Phoneme Ngrams Based on a Polish Newspaper Corpus

Bartosz Ziółko, Jakub Gałka, Mariusz Ziółko
Department of Electronics, AGH University of Science and Technology, Kraków, Poland

Abstract— *The phonetical statistics of Polish were collected from a newspaper corpus of around 110 000 000 words. The paper presents summarisation of the data which are phoneme ngrams and some phenomena in the statistics including a distribution of frequency of triphones occurring. Triphone statistics apply context-dependent speech units which have an important role in automatic speech recognition systems. The standard phonetic alphabet for Polish, SAMPA, and methods of providing phonetic transcriptions are described.*

Keywords: *phoneme statistics, triphone statistics, Polish*

1. Introduction to Phoneme Statistics

Statistical linguistics at the word and sentence level is popular for several languages [1], [2]. Any similar research on phonemes is rare [3], [4], [5] and almost purely for English. The frequency of phonetic units appearance in natural language is an important topic itself for every language. It can also be used in several speech processing applications, for example modelling in 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. Phonetic decision trees can be used [6], [7] for this task but the list of possible triphones has to be provided for a particular language along with phonemes categorisation. The triphone statistics can be also used to generate hypotheses used in recognition of out-of-dictionary words including names and addresses.

We have already presented some similar statistics [8], which were collected from around 10 000 000 words of mainly spoken language. Here we present statistical data collected from much larger Rzeczpospolita corpus containing articles from a well known in Poland, every day newspaper of quality and type like Times or Guardian. We conducted similar experiments on large literature and Internet corpora and their results were just accepted for publication. Experiments on different corpora will allow to compare these statistics to evaluate how representative and complete they are. The choice of a corpus results in type of found linguistic phenomena.

This paper describes several issues related to phoneme, diphone and triphone statistics which can be also called ngrams. The paper 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, our results and some phenomena we uncovered. Section 5 presents opportunities of applying statistics we collected in natural language and speech processing for artificial intelligence tasks like automatic speech recognition. The paper is summed up the paper with conclusions.

Table 1: *Phoneme transcription in Polish - SAMPA [9]*

SAMPA	example	transcr.	occurr.	%
#		#	110 475 957	14.99
a	pat	pat	59 808 483	8.12
o	pot	pot	57 141 107	7.76
e	test	test	57 017 162	7.74
r	ryk	rIk	29 150 243	3.96
t	test	test	28 433 077	3.86
n	nasz	naS	27 047 875	3.67
i	PIT	pit	26 568 213	3.61
v	wilk	vilk	23 911 455	3.24
I	typ	tIp	23 875 687	3.24
j	jak	jak	22 550 363	3.06
p	pik	pik	21 742 544	2.95
s	syk	sIk	21 478 890	2.91
u	puk	puk	20 869 623	2.83
d	dym	dIm	19 141 562	2.60
k	kit	kit	18 919 934	2.57
m	mysz	mIS	18 548 063	2.52
l	luk	luk	15 558 031	2.11
n’	koń	kon’	13 957 066	1.89
z	zbir	zbir	12 073 293	1.64
tˆs	cyk	tˆsIk	10 823 185	1.47
f	fan	fan	9 972 436	1.35
w	łyk	wIk	9 929 083	1.35
b	bit	bit	9 436 766	1.28
x	hymn	xImn	9 148 491	1.24
g	gen	gen	8 928 754	1.21
S	szyk	SIk	7 975 642	1.08
Z	żyto	ZIto	6 309 944	0.86
tˆS	czyn	tˆSIn	6 091 250	0.83
s’	świt	s’vit	6 077 420	0.82
w∼	ciąża	ts’ow∼Za	4 244 488	0.58
tˆs’	ćma	tˆs’ma	4 206 577	0.57
dˆz’	dźwig	dˆz’vik	3 916 493	0.53
c	kiedy	cjedy	3 694 721	0.50
J	giełda	Jjewda	2 026 765	0.27
N	pęk	peNk	1 950 677	0.26
dˆz	dzwoń	dˆzvon’	1 846 929	0.25
z’	źle	z’le	997 176	0.13
j∼	więź	vjej∼s’	651 376	0.09
dˆZ	dżem	dˆZem	218 975	0.03
q	-	-	1	0.00

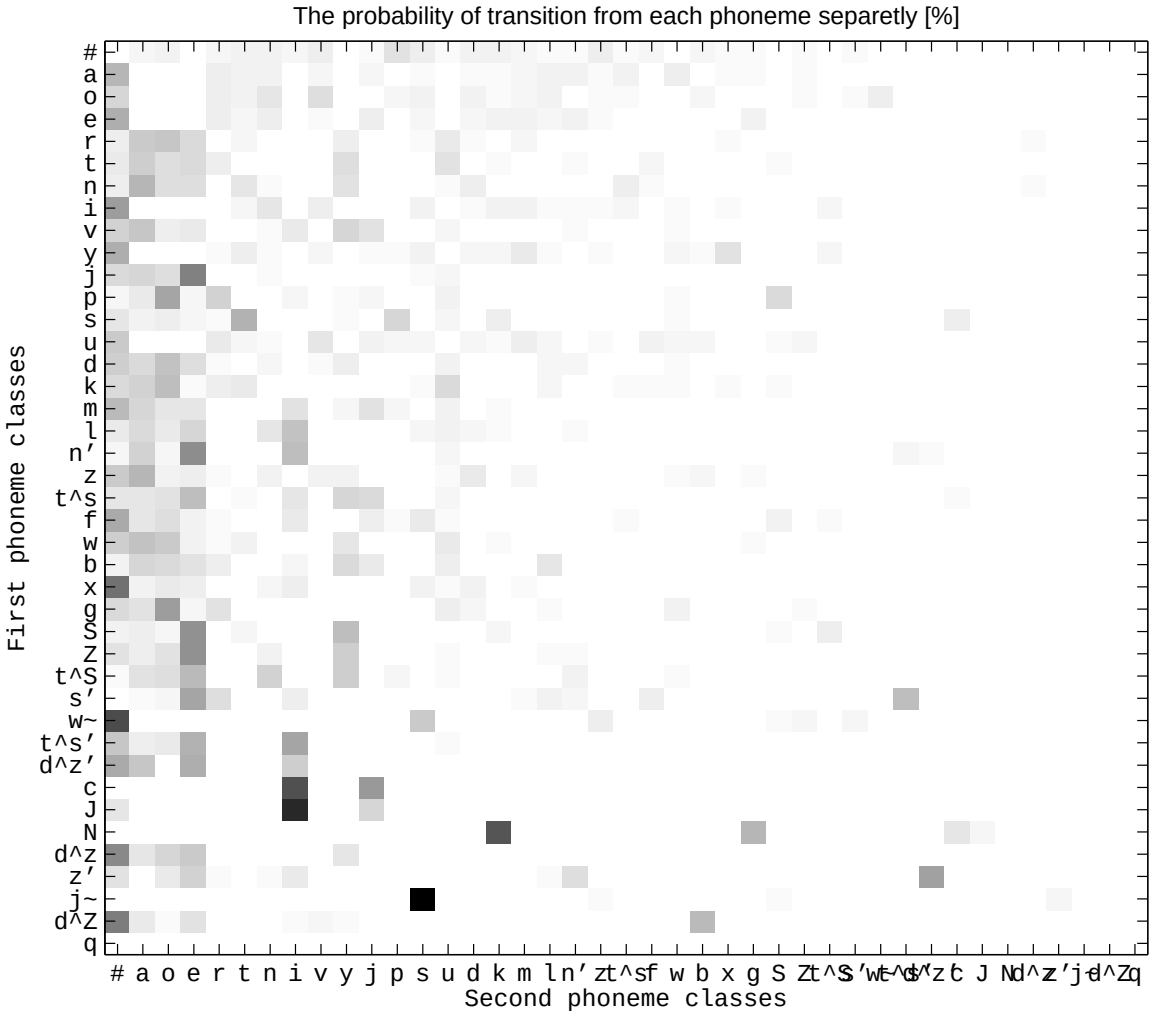

Fig. 1: Frequency of diphones in Polish (each phoneme separately)

2. Method Description

Sophisticated rules and methods are necessary to obtain of phonetic information from an orthographic text-data without simplifications which could cause errors [10], [11]. Transcription of text into phonetic data has to be applied first [12]. We used PolPhone [9] 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 Poznań and Kraków. For practical reasons we used our own digit symbols corresponding to SAMPA symbols, instead of typical ones, to distinguish phonemes easier while analysing received phonetic transcriptions. Linux stream editor (SED) was used to change original phoneme transcriptions into digits designed by us to simplify calculations with the script presented in Table 2.

Table 2: SED script for changing SAMPA transcriptions into one symbol only phonetic alphabet

s/##/#/g	s/z'/4/g
s/tˆs'/8/g	s/tˆs/7/g
s/dˆz'/X/g	s/dˆz/6/g
s/j∼/1/g	s/tˆS/0/g
s/w∼/2/g	s/dˆZ/9/g
s/s'/5/g	s/n'/3/g

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 a digit) what was conducted in Matlab on a high performance computer in the Academic Computer Centre CYFRONET AGH. The computer we

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 with an option of storing more in PBS queue.

3. 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 [13] or machine learning process [14]. 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.

The necessity of investigating large text corpus pointed to the use of the Polish phonetic transcription system PolPhone [15], [9]. 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 $T[1..m][1..n]$ is a *transcription table* and its cells meet the following requirements [9]. 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 precede currently transcribed character. The first row ($T[1][j]$, where $j = 1, ..., m$) contains all possible character strings that can proceed 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 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.

4. Rzeczpospolita Corpus and Results

Several newspaper articles in Polish were used as input data in our experiment. They are all from one newspaper called Rzeczpospolita from years 1993-2002. They are mainly on 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.

Total number of 736 715 777 phonemes were analysed. They are grouped into 40 categories (including space). Actually there is one more - q, which appeared once from a foreign name. Their distribution is presented in Table 1. Exactly 1 149 different diphones (Fig. 1 and Table 3) for 1 560 possible combinations were found, which is 74%. 17 278 different triphones (Table 4) were detected. Combinations like *#*, where * is any phoneme and # is space were removed. These triples should not be considered as triphones. The reason for it, is that first phoneme * and the second one are actually in 2 different words, while in this experiment we are interested in triphone statistics inside words. The list of the most common triphones is presented in Table 4. Assuming 40 different phonemes (including space) and subtracting mentioned *#* combinations, there are 62 479 possible triples. We found 17 278 different triphones. It leads to a conclusion that around 28% of possible combinations were actually found as triphones, which is similar to what we have found in our previous experiment [8] and now in other corpora. Young [7] estimates that in English, 60-70% of possible triples exist as triphones. However, in his estimation there is no space between words what changes distribution a lot. Some triphones may not occur inside words but may occur on combinations of an end of one word and beginning of another. We started to calculate such statistics without an empty space as the next step of our research. It is also possible that there are different numbers of triphones for different languages. Space (noted as #) frequency was 14.99. Let us divide 100 by 14.99 to receive an average length of words in phonemes as 6.7. The real average length is less then 6 because one space after each word is included.

We observed that all ngrams, even 1gram (Table 1), are different in this experiment then in the previous one [8]. They also differ slightly from yet unpublished statistics we collected from literature and Internet corpora. We used a slightly different version of SAMPA alphabet in [8], but the differences between experiments, in order of phonemes can be easily spotted. In [8] 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, tˆs, g, S, b, x, tˆS, dˆz, tˆs', dˆz', Z, s', o∼, N, w, z', dˆZ, 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 the possible applications 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. 2. We received another distribution than in the previous experiment [8] because

Table 3: *Most common diphones in Rzeczpospolita corpus*

diphone	no. of occurrences	percentage
e#	16 411 486	2.228
a#	15 503 774	2.105
#p	12 480 390	1.694
je	10 294 246	1.398
i#	9 298 146	1.262
o#	8 735 399	1.186
#v	7 658 002	1.040
na	7 119 701	0.9666
y#	7 083 354	0.9617
ov	6 990 033	0.949
#s	6 888 134	0.9352
po	6 885 441	0.9348
#z	6 336 099	0.8602
#o	6 088 722	0.8266
ro	5 978 333	0.8116
st	5 903 500	0.8015
n'e	5 720 903	0.7767
ra	5 711 314	0.7754
#d	5 548 842	0.7533
#t	5 274 406	0.7161
on	5 237 119	0.7110
ta	5 177 357	0.7029
#k	5 081 705	0.6899
#n	4 918 324	0.6677
va	4 876 548	0.6621
#m	4 717 016	0.6404
m#	4 612 790	0.6262
x#	4 589 623	0.6231
ko	4 577 042	0.6214
#r	4 460 984	0.6056
#i	4 338 869	0.5891
do	4 276 312	0.5806
#b	4 258 795	0.5782
v#	4 105 269	0.5573
u#	4 077 422	0.5536
#a	3 990 314	0.5417
ar	3 951 328	0.5364
#f	3 906 245	0.5303
re	3 865 551	0.5248
te	3 827 810	0.5197
or	3 786 968	0.5141
pr	3 668 247	0.4980
vy	3 646 770	0.4951
er	3 629 269	0.4927
ty	3 627 013	0.4924
to	3 605 958	0.4896
en	3 501 650	0.4754
ja	3 489 293	0.4737
li	3 482 998	0.4729
no	3450601	0.46847
aw	3450552	0.46846
ej	3437450	0.46668
ow∼	3323606	0.45123
sp	3313926	0.44991
d#	3307959	0.4491
ne	3305175	0.44873
n'i	3245003	0.44056
za	3224619	0.43779
Se	3166833	0.42994
al	3153450	0.42813

Table 4: *Most common triphones in Rzeczpospolita corpus*

triphone	no. of occurrences	percentage
#po	4 707 809	0.6393
#na	3 708 197	0.5035
n'e#	3 504 870	0.4759
na#	3 268 038	0.4438
#do	3 120 919	0.4238
ow∼#	2 707 879	0.3677
je#	2 670 609	0.3626
ej#	2 553 234	0.3467
#pr	2 539 370	0.3448
#za	2 525 949	0.343
#pS	2 508 259	0.3406
yx#	2 499 754	0.3394
ova	2 493 643	0.3386
ego	2 184 820	0.2967
go#	2 182 700	0.2964
pSe	2 093 032	0.2842
#ko	2 044 036	0.2776
#i#	2 006 665	0.2725
n'a#	1 998 177	0.2713
#vy	1 994 206	0.2708
#n'e	1 902 051	0.2583
sta	1 886 676	0.2562
#je	1 867 311	0.2536
vje	1 850 078	0.2512
#v#	1 846 576	0.2507
e#p	1 818 216	0.2469
#f#	1 716 208	0.2330
a#p	1 617 363	0.2196
ta#	1 548 535	0.2103
#ro	1 526 150	0.2072
#sp	1 504 621	0.2043
#re	1 498 372	0.2035
ne#	1 465 140	0.1989
ci#	1 462 658	0.1986
#s'e	1 457 281	0.1979
#te	1 457 057	0.1979
s'e#	1 456 304	0.1977
pro	1 422 882	0.1932
em#	1 417 226	0.1924
pra	1 399 453	0.1900
#o#	1 375 848	0.1868
cje	1 359 971	0.1847
Ze#	1 331 998	0.1809
#st	1 282 904	0.1742
#z#	1 271 576	0.1727
#ty	1 266 521	0.1720
ym#	1 262 608	0.1714
mje	1 231 848	0.1673
ovy	1 225 094	0.1664
ny#	1 211 519	0.1645
do#	1210360	0.16436
ent	1210223	0.16434
ontˆs	1195704	0.16237
tˆse#	1195072	0.16228
#Ze	1190593	0.16167
a#v	1156447	0.15704
e#v	1150563	0.15624
jon	1150364	0.15621
an'a	1123600	0.15258
o#p	1119044	0.15196

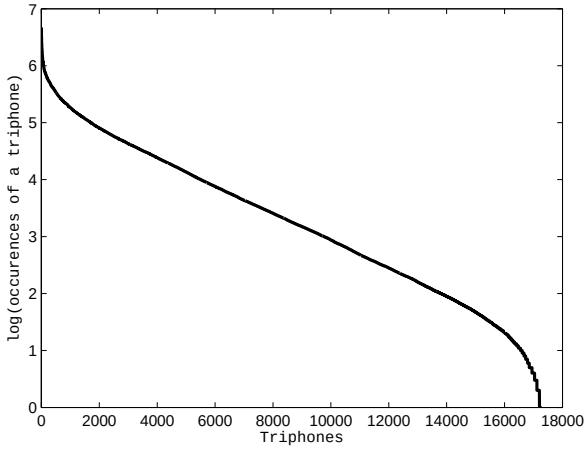

Fig. 2: Polish Phonemes in SAMPA alphabet

larger number of words were analysed. We have few triphones which occurred rarely, namely we found around 100 triphones with occurrences 1 to 8. 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.

5. N-gram Probability Model

Context-dependent modelling can improve speech recognition highly. Same 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. They often overlap each other. It results in phoneme waves interfering with others. Speech recognisers based on triphone models rather than phoneme ones are much more complex but give better results [16]. 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 [7] or a diphone (applying only left or right context) [16].

6. Conclusions

110 000 000 words from newspaper articles were analysed and statistics of Polish phonemes, diphones and triphones were created in this way. They are not fully complete but the corpus was large enough, that they can be successfully applied in language modelling. 28% of possible triples were detected as triphones, the very most of them at least 10 times. The full statistics are available on request by an email as a Matlab file.

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

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

[2] J. R. Bellegarda, "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] P. B. Denes, "Statistics of spoken English," *The Journal of the Acoustical Society of America*, vol. 34, pp. 1978–1979, 1962.

[4] E. J. Yannakoudakis and P. J. Hutton, "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] B. Kollmeier and M. Wesselkamp, "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment," *The Journal of the Acoustical Society of America*, vol. 102, pp. 2412–2421, 1997.

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

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

[8] B. Ziółko, J. Gałka, S. Manandhar, R. Wilson, and M. Ziółko, "Triphone statistics for Polish language," *Proceedings of 3rd Language and Technology Conference, Poznań*, 2007.

[9] G. Demenko, M. Wypych, and E. Baranowska, "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.

[10] J. Holmes, I. Mattingley, and J. Shearme, "Speech synthesis by rule," *Language and Speech*, vol. 7, pp. 127–143, 1964.

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

[12] D. Oliver, *Polish Text to Speech Synthesis, MSc. Thesis in Speech and Language Processing*. Edinburgh: Edinburgh University, 1998.

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

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

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

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