Language Technology, PTFon, Poznań, vol. 7, no. 17, 2003

[9] Young S.: *Large vocabulary continuous speech recognition: a review*, IEEE Signal Processing Magazine, vol. 13(5), pp. 45–57, 1996

[10] Rabiner L., Juang B. H.: *Fundamentals of speech recognition*. New Jersey: PTR Prentice-Hall, Inc., 1993

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

[12] Steffen-Batóg M., Nowakowski P.: *An algorithm for phonetic transcription of ortographic texts in Polish*, Studia Phonetica Posnaniensia, vol. 3, 1993

[13] Daelemans W., Bosch, van den, A.: *Language-independent data-oriented grapheme-to-phoneme conversion*, Progress in Speech Synthesis, New York: Springer-Verlag, 1997

[14] Jassem K.: *A phonemic transcription and syllable division rule engine*, Onomastica-Copernicus Research Colloquium, Edinburgh, 1996